

Towards Integration of Latvian Web Services

Normunds Grūzītis
Madars Virza

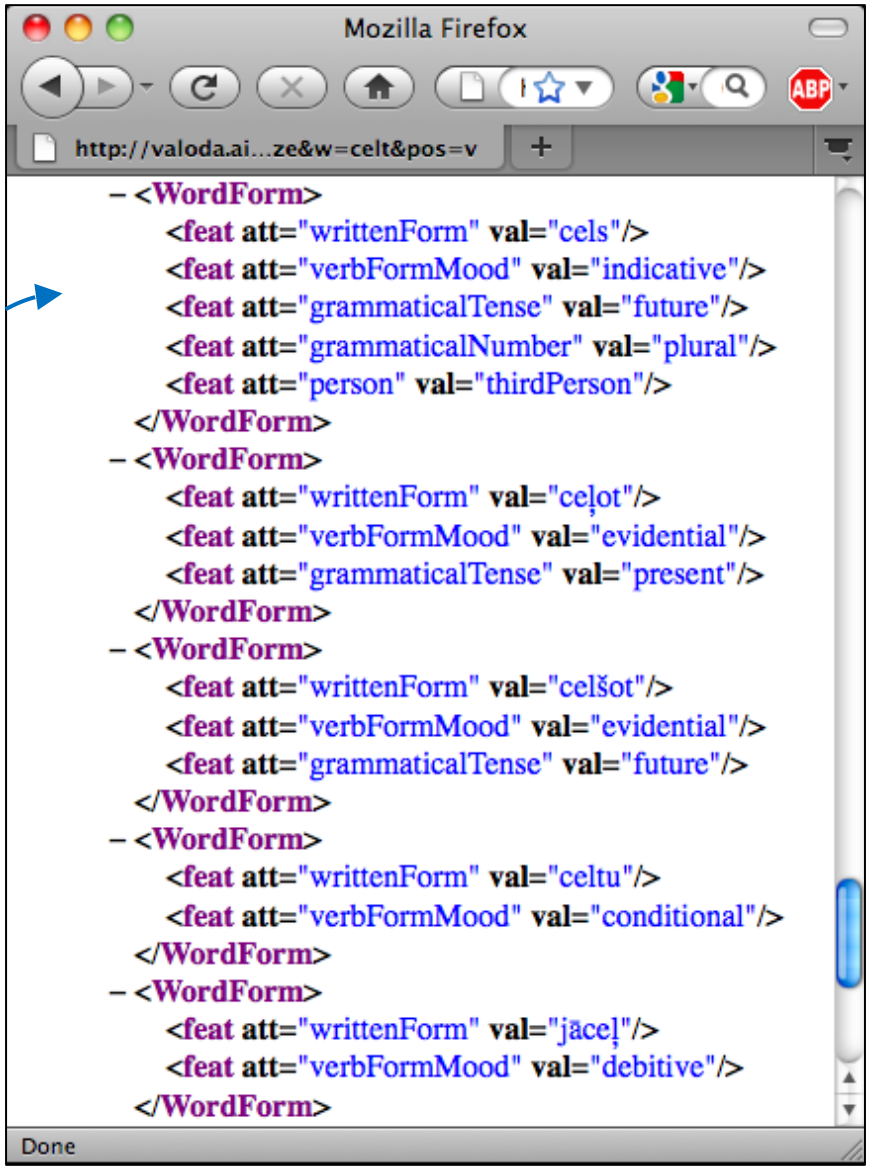
Institute of Mathematics and Computer Science
University of Latvia
(IMCS UL)

Workshop on Web Services, November 15–17, 2010, Freudenstadt, Germany

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



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 **PID**s 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**

http://85.254.250.22:8080/tagexplain/Ārzemju%20tūristi%20Latvijas%20suvenīrus%20pērkot%20aizvien%20mazāk%20.

http://85.254.250.22:8080/tagexplain/Ārzemju%20tūristi%20Latvijas%20suvenīrus%20pērkot%20aizvien

Google


```
{
    "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": "-"
            },
```


→

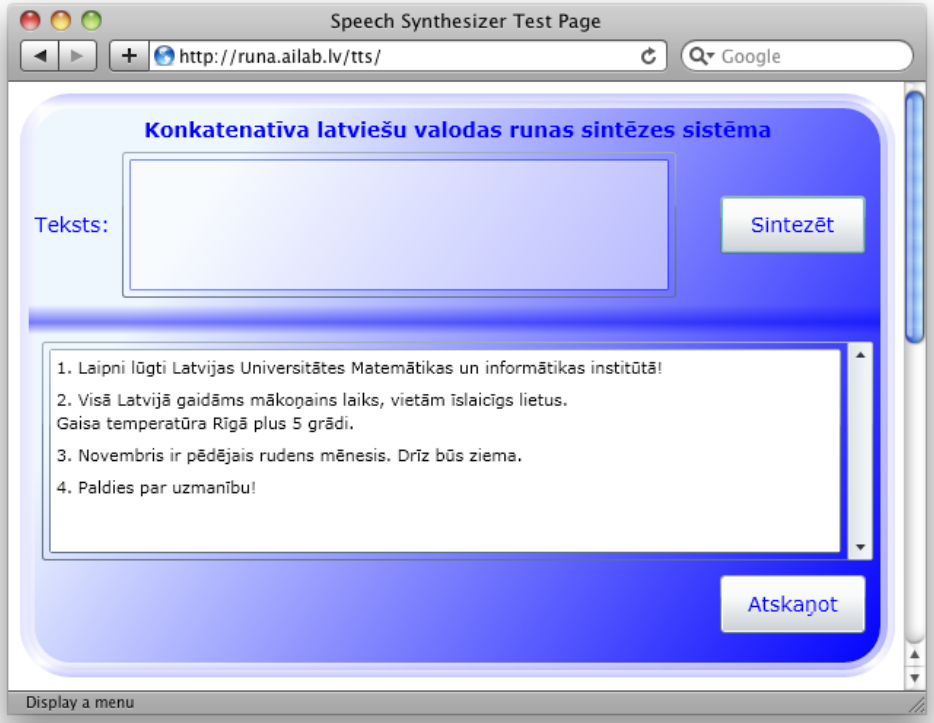
case = genitiveCase
partOfSpeech = noun
grammaticalGender = male
grammaticalNumber = plural

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

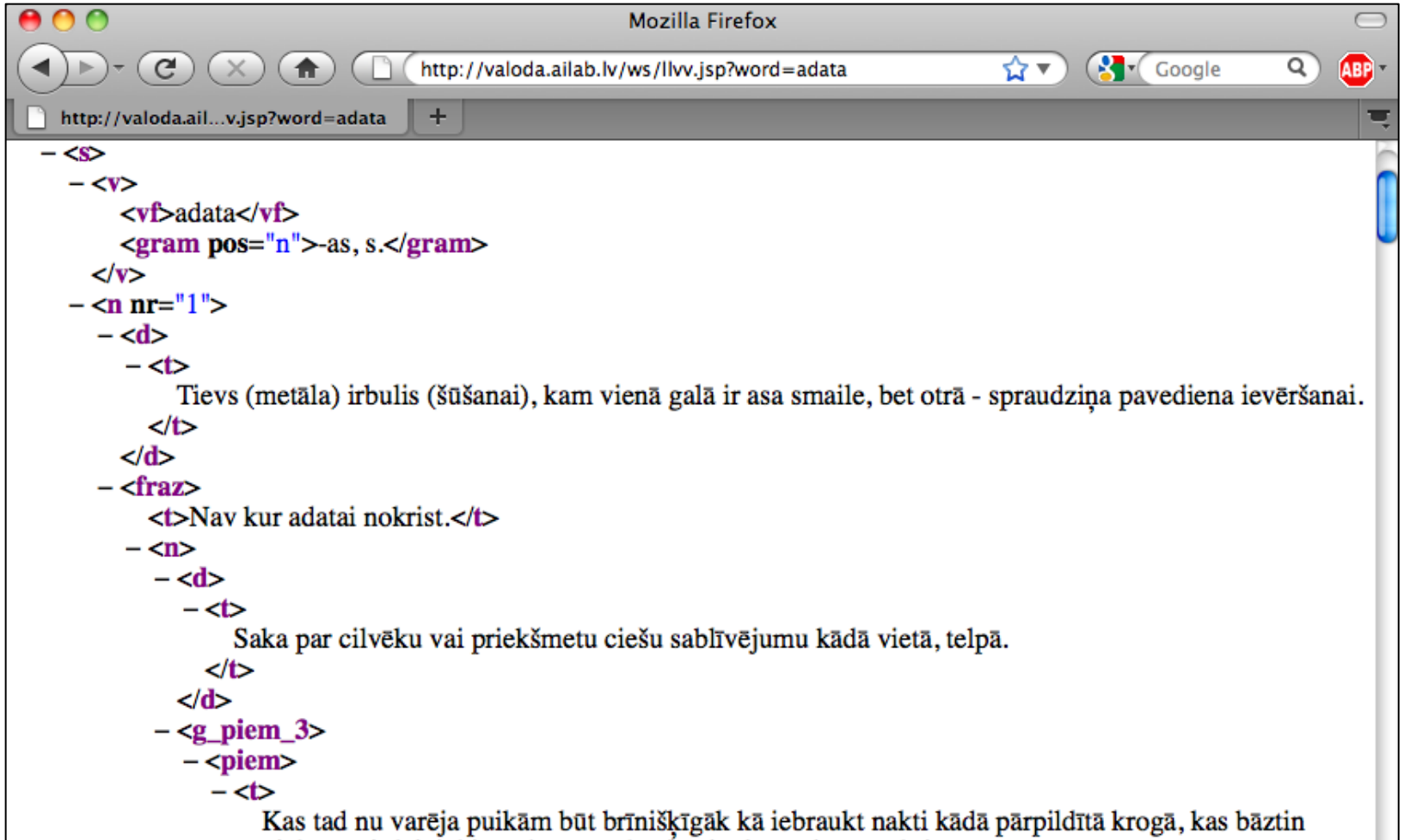
Text-to-Speech Service

- 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:

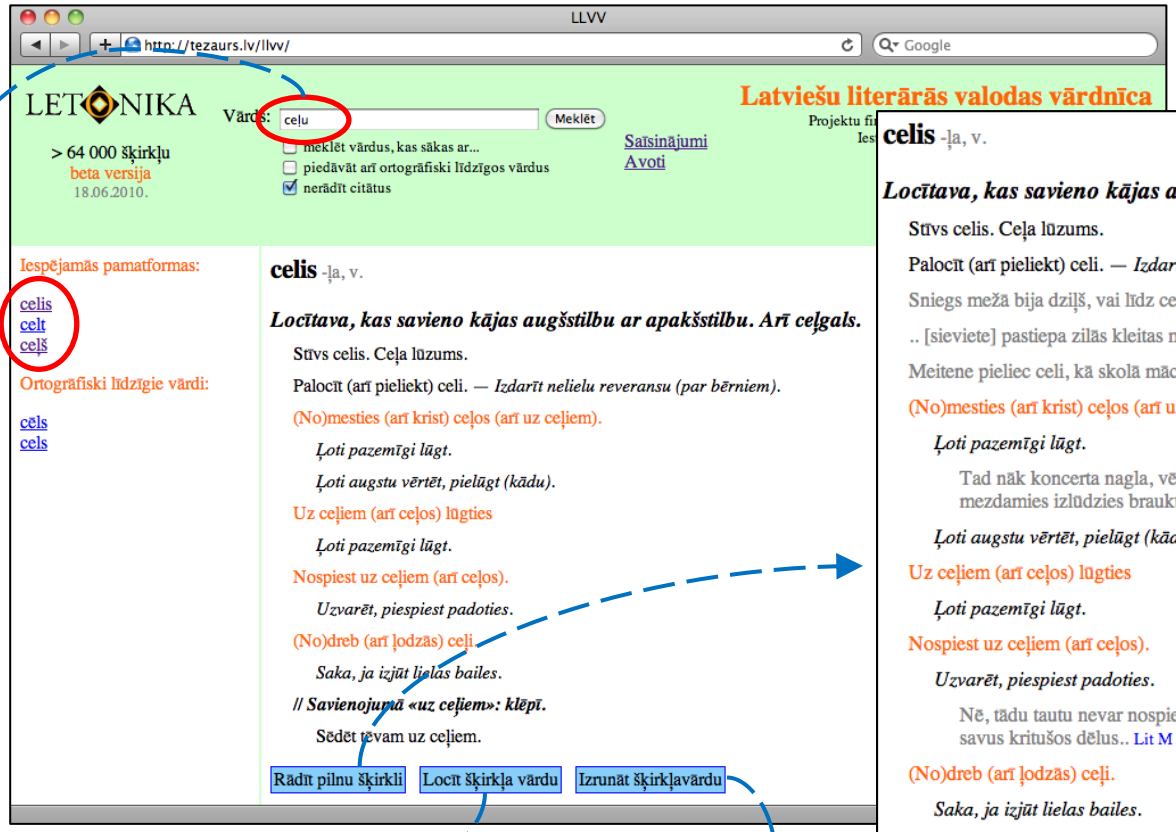


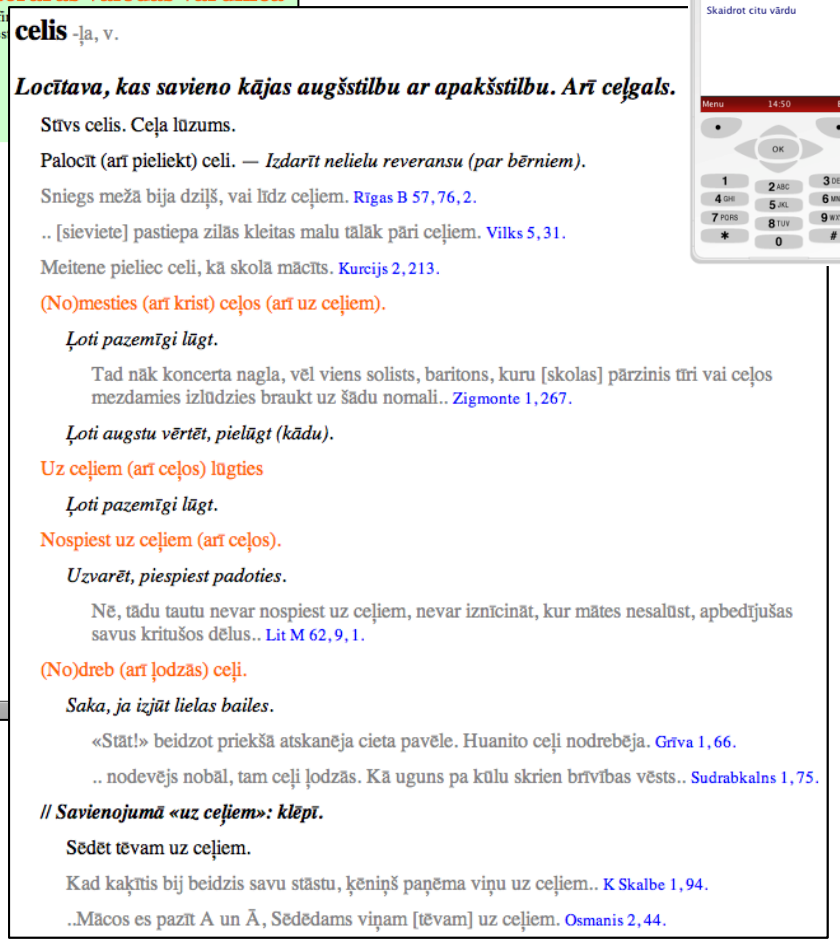
Dictionary of Latvian Language

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



Online Dictionary of Latvian: a Mash-up





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

Thank you!