Towards Development of Language Analysis Tools for the Written Latgalian Language

Daiga Deksne (Tilde, Latvia) and Anna Vulāne (Latvian Language Institute of the University of Latvia, Latvia)

Introduction

- The Latgalian written language is a historic variant of the Latvian language. 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.
- In this project, we create morphological analysis and spell checking tools for the written Latgalian language.
- Work is done by the joint effort of linguists and IT specialists.

Morphologically Marked Lexicon

Two files containing

- inflection pattern groups of a morphological system;
- words marked by the groups.

Example of verb records in the Latgalian lexicon

lemma	gr.	p1p	p2p	p3p	pst1p	pst3p	ppmsc	ppfem
bēgt	V12a	bāgu	bēdz	bāg	biegu	bāga	biedzs	bāguse
cyluot	V2uot							
badeit	V3eit							

POS	Num. of lemmas	Num. of groups
noun	5,010	29
verb	5,435	29
adj.	1,302	15
pron.	109	15
adverb	931	1
numeral	140	12
particle	34	1
conj.	23	1
Prep.	18	1
Interject.	137	1
Total	**13,139**	**105**

Development of Language Analysis Tools

- Words are represented as concatenation of a prefix, a stem, and an ending using *Stuttgart Finite-State Transducer Toolkit* .

Examples of noun declension class, noun stems and constituent part concatenation

```
$N5pl$ = <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>
```

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

```
zuoļ:l<altEnd1><N5pl><N5pl><altEnd1>{u}:{is}<415>:<n>
zuol<normEnd><N5pl><N5pl><normEnd>{em}:{is}<416>:<n>
```

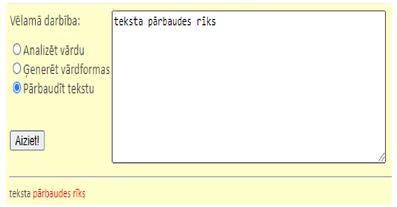
- Transducers compiled with *OpenFst toolkit*.

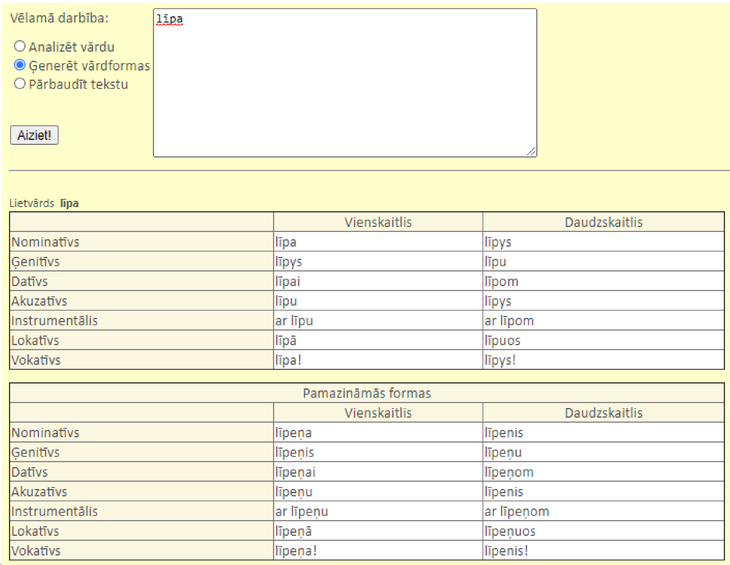
```
<document><source_info original='Afrika' />
<word pos='n' baseform='Afrika'>
<form descr='n0fsn000000000n0000000000000' spelling='Afrika' />
<form descr='n0fsg000000000n0000000000000' spelling='Afrikys' />
<form descr='n0fsd000000000n0000000000000' spelling='Afrikai' />
<form descr='n0fsa000000000n0000000000000'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

- The dictionary for the spelling checking in *HunSpell* format is compiled from the files prepared for the transducer.

Language Analysis Web Service

- morphological analysis of a word
- full inflection paradigm for a word
- spell checking





	Vienskaitlis	Daudzskaitlis
Nominatīvs	līpa	līpys
Ģenitīvs	līpys	līpu
Datīvs	līpai	līpom
Akuzatīvs	līpu	līpys
Instrumentālis	ar līpu	ar līpom
Lokatīvs	līpā	līpuos
Vokatīvs	līpa!	līpys!

Pamazināmās formas		
	Vienskaitlis	Daudzskaitlis
Nominatīvs	līpeņa	līpenis
Ģenitīvs	līpeņa	līpeņu
Datīvs	līpeņai	līpeņom
Akuzatīvs	līpeņu	līpenis
Instrumentālis	ar līpeņu	ar līpeņom
Lokatīvs	līpeņā	līpeņuos
Vokatīvs	līpeņa!	līpenis!

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