Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200616

Towards the Development of Language Analysis Tools for the Written Latgalian Language

Daiga DEKSNE [a,1] and Anna VULĀNE [b]
[a] *Tilde, Latvia*
[b] *Latvian Language Institute of the University of Latvia, Latvia*

Abstract. This paper reports on the development of spell checking and morphological analysis tools for Latgalian. The Latgalian written language is a historic variant of the Latvian language. There is a wide range of language analysis tools available for Latvian, whereas the Latgalian language lacks such tools. The work is done by the joint effort of linguists who work on morphologically marked lexicon creation and IT specialists who work on language tool development. For the creation of a morphological analysis tool, we reuse the FST technology used for the Latvian morphological analyzer. We create a spelling dictionary that can be used with the Hunspell engine. All tools are accessible via Web Service. For now, the Latgalian lexicon contains 13,139 lemmas marked by 105 inflection groups. The work of lexicon replenishment still continues.

Keywords. Proofing tools, morphological analysis, spell checking, FST transducer

1. Introduction

In the last decade, various essential language tools and technologies have been developed for the Latvian language, such as spelling checkers, morphological analyzers, taggers and parsers, speech recognition and synthesis tools, and machine translation systems [1]. Unfortunately, there is a lack of such tools for Latgalian.

There are three dialects in the territory of Latvia: the Livonian Dialect, the Middle Dialect, and the High Latvian dialect. Latvian literary language has been formed on the Middle Dialect. Latgalian is based on the Latgalian subdialects of the High Latvian dialect. It differs significantly from the Middle and Livonian Dialects as well as the Latvian literary language. The Latgalian written language is a historic variant of the Latvian language. In the 16th century, German clergymen began developing a language of writings on the basis of the low Latvian dialects. Later, this variety became the normed language of Latvian. On the other hand, in the 18th century, Catholic clerics began to form the language of Latgalian writings on the basis of the high Latvian (Latgalian) dialects.

Latgale is the eastern region of Latvia which was separated by a state border from the rest of Latvia's territory for almost 300 years, and it was then that the Latgalian written language developed. These historical circumstances determined the development

[1] Corresponding Author: Daiga Deksne; Tilde, Vienibas gatve 75a, Riga, Latvia, LV1004; E-mail: daiga.deksne@tilde.lv.

of a second written language tradition. Nowadays, the Latgalian written language is regularly used in Roman Catholic churches and some schools in the region, by book publishers and media, it can be heard in theatres and at concerts, on radio and television, at public events and gatherings around the country, as well as on the Web. Therefore, it is necessary to create a digital spelling tool to maintain the high quality of language use. Latgalians use either the native dialect or the Latvian literary language in their spoken communications. According to the Census 2011, 8.8 % of the Latvian population speak Latgalian on a daily basis, with 5.7 % of all children up to the age of 17. "Latgalian is spoken the most in Latgale region – 35.5 % of the population, although this reduces to 27 % amongst children up to the age of 17" [4].

The goal of this project is to create morphological analysis and spelling checking tools for the written Latgalian language. The work is done by the joint effort of linguists and IT specialists. Linguists with Latgalian language knolwedge develop a morphologically marked lexicon. The task of the IT specialists is to create tools that will help users to learn Latgalian and aid in text production without spelling mistakes.

The workflow for the creation of the language analysis tools consists of several steps:

- Establishing an environment for the creation of the morphologically marked lexicon (done by IT specialists and linguists);
- Creation of the lexicon (done by linguists);
- Development of language analysis tools (done by IT specialists);
- Development of a client-side application (done by IT specialists);
- Checking errors in the lexicon and identifying missing entries using a client-side application (done by linguists);
- Fixing errors and replenishing the lexicon (done by linguists);
- Rebuilding the tools including changes made in the lexicon (done by IT specialists).

There are several ways how morphologically marked lexicon can be created. The authors of *Grammatical Dictionary of Polish* [1] define inflection groups in a relational database. Words in the dictionary are linked to inflection groups. *Croatian Morphological Lexicon* [2] has three parts: a list of lemmas containing stems and inflectional pattern numbers, a list of endings including morphosyntactic category values, and a list of transformations applied on stems when concatenating words from morphemes. To facilitate the work of lexicographers, we used a simple approach. Template examples for paradigm definition for different part-of-speech were created specifying word forms used in Latvian. Linguists modified them and filled in with example word forms to cover all inflection patterns used in Latgalian. Defined inflection groups were used for word marking in the lexicon.

2. Creation of Morphologically Marked Lexicon

Some linguistic features of Latgalian not found in Latvian are as follows:

- Complete opposition of hard and soft consonants;
- The use of graphemes *y* and *uo*, which are not in the Latvian alphabet;
- Endings *-ys*, *-is* in place of Latvian *-as*, *-es*;
- Change of vowels under the influence of the next syllable vowel (*g**o**b**o**ls* and *g**a**b**a**leņš*);

- Prefix *da-* and the use of formants *-za-* and *-sa-* in reflexive prefixed verbs and nomens with endings *-inis*, *-ine*;
- A wider participle system;
- Many words have preserved more ancient meanings.

Given the fact that Latgalian written language is only taught in a handful of schools in Latgale on an optional basis or, alternatively, can only be learned through courses or self study, only a small proportion of the population are familiar with the orthographic norms of the Latgalian written language. Therefore, errors are prevalent in informal communications such as social media, text messaging, on-line comments, unedited literary works, etc. The most common of these include:

- The softening of consonants before the letters *e* or *i* (e.g. ***ņ**edeļa* : ***n**edeļa*);
- The use of an inappropriate root, suffix vowel or consonant, which is influenced by the specificity of pronunciation in a subdialect. For example, *ir – jir*, *jer*, *prīdē – prīdī*, *koln**i** – koln**y***, *gar**s** – gar**š***, *puišk**i**ns – puišk**y**ns*, *skrēj**a** – skrēj**e***;
- The use of dialect specific person forms of verbs, such as *ūgoiļom – ūgojom*, *dzīduo – dzīduoja*;
- The lack of consistency due to simultaneous use of the orthographic norms of 1929 [5] and 2007 [6];
- The phonetic translation of words from Latvian literary language;
- The use of written language according to the pronunciation in a subdialect, thereby, disregarding orthography.

A number of parallel forms were allowed to be in use during the transition to the improved and refined orthography. However, variants of graphemes, morphemes and forms were not created during the development of the spelling tool as unsubstantiated variability of word forms can lead to delayed embedding of the spelling rules.

Given the spelling tool has been developed based on the spelling rules of 2007, a number of typical mistakes were identified in the articles that were published in line with the orthographic norms of 1929 (see Table 1).

Table 1. Orthographic norms of 2007 and 1927

Language Unit	2007	1927
Diphthong designation	*uo – m**uo**te* *iu – **iu**diņs*	*ō – m**ō**te* *yu – **yu**diņs*
Diphthong of a root	*p**ie**rts*	*pērts*
Ending following the hard or palatalized consonant (in singular genitive, plural nominative, and plural accusative forms)	*-ys*, *-is* *lop**ys**, kuoj**is**, muot**is***	*-as*, *-es* *lop**as**, kōj**as**, muot**es***
Suffix	*-eja – Latv**eja***	*-ija – Latv**ija***
Past and future forms of 2nd and 3rd conjugation verbs with *-ēt*	*kavēt*, *redzēt – kav**ie**ju*, *redz**ie**ju*; *kav**ie**šu*, *redz**ie**šu*	*kavēju*, *redzēju*; *kavēšu* *redzēšu*

All parts of speech were covered and the following sources were used during the creation of the lexicon:

- Published Latgalian language dictionaries of spelling [7], [8];
- *Kalupe Subdialect Dictionary* [9];
- Scientific articles on Latgalian vocabulary and word-formation;
- Unpublished material on Latgalian subdialects and written language;
- Press and fiction texts;

- The meeting minutes and decisions of the Latgalian Written Language Subcommittee meetings on the correct spelling of work positions and professions, names of residents and toponyms, and other material.

In addition to the widely used vocabulary, different variants of the same concept encountered in dialects were included, however, spoken vocabulary was not provided. The notation on word-forms observed the spelling rules of the Latgalian language that stem from parallel forms and are found in dialects. Furthermore, older word-forms were included in the spelling tool in order to facilitate and preserve their use. In the case of homography, and if a word is part of two different paradigms, both words were included and the appropriate morphological classification was provided.

The language material was arranged in two files containing a description of a morphological system and a vocabulary. Paradigms of all word classes that can be inflected were developed and classified; a lexicon was created and morphologically marked. If applicable, the diminutive forms of nouns were included, as well as the present, past and participle forms of verbs. See Table 2 for examples of verb records in the lexicon. All columns are not filled for verbs belonging to the groups where inflected forms have the same stem as infinitive or can be derived by regular rules.

Table 2. Example of verb records in the Latgalian lexicon

word	group	pres1p	pres2p	pres3p	past1p	past3p	ppmasc	ppfem
bēgt	V12a	*bāgu*	*bēdz*	*bāg*	*biegu*	*bāga*	*biedzs*	*bāguse*
cyluot	V2uot							
badeit	V3eit							

Participles were not included as separate entries in the lexicon as they are automatically generated from the verb stems. See Table 3 for complete statistics of different part-of-speech words in the lexicon and the inflection groups defined.

Table 3. Statistics of the Latgalian lexicon

Part-of-speech	Number of lemmas	Number of groups
noun	5,010	29
verb	5,435	29
adjective	1,302	15
pronoun	109	15
adverb	931	1
numeral	140	12
particle	34	1
conjunction	23	1
preposition	18	1
interjection	137	1
Total	13,139	105

3. Development of Language Analysis Tools

We reuse Finite state transducer (FST) technology used in the development of the Latvian morphological analyzer [10].

3.1. Finite State Transducer

For the lexicon description, we use the Stuttgart Finite-State Transducer Toolkit (SFST) [11] as it allows the use of regular expressions, variables and different operators for text string transformation – concatenation, composition, insertion, and others. For transducer compilation, we use OpenFst toolkit[2]. For verbs, nouns and adjectives, we define inflection classes containing information about every word-form in the paradigm – form identifier, word-form's ending, corresponding lemma's ending, tags signalizing to which stem an ending can attach (see Figure 1). Once defined, this part of the transducer is reused when new entries are added to the lexicon.

```
$N5pl$ = <normEnd>{is}:{is}<414>:<n> |\
<altEnd1>{u}:{is}<415>:<n> |\
<normEnd>{em}:{is}<416>:<n> |\
<normEnd>{is}:{is}<417>:<n> |\
<normEnd>{em}:{is}<418>:<n> |\
<normEnd>{ēs}:{is}<419>:<n> |\
<normEnd>{is}:{is}<420>:<n>
```

Figure 1. Example of noun declension class definition

The dynamic part of the transducer is a set of stems linked to the declension groups (see Figure 2). This set is recreated when the lexicon is changed. Nouns can have up to three stems. Verbs have up to 11 stems according to conjugation paradigm. Stems not specified in the lexicon are generated according to regular palatalization rules.

```
<N5pl>     Dekšuor|is Dekšuor|u
<N5pl>     pušdīn|is  pušdīņ|u
<N5pl>     zuol|is    zuoļ|u
```

Figure 2. Example of noun stem representation

The non-inflectional part-of-speech words are represented as lexical entries followed by form identifiers. Adverbs, numerals, and pronouns are also included as lexical entries along with information on how to generate lemma from inflected form (see Figure 3).

```
tu<1456>:<p> |\
{teve}:{tu}<1457>:<p> |\
{tev}:{tu}<1458>:<p> |\
{tevi}:{tu}<1459>:<p> |\
{tevim}:{tu}<1460>:<p> |\
{tevi}:{tu}<1461>:<p> |\
```

Figure 3. Lexical entries for pronoun *tu* ('you')

Words in the transducer are represented as concatenation of separate parts. For example, verbs are represented as concatenation of items from a prefix set, a stem set,

[2] http://www.openfst.org/.

and an ending set. The prefix set has only three items – a prefix for negation, a prefix for the debitive mood form, and an empty prefix. The stem set has verb stems sorted by conjugation classes. The ending set has endings sorted by conjugation classes. The correct word forms are obtained by matching tags of constituent parts. For example, there are tags in the prefix part and the stem part that must match the ending part.

In the compiled transducer, the form identifiers are replaced by the morphological description strings that are based on MULTEXT-East format [12]. Each form's description is 28 symbols long string. Each position in a string is reserved for the value of a particular grammatical feature. The first position is reserved for part-of-speech, the second – for tense, the third – for gender, the fourth – for number, the fifth – for case, etc. Values of features are represented by a single symbol in a particular position. For example, in the second position, symbol 'p' (present tense), 's' (past tense), or 'f' (future tense) can be found. Not all positions are filled in for every word as each part-of-speech word has a different set of features. Verbs have tense, number, person, and mood. Nouns have gender, number, case, and diminutive marker. Adjectives have gender, number, case, definite ending marker, and comparative forms. Positions that are not relevant for the particular part-of-speech word are filled with value '0'.

We built two transducers. One provides morphological analysis description for a given word, whereas another generates all word-forms for a given lemma. The output is presented in XML format (see Figure 4).

```
<document><source_info original='Afrika' />
<word pos='n' baseform='Afrika'>
<form descr='n0fsn000000000n0000000000000' spelling='Afrika' />
<form descr='n0fsg000000000n0000000000000' spelling='Afrikys' />
<form descr='n0fsd000000000n0000000000000' spelling='Afrikai' />
<form descr='n0fsa000000000n0000000000000' spelling='Afriku' />
<form descr='n0fsi000000000n0000000000000' spelling='Afriku' />
<form descr='n0fsl000000000n0000000000000' spelling='Afrikā' />
<form descr='n0fsv000000000n0000000000000' spelling='Afrika' />
</word></document>
```

Figure 4. Example of form generation result in xml format

In the case of homoforms, description of every form for a given word is provided as well as the lemma of a particular word form. For example, the word 'molu' can be a verb in past tense, first person singular, indicative mood form and a noun in singular accusative, singular instrumental or plural genitive form (see Table 4).

Table 4. Morphological analysis of word 'molu'

Lemma	Part-of-speech	Form description
mola (side)	noun	<n0fsa000000000n0000000000000>
mola (side)	noun	<n0fsi000000000n0000000000000>
mola (side)	noun	<n0fpg000000000n0000000000000>
maļt (to grind)	verb	<vs0s00100i000000000000000000>

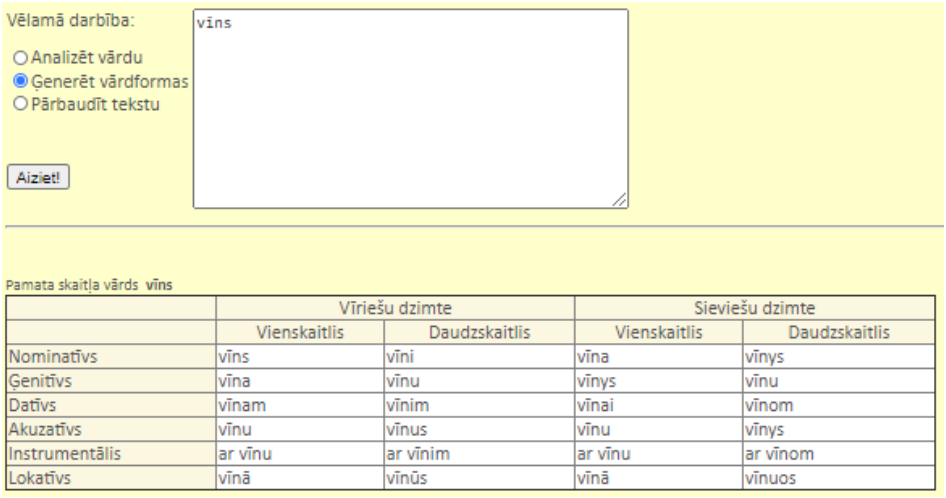
3.2. Hunspell Dictionary

We build the spell checking tool using the Hunspell library[3]. The dictionary for the spell checking tool is compiled from the files prepared for transducer compilation. The spelling tool checks text from the standard input (*stdin*). The produced output is in the HTML format. The misspelled words are included in <*span*> tags containing spelling suggestions in the *title* attribute. In such a way, spelling suggestions are shown as a tooltip when the mouse moves over a particular <*span*> element in any HTML browser application.

Another way how to use a compiled dictionary is by using plug-ins supporting Hunspell format, for example, DSpellCheck[4] plug-in.

4. Language Analysis Web Service

A Web Service is created to access the functionality of the developed tools from the Web environment. We have created an initial version of the Web form that enables users to check the correctness of a text, to see the morphological description of a desired word-form, and to see the tables displaying the full paradigm for a given word (see Figure 5).



	Vīriešu dzimte		Sieviešu dzimte	
	Vienskaitlis	Daudzskaitlis	Vienskaitlis	Daudzskaitlis
Nominatīvs	vīns	vīni	vīna	vīnys
Ģenitīvs	vīna	vīnu	vīnys	vīnu
Datīvs	vīnam	vīnim	vīnai	vīnom
Akuzatīvs	vīnu	vīnus	vīnu	vīnys
Instrumentālis	ar vīnu	ar vīnim	ar vīnu	ar vīnom
Lokatīvs	vīnā	vīnūs	vīnā	vīnuos

Figure 5. Web form with Inflection table for numeral *vīns* (‘one’)

5. Conclusion

In this paper, we described the creation of morphological analysis and spelling tools for the Latgalian written language. The work is still in progress. We have finished the first steps in the project as a result of which 105 inflection paradigms used in Latgalian have been defined and the basic lexicon containing 13,136 entries created. The definitions of inflection paradigms as well as the lexicon entries have been transferred to the finite state

[3] https://github.com/hunspell/hunspell.
[4] https://github.com/Predelnik/DSpellCheck.

transducer, and morphological analysis and spelling checking components created. The initial version of the Web Service allows analyzing a word, seeing its inflection paradigm, and checking the spelling of a text. The next phase involves an active work from linguists in checking the correctness of the words in the Latgalian lexicon and identifying missing entries.

Acknowledgments

Creation of the morphologically annotated Latgalian lexicon was supported by National Research Programme project "Latvian Language" (№ VPP-IZM-2018/2-0002).

References

[1] Woliński, M. A relational model of Polish inflection in Grammatical Dictionary of Polish. In: Language and Technology Conference; October 5-7, 2007; Poznan, Poland. Springer, Berlin, Heidelberg; c2007. p. 96-106.
[2] Tadić, M, Fulgosi, S. Building the Croatian morphological lexicon. In: Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages; April 13th, 2003; Budapest, Hungary. ACL; c2003. p. 41-45.
[3] Skadiņa I. Languages of Baltic Countries in Digital Age. In: Lupeikiene A., Vasilecas O., Dzemyda G. (eds) Databases and Information Systems. Proceedings of the 13th International Baltic Conference on DB&IS 2018; 1-4 July 2018, Trakai, Lithuania. Communications in Computer and Information Science, vol 838. Springer, Cham. p. 32-40.
[4] Spārīte L. (ed.) Bērnu tautība un mājās lietotā valoda. Bērni Latvijā; 2013. Available: http://www.csb.gov.lv/sites/default/files/publikacijas/nr_13_berni_lat-vija_2013_13_00_lv_en.pdf
[5] Strods, P. (kom. pr.) Nuteikumi par latgališu izluksnes ortografiju. In: Zīdūnis, 1929, Nr. 7, 21. (1669)–22. (1670) lpp.
[6] LR Tieslietu ministrijas Valsts valodas centrs. Latgaliešu rakstības noteikumi=Latgalīšu raksteibys nūsacejumi. Rīga, 2007. Available: https://m.likumi.lv/doc.php?id=164904
[7] Bukšs M. Placinskis J. Latgaļu volūdas gramatika un pareizraksteibas vōrdneica. Minhene: Latgaļu izdevnīceiba; 1973. 420 lpp.
[8] Strods P. Pareizraksteibas vōrdneica. Rēzekne: Dorbs un Zineiba; 1933. 213 lpp.
[9] Reķēna A. Kalupes izloksnes vārdnīca I–II. Rīga: Latviešu valodas institūts; 1998. 601 lpp.
[10] Deksne, D. Finite State Morphology Tool for Latvian. In: Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing; July 2013; St Andrews, Scotland. Association for Computational Linguistics; c2013; p. 49-53.
[11] Schmid H. A Programming Language for Finite State Transducers. In: Yli-Jyrä A., Karttunen L., Karhumäki J. (eds) Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005); 2005 September 1-2; Helsinki, Finland. Lecture Notes in Computer Science, vol 4002. Springer, Berlin, Heidelberg.
[12] Erjavec T. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources and Evaluation. 2012; 46/1:131-142.