



Towards Integration of Latvian Web Services

Normunds Grūzītis

Madars Virza

Institute of Mathematics and Computer Science

University of Latvia

(IMCS UL)

IMCS UL Web Services

- <http://valoda ailab.lv/ws/>
 - Morphological **analyzer & synthesizer**
 - Morphological **tagger**
 - Tokenizer
 - Sentence splitter
 - Tagger
 - **Text-to-speech** synthesizer
 - **Dictionary** of the Standard Latvian Language
- Future plans: corpus query service
- Forthcoming: parser/generator of controlled Latvian

Morphological Analyzer/Synthesizer

```
Mozilla Firefox
http://valoda.ai...=analyze&w=ce|u
- <LexicalResource dtdVersion="16">
  - <GlobalInformation>
    <feat att="languageCoding" val="ISO 639-3"/>
  </GlobalInformation>
  - <Lexicon>
    <feat att="language" val="lav"/>
    - <LexicalEntry>
      <feat att="partOfSpeech" val="noun"/>
      - <Lemma>
        <feat att="writtenForm" val="celis"/>
      </Lemma>
    </LexicalEntry>
    - <LexicalEntry>
      <feat att="partOfSpeech" val="verb"/>
      - <Lemma>
        <feat att="writtenForm" val="celt"/>
      </Lemma>
    </LexicalEntry>
    - <LexicalEntry>
      <feat att="partOfSpeech" val="noun"/>
      - <Lemma>
        <feat att="writtenForm" val="ceļš"/>
      </Lemma>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

```
Mozilla Firefox
http://valoda.ai...ze&w=celt&pos=v
- <WordForm>
  <feat att="writtenForm" val="cels"/>
  <feat att="verbFormMood" val="indicative"/>
  <feat att="grammaticalTense" val="future"/>
  <feat att="grammaticalNumber" val="plural"/>
  <feat att="person" val="thirdPerson"/>
</WordForm>
- <WordForm>
  <feat att="writtenForm" val="ceļot"/>
  <feat att="verbFormMood" val="evidential"/>
  <feat att="grammaticalTense" val="present"/>
</WordForm>
- <WordForm>
  <feat att="writtenForm" val="ceļot"/>
  <feat att="verbFormMood" val="evidential"/>
  <feat att="grammaticalTense" val="future"/>
</WordForm>
- <WordForm>
  <feat att="writtenForm" val="celtu"/>
  <feat att="verbFormMood" val="conditional"/>
</WordForm>
- <WordForm>
  <feat att="writtenForm" val="jaceļ"/>
  <feat att="verbFormMood" val="debitive"/>
</WordForm>
```

Morphological Lexicon

POS	Lemmas	Word forms
Nouns	32 537	294 647
Verbs	12 007	347 874
Adjectives	6 098	682 976
Adverbs	6 508	6 508
Pronouns	51	472
Other	412	412
Total	57 613	1 332 889

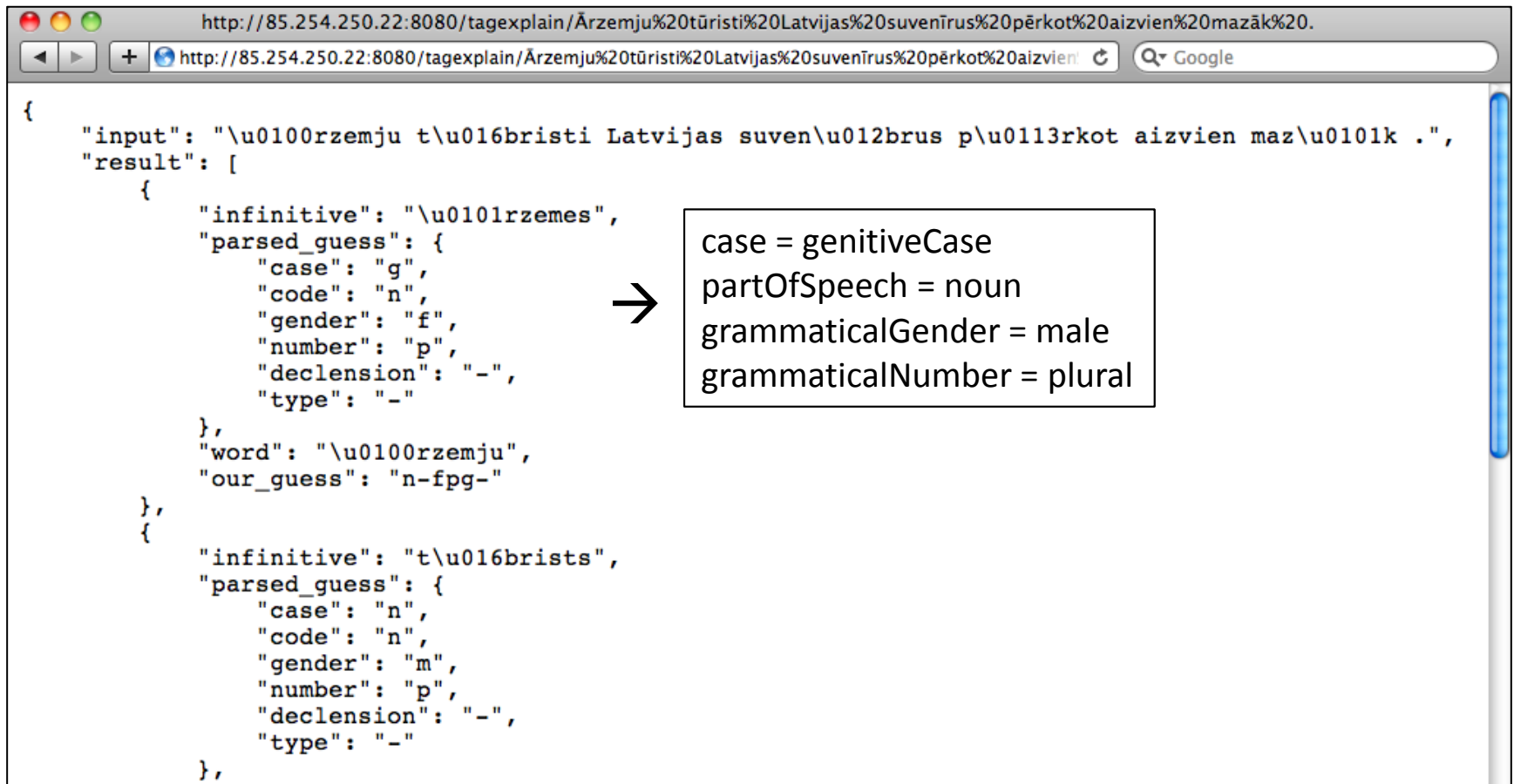
- **Past participle** forms currently are not supported
- **Analytical verb forms** (syntactic units) are not taken into account
 - Atomic constituents only (either functional or content words)

Annotation Standards Used

- Done: an ad-hoc (POS-specific) format → **LMF**
 - The **feature structure model** is generic, flexible and simple
 - For each particular category/word form: provide only those features that are actually **appropriate**
 - For a particular set of feature-values (for the same lemma) **alternative** word forms can be provided in the uniform way
- “Done”: compliance with **DCR (ISOcat)**
 - With slight **extensions**: e.g., values **evidential** and **debitive** are also used for the **verbFormMood** category
 - **Missing** concepts (terms)
 - **Different** linguistic traditions: how to categorize the same phenomenon
 - **Which** concept to choose: which is the primary component to use for classification
 - Use **PIDs** or the lexical identifiers?

POS Tagger

- Current output: ad-hoc feature structure in JSON
 - Tagset: kind of **Multext-East**
- Under development: standardization towards **ISOCat** and **MAF**



```
{
  "input": "\u0100rzemju t\u016bristi Latvijas suven\u012brus p\u0113rkot aizvien maz\u0101k .",
  "result": [
    {
      "infinitive": "\u0101rzemes",
      "parsed_guess": {
        "case": "g",
        "code": "n",
        "gender": "f",
        "number": "p",
        "declension": "-",
        "type": "-"
      },
      "word": "\u0100rzemju",
      "our_guess": "n-fpg-"
    },
    {
      "infinitive": "t\u016brists",
      "parsed_guess": {
        "case": "n",
        "code": "n",
        "gender": "m",
        "number": "p",
        "declension": "-",
        "type": "-"
      }
    }
  ]
}
```

→

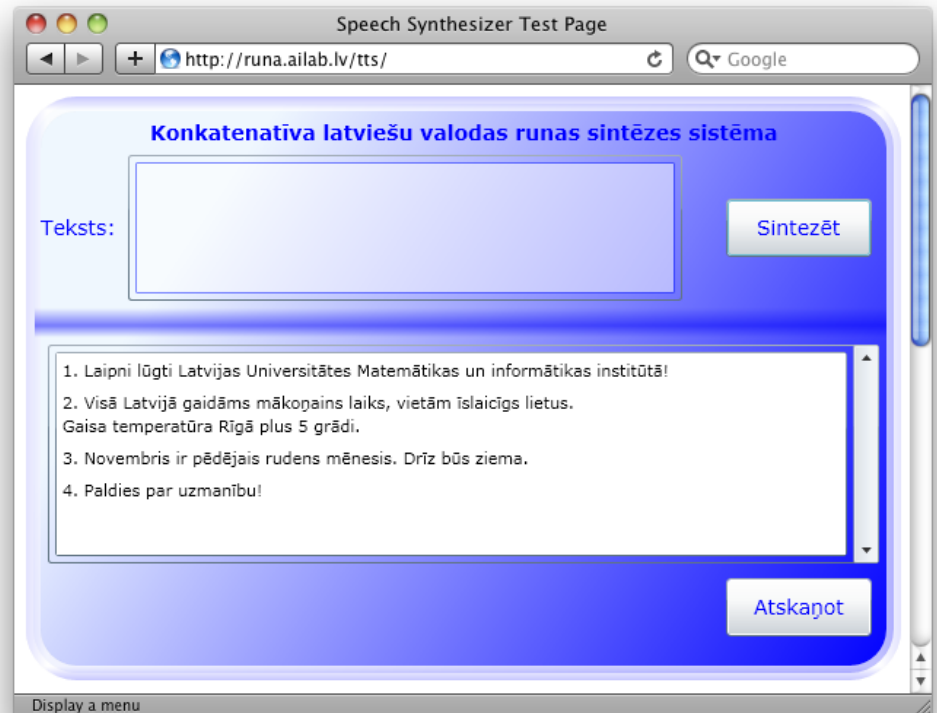
case = genitiveCase
partOfSpeech = noun
grammaticalGender = male
grammaticalNumber = plural

TODO: Text Processing Chain

- ISO (LAF/MAF): *stand-off* annotations
 - Tokenization
 - Sentence splitting
 - Tagging
- Looking towards integration in WebLicht
 - TCF → MAF; and vice versa?
- P.S. The current ISO status of LAF and MAF: international draft
 - Hopefully, converters to/from other well-known formats will be eventually available

Text-to-Speech Service

- The system is specifically adapted for **weather news**
 - Can be used for general language as well
- Response: **audio/mpeg** (MP3) or **text/plain** (URL)
- UI also available:



Dictionary of Latvian Language

- Rich entries (> 64 000), fine-grained structural annotations
- TODO: conversion from the proprietary format into LMF

```
- <s>
- <v>
  <vf>adata</vf>
  <gram pos="n">-as, s.</gram>
</v>
- <n nr="1">
  - <d>
    - <t>
      Tievs (metāla) irbulis (šūšanai), kam vienā galā ir asa smaile, bet otrā - spraudziņa pavediena ievēršanai.
    </t>
  </d>
- <fraz>
  <t>Nav kur adatai nokrist.</t>
- <n>
  - <d>
    - <t>
      Saka par cilvēku vai priekšmetu ciešu sablīvējumu kādā vietā, telpā.
    </t>
  </d>
- <g_piem_3>
  - <piem>
    - <t>
      Kas tad nu varēja puikām būt brīnišķīgāk kā iebraukt nakti kādā pārpildītā krogā, kas bāztin
```

Online Dictionary of Latvian: a Mash-up

LETONIKA Vārds: **celis** Meklēt

> 64 000 šķirklju
beta versija
18.06.2010.

Latviešu literārās valodas vārdnīca

Projektu finansē Latvijas Kultūras ministrija

Iespējamās pamatformas:
celis
celt
ceļš

Ortogrāfiski līdzīgie vārdi:
celš
cels

celis -ļa, v.

Locītava, kas savieno kājas augšstilbu ar apakšstilbu. Arī ceļgals.

Sīvs celis. Ceļa lūzums.

Palocīt (arī pieliekt) celi. — *Izdarīt nelielu reveransu (par bērniem).*

(No)mesties (arī krist) ceļos (arī uz ceļiem).

Ļoti pazemīgi lūgt.

Ļoti augstu vērtēt, pielūgt (kādu).

Uz ceļiem (arī ceļos) lūgties

Ļoti pazemīgi lūgt.

Nospiest uz ceļiem (arī ceļos).

Uzvarēt, piespiest padoties.

(No)dreb (arī ļodzās) ceļi

Saka, ja izjūt lielas bailes.

// Savienojumā «uz ceļiem»: *klēpt.*

Sēdēt tēvam uz ceļiem.

[Rādīt pilnu šķirkli](#) [Locīt šķirklja vārdu](#) [Izrunāt šķirkljavārdu](#)

Locījums	Vienskaitlis	Daudzskaitlis
Nominatīvs	celis	ceļi
Ģenitīvs	ceļa	ceļu
Datīvs	celim	ceļiem
Akuzatīvs	celi	ceļus
Lokatīvs	celī	ceļos

celis -ļa, v.

Locītava, kas savieno kājas augšstilbu ar apakšstilbu. Arī ceļgals.

Sīvs celis. Ceļa lūzums.

Palocīt (arī pieliekt) celi. — *Izdarīt nelielu reveransu (par bērniem).*

Sniegs mežā bija dziļš, vai līdz ceļiem. *Rīgas B 57, 76, 2.*

.. [sieviete] pastiepa zilās kleitas malu tālāk pāri ceļiem. *Wilks 5, 31.*

Meitene pieliec celi, kā skolā mācīts. *Kureijs 2, 213.*

(No)mesties (arī krist) ceļos (arī uz ceļiem).

Ļoti pazemīgi lūgt.

Tad nāk koncerta nagla, vēl viens solists, baritons, kuru [skolas] pārzinis tīri vai ceļos mezdamiem izlūdzies braukt uz šādu nomali.. *Zigmonte 1, 267.*

Ļoti augstu vērtēt, pielūgt (kādu).

Uz ceļiem (arī ceļos) lūgties

Ļoti pazemīgi lūgt.

Nospiest uz ceļiem (arī ceļos).

Uzvarēt, piespiest padoties.

Nē, tādu tautu nevar nospiest uz ceļiem, nevar iznīcināt, kur mātes nesalūst, apbedījušas savus kritušos dēlus.. *Lit M 62, 9, 1.*

(No)dreb (arī ļodzās) ceļi.

Saka, ja izjūt lielas bailes.

«Stāt!» beidzot priekšā atskanēja cieta pavēle. Huanito ceļi nodrebēja. *Grīva 1, 66.*

.. nodevējis nobāl, tam ceļi ļodzās. Kā uguns pa kūlu skrien brīvības vēsts.. *Sudrabkalns 1, 75.*

// Savienojumā «uz ceļiem»: *klēpt.*

Sēdēt tēvam uz ceļiem.

Kad kaķītis bij beidzis savu stāstu, ķēniņš paņēma viņu uz ceļiem.. *K Skalbe 1, 94.*

..Mācos es pazīt A un Ā, Sēdēdams viņam [tēvam] uz ceļiem. *Osmanis 2, 44.*



Thank you!