The modern electronic dictionary that always provides an answer

Daiga Deksne, Inguna Skadiņa, Andrejs Vasiļjevs

Tilde, Vienības gatve 75a, Riga, Latvia
E-mail: daiga.deksne@tilde.lv, inguna.skadina@tilde.lv, andrejs@tilde.lv

Abstract

This paper presents the Tilde Dictionary Browser (TDB), an innovative dictionary browsing environment for a wide range of users: language learners, language teachers, translators, and casual users. We describe several techniques to maximise the likelihood of providing users with a useful result even when searched items do not have a direct match in the dictionary due to misspellings, inflected words, multi-word items or phrase fragments, or where there is a lack of data in the main dictionary. TDB is targeted for broad use on multiple platforms and is implemented as desktop software, and a Web and mobile application. The desktop version of TDB currently contains dictionaries for more than 20 language pairs, including the languages of the Baltic countries, and is easily extendable to other languages. Besides the data from translation dictionaries, TDB also provides information from different online resources, such as terminology dictionaries, as well as integrates the machine translation facility.

Keywords: electronic dictionaries; machine translation; spelling checker; morphological analyser; text-to-speech synthesis.

1. Introduction

In the last two decades, electronic dictionaries have been established among the most widely used software applications for non-English speakers, and the majority of users prefer electronic dictionaries to printed ones (Koren 1997). Different models for electronic dictionaries have been of interest to researchers for a long time (for an overview, see de Schryver, 2003). In their work, Oppentocht and Schutz (2003) describe the advantages of electronic dictionaries (e.g., explicit information, consistency, reusability, etc.). Detailed analysis of electronic dictionaries from different viewpoints is presented by Müller-Spitzer; her later findings are related to user needs and usage scenarios of electronic dictionaries (Müller-Spitzer et al., 2011). There has also been a lot of research on the typology of electronic dictionaries (e.g., Ide 1993; Sharpe 1995; Lehr 1996) and the different types of users.

When using a paper dictionary, the user usually must flip through pages to find the sought-after entry, whereas when using an electronic dictionary, the user can type the word in a search field or choose an entry from a word list. However, several authors (Měchura 2008; Nessi and Haill 2002) point out that users often fail to locate the information that they need. Users often search dictionaries for words that cannot be found in them, or cannot be found in the form in which they have typed them:

misspellings, inflected words, multi-word items, phrase fragments or even whole sentences. Many electronic dictionaries fail to return useful results when being searched for anything other than exactly matching units.

The aim of our work was to develop a dictionary software that is able to provide useful information for all types of search queries and information needs, including many problematic cases, i.e., when searched items do not have any direct matches in the dictionary data.

In the dictionary software, Tilde Dictionary Browser (TDB), that is presented in this paper, we have applied several techniques to maximise the likelihood of providing users with useful results:

- The entries from a main dictionary and possibly several terminology and explanatory dictionaries are merged in a single list, allowing users to get consolidated information from **several dictionaries simultaneously**.
- In the case of **incorrect spelling,** TDB suggests possible corrections and provides their translations.
- **For languages with rich morphology,** users can find translations for words that are not in base form, as usually dictionary entries are. With the help of the morphological analyser, possible base forms are obtained and their translations are displayed.
- Users can also **see** all of the **inflectional forms** for a particular word.
- If a user wants to see **usage examples** for a particular word, the search engine will show all dictionary entries containing this word, even if it is not a headword or translation, but part of a longer multi-word phrase.
- Users can also **search** terminology dictionaries **in the Web,** and the results will be displayed in the same uniform way along with the local dictionary entries.
- If there is **no entry in lexicon** to a user's request, the request can be redirected to a machine translation (MT) system on the Web, which will then translate and present the translation in TDB translation view.
- For those who are learning a language, TDB provides a **text-to-speech facility** that allows to hear the pronunciation of the selected dictionary entry.

Currently, TDB includes numerous general and specialised dictionaries for 19 translation directions: from English, French, German and Russian into Latvian and vice versa, from English, French, German and Russian into Lithuanian, as well as Latvian-Lithuanian, Lithuanian-Latvian and Estonian-Latvian. More than 25 terminology dictionaries are integrated into the TDB.

The dictionary content is licensed from leading lexicographers (authors of printed

dictionaries). The cooperation with authors goes beyond the licensing of existing content of printed dictionaries: using corpora processing techniques we provide lexicographers with lexical items that are not included into dictionaries as they have appeared recently. Such lexical items are then investigated by lexicographers and after validation added to the corresponding TDB lexicon. As a result TDB allows the location of lexical items that are not yet available from any printed dictionary.

TDB has been incorporated into several commercial products (Tildes Birojs, Tildes Biuras) and is also extended (while maintaining the same functionality) for dictionary look-up on the Web and on mobile phones. It is one of the most popular software applications in the Baltic countries, with about 400 000 users.

In this paper, we describe the functionality of the Tilde Dictionary Browser in detail, demonstrate the importance of language technologies in a modern electronic dictionary, discuss scalability and interoperability issues in different media, and present common application scenarios for a modern electronic dictionary.

2. Consolidation of data in dictionary entry creation

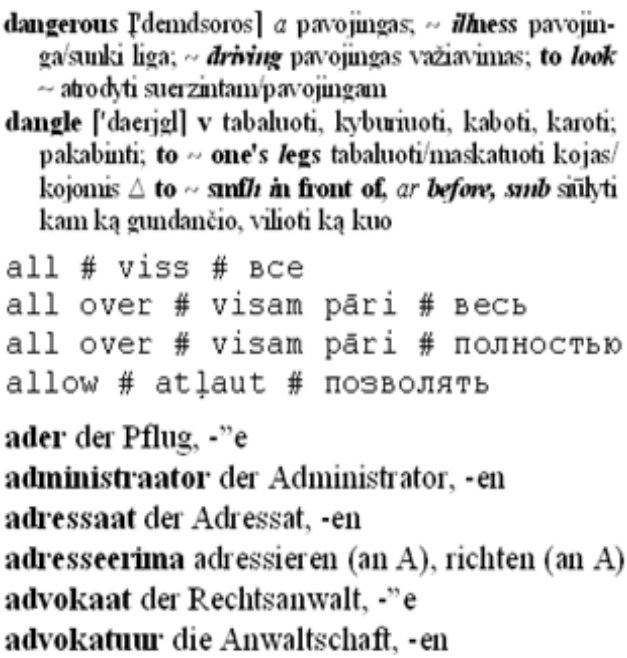
While a printed dictionary limits a search to the particular dictionary, electronic dictionaries can provide users with the ability to work with **several dictionaries simultaneously**. For this, entries from a main dictionary, and possibly several terminology, explanatory and synonym dictionaries, are merged in a single alphabetical list. Users can browse the entry just by clicking on a particular word in a list or search for a particular word or phrase by typing or copying it in a search field.

2.1 Forming a lexical entry: merging different sources

A logical part of a dictionary is an entry. However, dictionary entries may have very diverse formats. Some entries are very simple – just a word in a source language and a single or several translations in a target language.

More complex entries may contain translations grouped into several meanings, pronunciations, grammatical information, comments, usage samples and their translations, and explanations. Explanatory and synonym dictionaries usually have entries in a single language, while entries in translation and terminology dictionaries usually are in two or more languages.

The original formatting of dictionary entries is also very different: from simple tab or space separated words to entries with a rich formatting. Some samples of diverse dictionary formats are shown in Figure 1.



dangerous ['deindʒərəs] *a* pavojingas; ~ ***illness*** pavojinga/sunki liga; ~ ***driving*** pavojingas važiavimas; **to** ***look*** ~ atrodyti suerzintam/pavojingam

dangle ['daeŋgl] *v* tabaluoti, kyburiuoti, kaboti, karoti; pakabinti; **to** ~ **one's** ***legs*** tabaluoti/maskatuoti kojas/kojomis △ **to** ~ ***smth*** **in front of,** *or* ***before, smb*** siūlyti kam ką gundančio, vilioti ką kuo

```
all # viss # все
all over # visam pāri # весь
all over # visam pāri # полностью
allow # atļaut # позволять
```

ader der Pflug, -"e
administraator der Administrator, -en
adressaat der Adressat, -en
adresseerima adressieren (an A), richten (an A)
advokaat der Rechtsanwalt, -"e
advokatuur die Anwaltschaft, -en

Figure 1: Samples of different dictionary formats in printed dictionaries

The task of a modern electronic dictionary browser is to present the entries from different sources in a uniform way. This is achieved by parsing original dictionaries and internal representation of their entries in an XML format.

We have developed a special XML format for dictionary entry representation (Figure 2). This format differs from Text Encoding Initiative (TEI) guidelines, however, it can be transformed to TEI rather easily. About twenty different XML tags mark the different semantic parts of an entry, but not all of them are used in every dictionary.

```
<entry title="ābece">
<title>ābece</title>
<gram>n</gram>
<mean digits="1" symbol="." />
<transl>ABC</transl><comment>(book)</comment>
<transl>primer</transl>
<mean digits="2" symbol="." />
<transl>ABC</transl>
<usage>pārn</usage>
<transl>the rudiments</transl>
<from_sample>ābeces patiesība</from_sample>
<to_sample>platitude</to_sample>
<to_sample>self-evidence</to_sample>
<to_sample>truism</to_sample><comment>(man)</c
    omment>
<idiom />
<from_sample>tā ir ķīniešu ābece</from_sample>
<to_sample>it is all
    Greek</to_sample><comment>(to
    me)</comment>
</entry>
```

ābece
n **1.** ABC *(book)*, primer; **2.** ABC *pārn* the rudiments; ābeces patiesība - platitude, self-evidence, truism *(man)*; ♦ tā ir ķīniešu ābece - it is all Greek *(to me)*

Figure 2: Sample of dictionary entry in printed dictionary (right) and XML format (left) for the dictionary entry *ābece*.

Each entry is included in <entry> tag. Every entry starts and must have at least one <title> tag that represents the lexical entry. Other possible tags include:

- part of speech and other grammatical information, enclosed by a <gram> tag;
- in bi/multi-lingual dictionaries, there usually are one or several <transl> tags which are used to describe the translation;
- <link> tag, used to point at another related entry;
- <from_sample> tag, enclosing a sample in the source language, and the following <to_sample> tag, enclosing its translation into the target language. In case of a monolingual dictionary, only the <from_sample> tag is used;
- <comment> tag, enclosing additional contextual information that is specific to the entry, its translation, or sample phrase.

Diversity of XML tags helps to preserve the rich content of a dictionary, very close to its original view. When a dictionary entry is presented to a user, the dictionary entry is transformed from XML format to HTML view, and different XML tags are specifically formatted: bold, italic, different font size and different font colour (Figure 3).

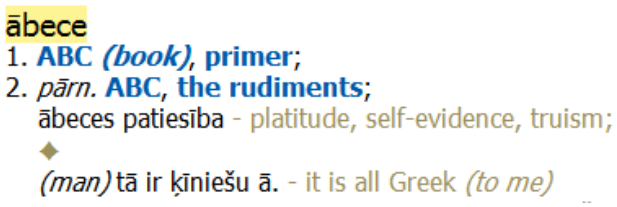


Figure 3: Dictionary entry in the electronic dictionary for the word *ābece*.

Although dictionary entries are merged, a user still has the possibility to search in a single dictionary (or several dictionary sources), as TDB allows all dictionary sources to be seen for each translation direction, or select a particular dictionary (or dictionaries).

2.2 Adding terminology data

In addition to general language dictionaries, terminological data is another type of resource that can be very useful for translation or comprehension of lexical units, particularly if a user is dealing with a text in a specialized domain.

TDB provides two options for integrating terminological data. A terminology resource can be added as an additional local terminology dictionary or accessed as a remote online resource.

Local terminology dictionaries are provided in a similar manner, as lexical dictionaries. Terms are automatically added to the list of all headwords for the source language that is displayed on the left side pane of the Dictionary Browser (excluding

duplicates, in case some similar general language headword is already present in the list). Users can also access a terminological entry using the search feature.

Terminology entries that match the selected headword or a search query are displayed in a separate terminology section on the right side pane of TDB.

Although representation of terminological entries is similar to that of lexical entries, there are important conceptual differences. While in a lexical entry all of the meanings are grouped under one headword, in the case of terminology data, there are separate entries displayed for each term corresponding to the search criteria. This approach is chosen because we follow the concept based principle for the organisation of terminological data. According to this approach, every terminological entry corresponds to one concept. One concept may have several lexical units denoting it, but a single terminology entry may not depict more than one concept.

Figure 4 shows this approach for an example of terminology data found for the search-word *communication*. Several terminology entries are displayed from a number of terminology dictionaries on different subject fields.

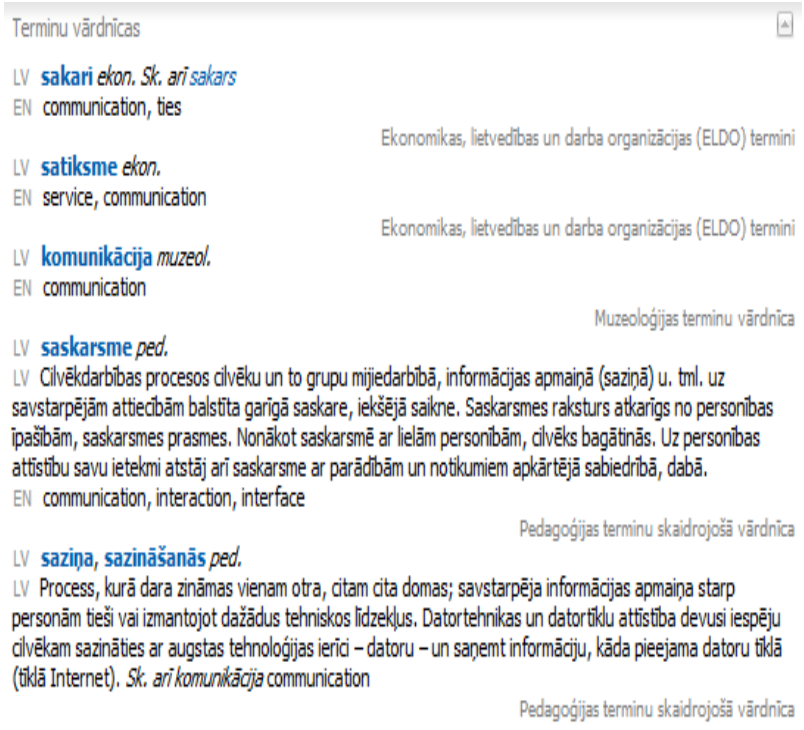


Figure 4: Representation of terminological data from multiple resources and domains

Terminological data of multiple domains can be very voluminous with many specific and rarely used terms. This makes it impractical to provide all of these data locally. Our approach is to limit the data stored on a user's computer only to the most-used domains, such as economics and finance, law, legislation and information technology. Other terminological resources are accessed through dynamic querying of online

sources. This also ensures the up-to-datedness of information, as new terms are being introduced, and some older terms become depreciated or changed.

For TDB, such an external terminology resource is EuroTermBank[1]. It provides free web-based access to the richest collection of European multilingual terminology from a variety of collections and domains (Vasiljevs et al. 2008). Its database currently contains approximately 2.6 million terms from 137 terminology resources in more than 30 languages. EuroTermBank provides not only terms stored in its repository, but also matching terms retrieved from external online terminology databases, such as the database of the Terminology Commission of Latvia[2] and EU inter-institutional terminology database IATE[3].

EuroTermBank provides a common application programming interface (API) to query its data by external systems. This API returns terminology data in the TBX format. TBX (TermBase eXchange) is a standard format for terminology exchange developed by the Terminology Special Interest Group of the recently dissolved Localization Industry Standards Association (LISA). In 2008, this format was adapted by ISO as international standard ISO 30042:2008. Terminological data is organized in data categories that are compliant to ISOcat data category registry as defined in ISO 12620.

TDB queries EuroTermBank for the word searched by the user and processes the received result to represent it in a way similar to that of terminology data from locally stored resources. As online querying of EuroTermBank may take some time depending on the speed of the user's Internet connection, it is optional, and the user can easily switch it on or off.

The terminology entry represented to a user includes such data as the term in the source language, its equivalent in the target language, subject domain, definition (if provided) and the source of data, e.g., information about the terminology resource from which this particular entry originates.

3. Integration of language technologies

While the basic functionality of the electronic dictionary is realized through a common data format and efficient search algorithms, the more advanced and important features are realised through integration of several language technology solutions. For different tasks, TDB uses spelling checker, morphological analyser, text to speech engine, and machine translation services.

[1] http://www.eurotermbank.com

[2] http://termnet.lv

[3] http://iate.europe.eu

3.1 Language technologies that enrich search facilities

The integration of **spelling checker** into TDB plays an important role for users in two cases: (1) for a language with rich diacritics, a spelling checker helps to correct mistakes of forgotten diacritics (see Figure 5), and (2) for users with insufficient knowledge of a language (e.g. a foreign language learner or a child), spelling checker helps to correct errors in words with complicated spelling. In both cases, the task of spelling checker is to help the user find a translation in cases when an incorrect lexical entry is requested.

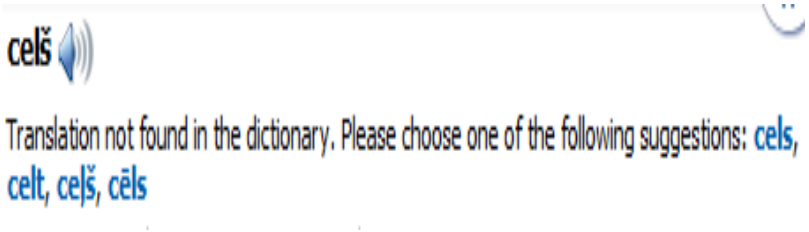


Figure 5: Suggestion from spelling checker for incorrect Latvian word *celš*

More advanced, but similar functionality is provided by the **lemmatizer** and **morphological analysis tools.** These tools allow a user to find translations for forms that differ from the lexical entry. This feature is very useful for highly inflected languages where word form can vary significantly from the base form, as illustrated in Table 1 for the verb *iet* (*to walk).*

	Present	*Past*	*Future*
1st pers. sing.	*eju*	*gāju*	*iešu*
2nd pers. sing	*ej*	*gāji*	*iesi*
3rd pers. sing.	*iet*	*gāja*	*ies*
1st pers. plur.	*ejam*	*gājām*	*iesim*
2nd pers. plur.	*ejat*	*gājāt*	*iesiet*
3rd pers. plur.	*iet*	*gāja*	*ies*

Table 1: Inflected forms for verb *iet* (*to walk)*

The morphological analyser can also play the role of disambiguator in a dictionary. In the case of the entered word form corresponding to several base forms, the morphological analysis tool allows to choose between them and leads to the most appropriate translation (see Figure 6).

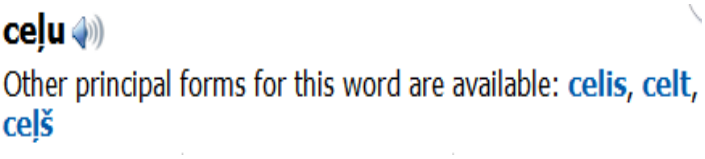


Figure 6: Suggestions of the morphological analyser for word form *ceļu*.

Finally, the morphological analyser is used as a reference tool that allows all inflectional forms of the word to be seen. As mentioned before, this is an important feature for inflected languages with a rich morphology. For instance, in the Latvian language, many palatalised forms occur for nouns. Although palatalisation rules are rather regular, some exceptions exist for each particular case, forming a set of exceptions, words which in many cases are spelled incorrectly even by native speakers.

3.2 Content enrichment through machine translation

The language technologies described above enrich search facilities in dictionary content and help users find a necessary dictionary entry. However, all dictionaries are limited in size and content and no dictionaries contain all possible words for a particular language and their translations. One possibility of how to extend coverage of translation dictionary content is to apply machine translation. Translations suggested by the machine translation system are not always perfect, but in many cases, they provide an added value for the user. Moreover, integration of the machine translation system into the dictionary software allows a user to translate a phrase or sentence with a particular word, thus allowing the user to find its contextual meaning (Figure 7).

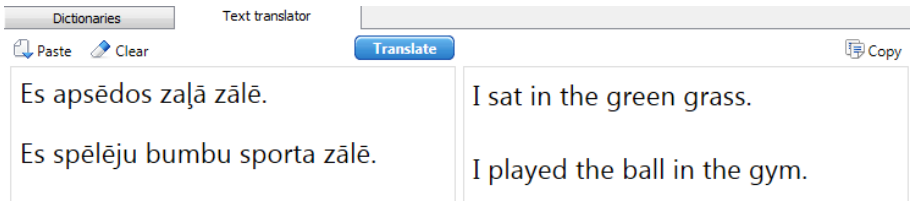


Figure 7: Machine translated samples for word *zāle (grass/hall)*

4. Dictionary content in different media

As there are more and more different devices where dictionaries could be presented, it is important to develop a dictionary browser that is interoperable between different platforms and devices. TDB is implemented not only as a desktop application, but also as a Web dictionary and mobile application. The same data modules are searched to translate a word or phrase upon user request. Only the way in which results are presented differs. The form in which results are presented depends on the size of the device, Internet access and other limitations.

As a desktop application, TDB has no limitation in the presentation of results. If a result does not fit on a visible part of the window, the result window has a scroll bar. The results from main dictionaries, term dictionaries, and synonym dictionaries are on separate foldable panels, which, if opened, show translations of particular types

and while in a folded state, do not take up much space in the result window (see Figure 8).

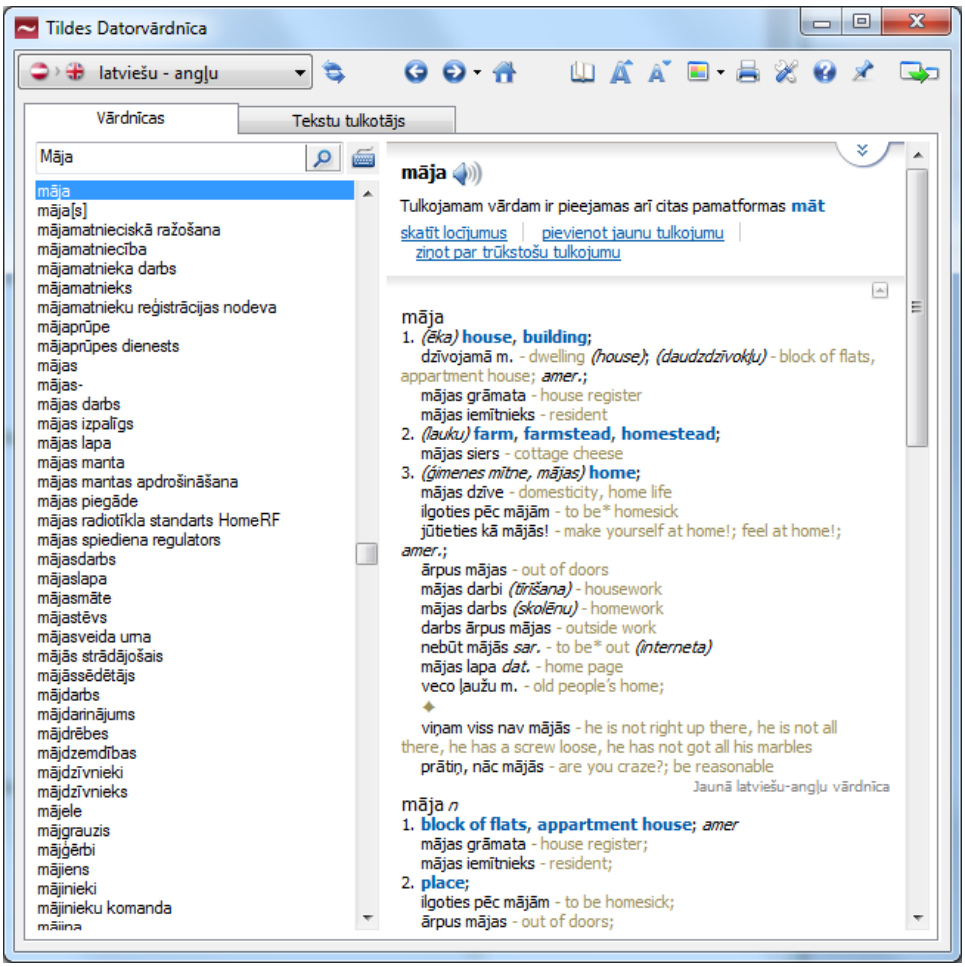


Figure 8: Results for word māja in Tilde Dictionary Browser

In TDB, a user can click on a link and add a new translation to the current entry or send a report to dictionary creators about a missing translation. A user can also switch to the Text translation tab, which allows the user to translate texts with an online Machine Translation service.

All dictionaries available from TDB are also available from the Web portal *letonika.lv* (Figure 9). Here, advanced search options are also available.

In mobile devices, the window for result presentation is much smaller than for a computer screen, and accordingly, less information can be displayed. Therefore we show a limited number of translations from the main dictionary and a limited number of usage samples (see Figure 10).

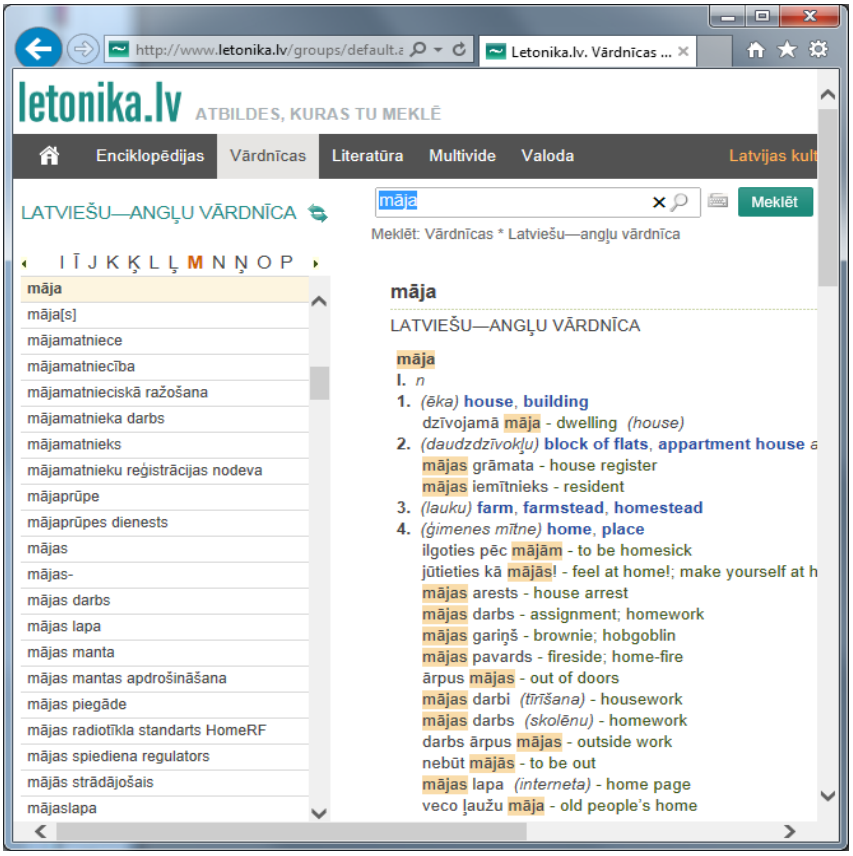


Figure 9: Results for the word *māja* in Web application

Figure 10: Results for word *māja* on a mobile phone

5. Other features to increase applicability

A number of usability features are implemented to facilitate fast and efficient work with TDB. Users can switch between full view and compact view that provides only the essential translation information in a smaller window. Compact view is particularly useful if a user needs to consult a dictionary very often. Then, TDB can stay open as a foreground application (always on top of other open windows) that occupies relatively little space on the screen.

If a user is reading text in a Web browser, text editor, or some other application and needs to quickly find a translation of a particular word, then TDB can be easily accessed by pressing a hot-key combination. In several applications like popular Web browsers and MS Word versions, the translation command is also included in the context menu evocable by the right-click of a mouse.

To facilitate the typing of search words, the keyboard is automatically switched to the target language layout. Special characters can also be typed by using an integrated on-screen keyboard.

A user can also create user dictionaries that can be local or shared throughout an organization. New entries in a user dictionary can be created from the TDB interface or by directly writing into the dictionary file that has a simple to understand text-based format.

Besides phonetic transcription of headword pronunciation, TDB makes it possible to listen to a particular translation, a sample of usage, or even a fragment of text. This feature is enabled through the integration of a **text-to-speech** engine. Currently, TDB integrates Latvian TTS developed by Tilde (Goba and Vasiljevs 2007) and English TTS provided by Microsoft. Microsoft Speech API is used for the TTS integration making it easy to extend language support with other MS SAPI compliant TTS engines.

6. Conclusion and tasks for the future

In this paper, we presented the electronic dictionary software TDB, that, in addition to simple search and browsing, also supports different language technology driven services that facilitate better retrieval of requested entries in non-trivial cases.

TDB can be used on different platforms, including mobile devices and the Web. Currently, 20 language pairs are supported for general content dictionaries. However, more language pairs can be easily incorporated, and additional dictionaries for current language pairs can be added.

Development of a user-friendly dictionary is a never-ending process. Our development plans include two directions: extension in content and extension in

functionality.

With respect to functionality, two major extensions are planned. Firstly, we plan to support specialists and language learners with extended context for a selected lexical item by providing concordances from corpora. Secondly, closer integration with machine translation is planned, thus allowing users to translate a full document instead of a phrase, sentence, or small fragment of text.

7. Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center " of EU Structural funds, contract nr. L-KC-11-0003, signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.8 "Research of automatic methods for text structural analysis".

8. References

Bogaards, P. (2003). Uses and users of dictionaries. In van Sterkenburg, Piet (ed.), *A practical guide to lexicography, Terminology and Lexicography Research and Practice 6*, pp. 26-33. Amsterdam: John Benjamins.

Burke, S. M. (1998). *The Design of Online Lexicons*. Master's thesis: Northwestern University, Evanston, IL.

de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. In *International Journal of Lexicography*, 16(2), pp. 143-199.

Deksne, D., Skadiņa, I., Skadiņš, R., Vasiļjevs, A. (2005). Foreign Language Reading Tool – First Step Towards English-Latvian Commercial Machine Translation. In *Proceedings of Second Baltic Conference „Human Language Technologies – the Baltic Perspective"*, Tallinn, 2005.

Goba, K., Vasiljevs, A. (2007). Development of Text-To-Speech System for Latvian. In Joakim Nivre, H.-J. Kaalep, K. Muischnek, & M. Koit (Eds.), In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, Tartu, pp. 67–72.

Ide, K. (1993). A Catalogue of Electronic Dictionaries. *Language* 22.5, pp. 42-49.

Koren, S. (1997). Quality versus convenience: comparison of modern dictionaries from the researcher's, teacher's and learner's points of view. In TESL-EJ 2 (3).

Lehr, A. (1996). Electronic Dictionaries. In *Lexicographica* 12, pp. 310-17.

Lew, R. (2004). Which dictionary for whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English. Poznan: Motivex.

Měchura, M. B. (2008). Giving them what they want: search strategies for electronic dictionaries. In *Proceedings of the 13th Euralex International Congress*, pp.

1295-1299.

Müller-Spitzer, C. (2011). Textual Structures in Electronic Dictionaries compared with Printed Dictionaries. A Short General Survey. In: Gouws, Rufus H./Heid, Ulrich/Schweickhard, Wolfgang/Wiegand, Herbert Ernst (Hgg.): *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter.

Müller-Spitzer, C., Koplenig, A., Töpel, A. (2011). What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. In *Proceedings of eLex 2011, pp. 203-2008*.

Nesi, H. and Haill, R. (2002). A Study of Dictionary Use by International Students at a British University. In *International Journal of Lexicography*, 15.4: 277-305.

Oppentocht, L. and Schutz, R. (2003). Developments in electronic dictionary design. In van Sterkenburg, Piet (ed.), *A practical guide to lexicography, Terminology and Lexicography Research and Practice* 6, 215-227. Amsterdam: John Benjamins.

Sharpe, P. (1995). 'Electronic Dictionaries with Particular Reference to the Design of an Electronic Bilingual Dictionary for English-speaking Learners of Japanese. *International Journal of Lexicography* 8.1, pp. 39-54.

Skadiņa, I., Vasiļjevs, A., Deksne, D., Skadiņš, R., Goldberga, L. (2007). Comprehension Assistant for Languages of Baltic States. In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, Tartu, pp. 167-174.

Vasiljevs, A., Rirdance, S., & Liedskalnins, A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources* (ICGL 2008), Hong Kong, pp. 213–220.