Tezaurs.lv: the Largest Open Lexical Database for Latvian

Andrejs Spektors, Ilze Auzina, Roberts Dargis, Normunds Gruzitis,
Peteris Paikens, Lauma Pretkalnina, Laura Rituma, Baiba Saulite
University of Latvia, Institute of Mathematics and Computer Science
Raina blvd 29, Riga, Latvia
name.surname@lumii.lv

Abstract

We describe an extensive and versatile lexical resource for Latvian, an under-resourced Indo-European language, which we call Tezaurs (Latvian for 'thesaurus'). It comprises a large explanatory dictionary of more than 250,000 entries that are derived from more than 280 external sources. The dictionary is enriched with phonetic, morphological, semantic and other annotations, as well as augmented by various language processing tools allowing for the generation of inflectional forms and pronunciation, for on-the-fly selection of corpus examples, for suggesting synonyms, etc. Tezaurs is available as a public and widely used web application for end-users, as an open data set for the use in language technology (LT), and as an API – a set of web services for the integration into third-party applications. The ultimate goal of Tezaurs is to be the central computational lexicon for Latvian, bringing together all Latvian words and frequently used multi-word units and allowing for the integration of other LT resources and tools.

Keywords: Lexicon, Dictionary, Thesaurus, Morphology, Latvian, API

1 Introduction

Tezaurs,[1] a machine-readable lexicon and an online dictionary for Latvian, one of the 24 official EU languages, has been around for a while. The initial human-oriented version of this resource was made publicly available in 2009, comprising more than 125,000 entries that were consolidated from around 40 sources: modern and historical dictionaries, mostly available in a printed form. Since then, Tezaurs has been updated once every three months, and so far it has grown to more than 250,000 entries referring to more than 280 sources.

Tezaurs has attracted a large end-user base[2] and an increasing interest from third-party application developers, however, this work has not been published before.

The ultimate goal of Tezaurs is to be the central open computational lexicon for Latvian, allowing for the integration of other resources and tools for language technology (LT). An analogy can be drawn to SALDO (Borin et al., 2013), a lexical database for Swedish, the central component in an integrated infrastructure for computational lexical resources.

The idea, theoretically, is to bring together all the Latvian words and frequent multi-word units, along with their morpho-syntactic features and meaning, that have been used in the written texts. A secondary aim is to create and maintain a reliable source for language users, where they can verify and learn word forms, senses, and the lexical and grammatical valency.

For the language users, Tezaurs is already a highly popular online reference dictionary.[3] In addition to the fact that it is derived and consolidated from existing sources, Tezaurs provides added value: inflectional tables, phonetic transcriptions, synonym sets, and corpus examples. All the data and the accompanying web services are open-source and open-access.

2 Wordlist

Tezaurs is already a useful LT resource even only as an extensive authoritative vocabulary with (optionally) additional attributes for each word: the homonym index, the part-of-speech (POS) category, the inflectional paradigm, the phonetic transcription, domains of usage, stylistic markers and usage restrictions (dialecticism, archaic, colloquial, slang, vulgarity, child speech, etc.), as well as references to the sources.

The additional features allow for calling the Tezaurs web services, e.g. to generate a table of possible word forms based on the lemma and the inflectional paradigm, and for selecting a sub-vocabulary depending on the particular use case and application. Tezaurs has already been used as a source of general-purpose or customized wordlists in various text analysis pipelines that tend to have conflicting requirements on inclusion or exclusion of e.g. slang, archaisms or specific domains. To mention a few examples, Tezaurs' wordlists have been exploited in a newswire information extraction system (Paikens, 2014), in the transliteration and correction of OCR errors in historical texts (Pretkalnina et al., 2012), in an open-source spell checker, in various word games like Scrabble, and in other smaller research and commercially oriented applications.

Currently, a list of headwords along with their homonym indices, part-of-speech categories, inflectional paradigms and source references is available in the public repository of Tezaurs open data.[4] The remaining word attributes are under revision.

The wordlist is available also a web service that returns either the whole wordlist[5] or a detailed set of the above mentioned attributes for a particular word[6] along with homonyms, if any.

3 Morphological Information

The current end-user interface integrates a morphological web service, an extension of an open-source morpholog-

[1] `http://tezaurs.lv`

[2] Around 195,000 unique visitors (78% returning) over the last 12 months; around 67,500 sessions per month.

[3] More than 4.5 million page (entry) views per year. (There are about 2 million Latvian speakers worldwide.)

[4] `https://github.com/LUMII-AILab/Tezaurs`

[5] `http://api.tezaurs.lv/v1/words/`

[6] `http://api.tezaurs.lv/v1/words/doma`

ical analyzer for Latvian (Paikens et al., 2013), as a way of generating inflection tables for the lexical entries. Consequently, it also supports the inclusion of the Tezaurs wordlist (or a subset of it) as a lexicon for POS and morphological tagging and for lemmatization.

Although the source dictionaries do not include the morphological information of the headwords, or they include only a partial information, we can semi-automatically detect the POS category and the inflectional paradigm for each word. In most cases this can be done automatically, although quite a few cases have a chance for errors or uncertainty until the particular word groups are manually reviewed.

The main challenge is due to the tradition in the Latvian lexicography, which typically does not specify the POS category (a consequence of a highly inflected language). As of authors knowledge, the only Latvian dictionary that consistently includes POS tags is the Dictionary of Modern Latvian Language, MLVV.[7] MLVV is only now being transformed into a machine-readable form. When this is done, it will cover about 20% of entries in Tezaurs. Thus, in cases where the POS category cannot be unambiguously determined by the formal indications such as the word ending, the detection of the POS category and the specific inflectional paradigm of that category requires taking the meaning of the word (homonym) into account.

Another challenge is the need for manual reviewing of entries that include hints for non-standard inflectional paradigms, particularly in case of archaic and dialectal words whose inflection might not be aligned with the modern (standard) grammar, e.g. they can lack some word forms. Note that Tezaurs includes more than 90,000 dialectal and archaic words.

The morphological features of each word form included in the inflection table (returned by the web service) are only partially included in the end-user interface. The service provides the detailed morphological descriptions either in a form of MULTEXT-East morphosyntactic tags (Erjavec, 2004) or as an ISOcat feature matrix (Windhouwer and Wright, 2012) which is exemplified in Figure 1. The web service can be integrated in third-party applications in combination with the features provided by the Tezaurs wordlist (particularly, the inflectional paradigm).

4 Phonetic Transcription

In most cases, there is a one to one mapping between graphemes and phonemes in Latvian. Therefore the source dictionaries typically do not include information about the pronunciation of headwords, except in rare cases. Such cases include, for instance, words with contrastive syllable tones which can change the meaning of orthographically identical words, e.g. *zāle*: [zãle] (level tone) 'hall, large room' vs. [zâle] (broken tone) 'grass, herb'. However, two specific graphemes – 'e' pronounced as 'e' or 'æ', and 'o' pronounced as 'u͜o' (as in *doma* 'thought'), 'ɔ' or 'ɔː' – require an informed choice to pronounce the word correctly, and the pronunciation may vary across inflectional forms, even with the same spelling.

Our recent research on Latvian speech processing has resulted in a rule-based system that captures the pronunciation patterns and generates a machine-readable phonetic transcription for the given isolated word (Auzina et al., 2014). The system is now accessible as a Tezaurs web service[8], and it is being integrated in the Tezaurs website and the data sets (starting with the wordlist). In combination with a text-to-speech service (Pinnis and Auzina, 2010), this will make Tezaurs a more useful resource for language learners.[9] The transcription service, however, occasionally makes mistakes in case of the 'e' and 'o' graphemes. Again, after processing and integrating the MLVV data, this issue will be fixed at least for frequently used words.

In future, the morphological service (Section 3) can be extended by the transcription service to generate inflectional tables that are enriched with the phonetic transcriptions. Note that for verbs the pronunciation of the stem may change across inflectional forms.

5 Dictionary Entries

Another primary facet of Tezaurs: it is an extensive explanatory online dictionary. An entry generally represents a partial morphological information of the headword, usage restrictions (if any), the sense split, multi-word units and idioms, and source references. Homonyms and homographs (for more than 4,500 words) are given as separate entries with different indices.

Entries are internally organized by word senses (around 325,000 senses in total; 1.3 senses per headword). Each sense is explained by a full definition or a synonymous cross-reference. Morphological and stylistic restrictions can be specified also at the sense level. Senses often include embedded micro-entries of multi-word units along with their usage restrictions and glosses (around 32,000 in total). Some entries embed also idiomatic micro-entries (more than 11,000 in total) which are related to the whole entry. Usage examples are generated on-the-fly from a balanced corpus, where possible, as described in Section 7.

An example entry, as presented for the end-user, is given in Figure 2.

There is a web service available[10] that returns the dictionary entries in the LMF format, the standard interchange format for lexical resources (Hayashi et al., 2013).

6 Semantic Relations

Last but not least, Tezaurs is an extensive source for synonyms and other related concepts. Currently, we have put the focus on the synonymy relations which are automatically extracted from the implicit cross-references in the glosses which in turn follow traditional lexicographic guidelines. An issue is that although the sense split is obvious for the outgoing synonym sets (synsets), the incoming

[7] *Mūsdienu latviešu valodas vārdnīca*. University of Latvia, Institute of Latvian Language, 2004–2014 [`http://tezaurs.lv/mlvv/`]

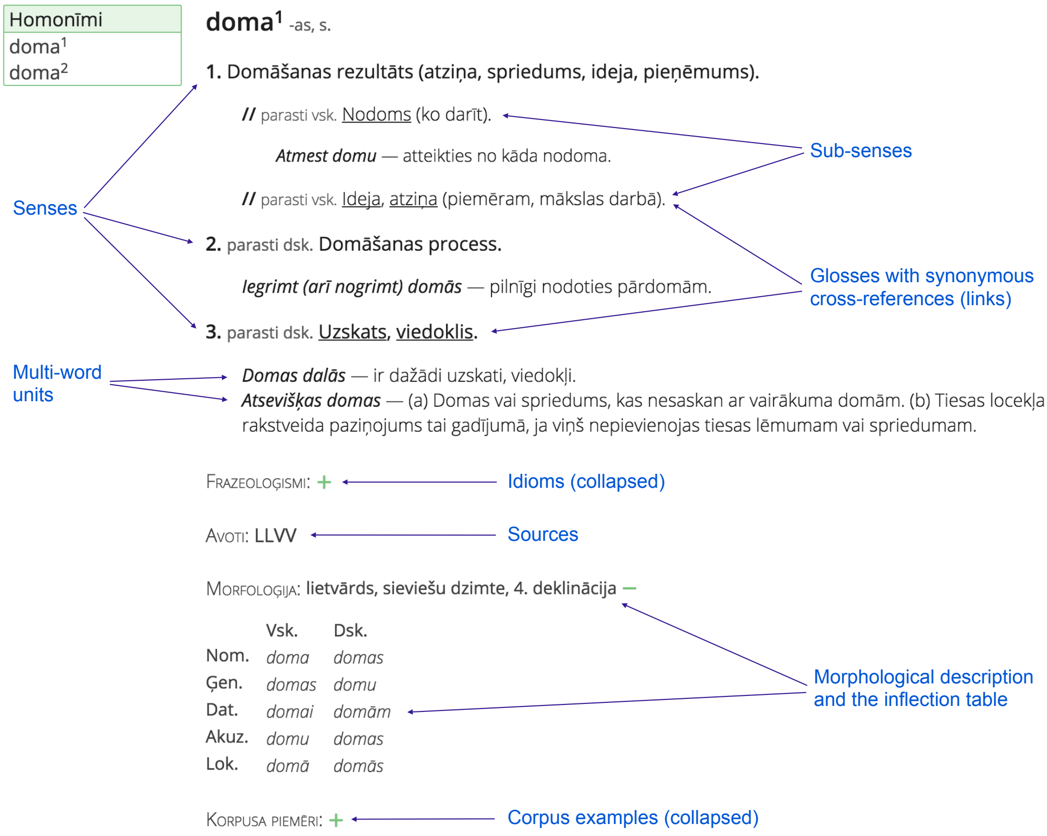
[8] `http://api.tezaurs.lv/v1/transcriptions/doma?encoding=ipa`

[9] `http://api.tezaurs.lv/v1/pronunciations/doma`

[10] `http://api.tezaurs.lv/v1/entries/doma/1`

```
[{
  "lemma" : "doma",
  "grammaticalGender" : "feminine",
  "declension" : "4",
  "partOfSpeech" : "noun",
  "wordForms" : [
    {"wordForm" : "doma",  "case" : "nominativeCase", "grammaticalNumber" : "singular"},
    {"wordForm" : "domas", "case" : "genitiveCase",   "grammaticalNumber" : "singular"},
    {"wordForm" : "domai", "case" : "dativeCase",     "grammaticalNumber" : "singular"},
    {"wordForm" : "domu",  "case" : "accusativeCase", "grammaticalNumber" : "singular"},
    {"wordForm" : "domā",  "case" : "locativeCase",   "grammaticalNumber" : "singular"},
    {"wordForm" : "doma",  "case" : "vocativeCase",   "grammaticalNumber" : "singular"},
    {"wordForm" : "domas", "case" : "nominativeCase", "grammaticalNumber" : "plural"},
    {"wordForm" : "domu",  "case" : "genitiveCase",   "grammaticalNumber" : "plural"},
    {"wordForm" : "domām", "case" : "dativeCase",     "grammaticalNumber" : "plural"},
    {"wordForm" : "domas", "case" : "accusativeCase", "grammaticalNumber" : "plural"},
    {"wordForm" : "domās", "case" : "locativeCase",   "grammaticalNumber" : "plural"},
    {"wordForm" : "domas", "case" : "vocativeCase",   "grammaticalNumber" : "plural"}
  ]
}]
```

Figure 1: A slightly simplified representation of `http://api.tezaurs.lv/v1/inflections/doma?paradigm=7` (‘thought’).



Homonīmi

doma^1

doma^2

doma^1 -as, s.

1. Domāšanas rezultāts (atziņa, spriedums, ideja, pieņēmums).

// parasti vsk. Nodoms (ko darīt).

Atmest domu — atteikties no kāda nodoma.

// parasti vsk. Ideja, atziņa (piemēram, mākslas darbā).

2. parasti dsk. Domāšanas process.

Iegrimt (arī nogrimt) domās — pilnīgi nodoties pārdomām.

3. parasti dsk. Uzskats, viedoklis.

Domas dalās — ir dažādi uzskati, viedokļi.

Atsevišķas domas — (a) Domas vai spriedums, kas nesaskan ar vairākuma domām. (b) Tiesas locekļa rakstveida paziņojums tai gadījumā, ja viņš nepievienojas tiesas lēmumam vai spriedumam.

Frazeoloģismi: +

Avoti: LLVV

Morfoloģija: lietvārds, sieviešu dzimte, 4. deklinācija —

	Vsk.	Dsk.
Nom.	*doma*	*domas*
Ģen.	*domas*	*domu*
Dat.	*domai*	*domām*
Akuz.	*domu*	*domas*
Lok.	*domā*	*domās*

Korpusa piemēri: +

Senses
Sub-senses
Glosses with synonymous cross-references (links)
Multi-word units
Idioms (collapsed)
Sources
Morphological description and the inflection table
Corpus examples (collapsed)

Figure 2: A slightly simplified end-user presentation of the entry `http://tezaurs.lv/#/sv/doma/1` (‘thought’).

sense is usually not specified in the glosses and, in general, has to be decided heuristically. In the long term, this will be a motivation to fix the ambiguous glosses manually.

The extracted synsets will be provided as open data along with the Tezaurs wordlists. We also intend to provide a corpus-driven list of semantically related words based on the *word2vec* approach (Mikolov et al., 2013). This does not necessarily reveal synonyms, but is interesting for human exploration and also as a feature for NLP tools.[11]

7 Corpus Examples

Availability of usage examples helps in understanding the meaning and customary usage of the words, however, appropriate sample sentences have generally not been available. Many source dictionaries do not include them, and for those that do, there are various problems that preclude directly using this data in Tezaurs - copyright issues, outdated usage, unavailability of the primary sources.

We currently provide[12] usage examples automatically retrieved from a balanced text corpus (Levane-Petrova,

[11] A demo of the already acquired vectors for Latvian is available at `http://api.tezaurs.lv/v1/embeddings/`

[12] `http://api.tezaurs.lv/v1/examples/doma`

2012), which provides adequate examples of contemporary usage for common words. The major issue that we encounter is the handling of homographs: morphological tagging and automatic word sense disambiguation helps, but is not perfect and needs manual review of such results.

While this provides useful results for common words, the coverage is limited by the size of corpus and for rare words usage examples are arguably even more important. This is an active direction of ongoing work to integrate data available from large unbalanced corpora of varying quality and/or web searches.

8 Sources

The primary source that has been used to derive the Tezaurs entries is the Dictionary of Standard Latvian Language, LLVV.[13] Almost 65,000 entries have been derived from LLVV (more than 25% of all Tezaurs entries).

There are about 20 secondary sources, each of them used in at least 1% of all entries (in total, around 149,000 entries refer to the secondary sources). The rest is a long tail of about 260 peripheral sources, each of them used in less than 1% of all entries; about 62,000 entries in total. Among them, less than 60 sources are used in 0.1–1.0% of all entries (each); about 55,000 entries in total.

9 Conclusion and Future Tasks

Tezaurs has acquired an important role for the human consumption (incl. professional translators, students, researchers, terminologists). We have also used this data set internally in the development of NLP tools, e.g. to extend the coverage of the POS-tagger (Paikens et al., 2013), to validate the correction of OCR errors (Pretkalnina et al., 2012), etc. We are anticipating an interest from researchers and application developers in the Tezaurs open machine-readable data and web services. The database attracts more and more interest from third-party application developers, both open-source and commercial, e.g. to be integrated in information retrieval systems, spellcheckers, style checkers, language games etc.

Future tasks include separate research problems that can be addressed based on this work. To mention some of them:

- Integration with a verb valency lexicon for Latvian (Nespore et al., 2012). The mapping of particular word senses to verb valencies needs to be done manually, which is feasible for the frequently used verbs.
- Providing corpus-based typical collocation information for each word.
- Further development of the semantic relations between word senses towards a WordNet-like semantic network.
- Integration with Linked Open Data to allow for word-sense grounding etc.
- Linking corpus usage examples to specific word senses by using word embeddings or similar techniques.

Acknowledgements

This work has been partially supported by Latvian State Research Programmes: Letonika (Project No. 3), NexIT (Project No. 1) and SOPHIS (Project No. 2).

References

Auzina, I., Pinnis, M., and Dargis, R. (2014). Comparison of rule-based and statistical methods for grapheme to phoneme modelling. In *Human Language Technologies – The Baltic Perspective*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 57–60. IOS Press.

Borin, L., Forsberg, M., and Lonngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1535–1538, Lisbon, Portugal.

Hayashi, Y., Monachini, M., Savas, B., Soria, C., and Calzolari, N., (2013). *LMF as a foundation for servicized lexical resources*, pages 201–213. Wiley.

Levane-Petrova, K. (2012). The balanced corpus of modern Latvian and the text selection criteria. *Baltistica*, VIII Priedas:89–98.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)*.

Nespore, G., Saulite, B., Gruzitis, N., and Garkaje, G. (2012). Towards a Latvian valency lexicon. In *Human Language Technologies – The Baltic Perspective*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 154–161. IOS Press.

Paikens, P., Rituma, L., and Pretkalnina, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 267–277, Oslo, Norway.

Paikens, P. (2014). Latvian newswire information extraction system and entity knowledge base. In *Human Language Technologies – The Baltic Perspective*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 119–125. IOS Press.

Pinnis, M. and Auzina, I. (2010). Latvian text-to-speech synthesizer. In *Human Language Technologies – The Baltic Perspective*, volume 219 of *Frontiers in Artificial Intelligence and Applications*, pages 69–72. IOS Press.

Pretkalnina, L., Paikens, P., Gruzitis, N., Rituma, L., and Spektors, A. (2012). Making historical Latvian texts more intelligible to contemporary readers. In *Proceedings of the LREC Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects*, Istanbul, Turkey.

Windhouwer, M. and Wright, S. E., (2012). *Linking to Linguistic Data Categories in ISOcat*, pages 99–107. Springer.

[13] *Latviešu literārās valodas vārdnīca*. 1.–8. Riga: Zinātne, 1972–1996 [`http://tezaurs.lv/llvv/`]