The Dictionary of Lithuanian (LKŽ) and its Future in Databases and Electronic Versions

Elena Jolanta Zabarskaitė and Gertrūda Naktinienė
Institute of the Lithuanian Language, Vilnius, Lithuania

The paper deals with the Dictionary of the Lithuanian Language (Vol. 1-20, 1941–2002): electronic release, 2005 (renewed version 2008) and its new version on the CD. A three-level lexical database: exhaustive for academic purposes, medium for the broad public, and more narrow for schools, is being created at the Institute of the Lithuanian Language. Its core consists of an electronic version of the Dictionary of the Lithuanian Language (about 0.5 million dictionary entries) and its card index (about 5 million cards), which is in the process of being computerized.

1. Introduction

The final 20th volume of the Dictionary of Lithuanian (*Lietuvių kalbos žodynas,* further LKŽ) was released in 2002, and the biggest work in Lithuanian linguistics of the 20th c. that several generations of linguists worked on was completed. LKŽ was compiled using the paper card index of 4.5 mn words that dates back to 1902. The initiator was Kazimieras Būga, professor of the Universities of Saint Petersburg, Perm, Tomsk (Russia), and afterwards Kaunas (Lithuania).

The history of LKŽ coincides with the complicated history of the State of Lithuania. LKŽ started writing in 1930. Volume I of the dictionary was released in 1941 and Volume II in 1947. The second volume came out at the beginning of the second Soviet occupation and was destined for the Soviet censorship. The Soviet authorities made demands for the introduction of examples in dictionary entries as illustrations from Soviet literature, Lenin works, documents of the Communist Party. Only after these, words could be illustrated by sentences of living dialects, old writings, and folklore. With the coming of 'Perestroika', ideological illustrative examples were discarded.

The dictionary was completed after Lithuania regained its independence. Completion of the dictionary was celebrated all over Lithuania.

Its significance to researches of Baltic linguistics, Indo-European linguistics, Baltic mythology and culture is acknowledged throughout the world (for further details see Schmalstieg 1996; Kažukauskaitė 2002; Топоров 2004).

2. The Dictionary of Lithuanian (LKŽ)

It is the biggest work in Lithuanian linguistics. It is a mixed-type dictionary including both lexis of writings and living language (dialects). Lexis of writings comprises the period between 1547-2001, i.e. from the release of the first Lithuanian book[1] to the time of dictionary completion. Lexis from living dialects includes words dating back to 1902 and up to 2001. The lexis of dialects is transposed into the standard language according to phonetic laws.

[1] The oldest known printed Lithuanian book is the catechism by Martynas Mažvydas published in Königsberg (Karaliaučius, East Prussia, now Kaliningrad, Russia) in 1547. The catechism also includes the first Lithuanian ABC-book that introduces the first Lithuanian alphabet.

The Dictionary makes up 22 thousand pages and comprises half a million lexicographic entries and over 11 million words of text. It presents the origin, history, and distribution of the word together with its accentuation, grammatical forms, categories, and its peculiarities with respect to word formation, semantic structure, stylistic usage, and etc. Lexicographic entries are illustrated with sentences quoted from religious, scientific, political, fictional, and journalistic literature, and the material of dialects example of folklore. Lots of proverbs, riddles, figurative phrases and sayings and examples of phraseology are presented. It serves as a source of national identity and linguistic investigation of the development of written and spoken Lithuanian, and addresses the world community working in the humanities and the society at large.

3. Card Index of LKŽ

The Institute of the Lithuanian Language houses the largest paper card index of lexical items in Lithuania that is continually supplemented. It consists of two parts, the *Main* card index of 4.5 mn lexical items with usage examples and card index of *Supplements* of 0.5 mn lexical items with usage examples that were collected after the release of the corresponding volumes of the dictionary for the preparation of the new dictionary releases. The card index is collected from almost 1,000 lexicographic sources (both hand-written and printed); dialectal words have been recorded in more than 500 Lithuanian settlements. The card index was compiled by many people of different education background therefore it is heterogeneous. Nearly all cards are hand-written. The digitisation of the card index of *Supplements* started in 1999 using 'CardScan' software created specifically for this purpose. The catalogue of keyboard-typed headwords can be created and images of scanned authentic cards can be saved. There is an option to launch a search of the scanned card images by the headword:

Figure 1. Database of the card index of *Supplements*

The work of creating the database of the *Main* card index began by describing it, i.e. by entering it into the database together with the image of a card. The description scheme is much the same as that of LKŽ headword data with additional parameters. Illustrations are scanned, for reselecting different images of the card index is inexpedient. Additional descriptive parameters are comments on degrees of academic value, presenter's last name, year of presentation (this enormously increases the value):

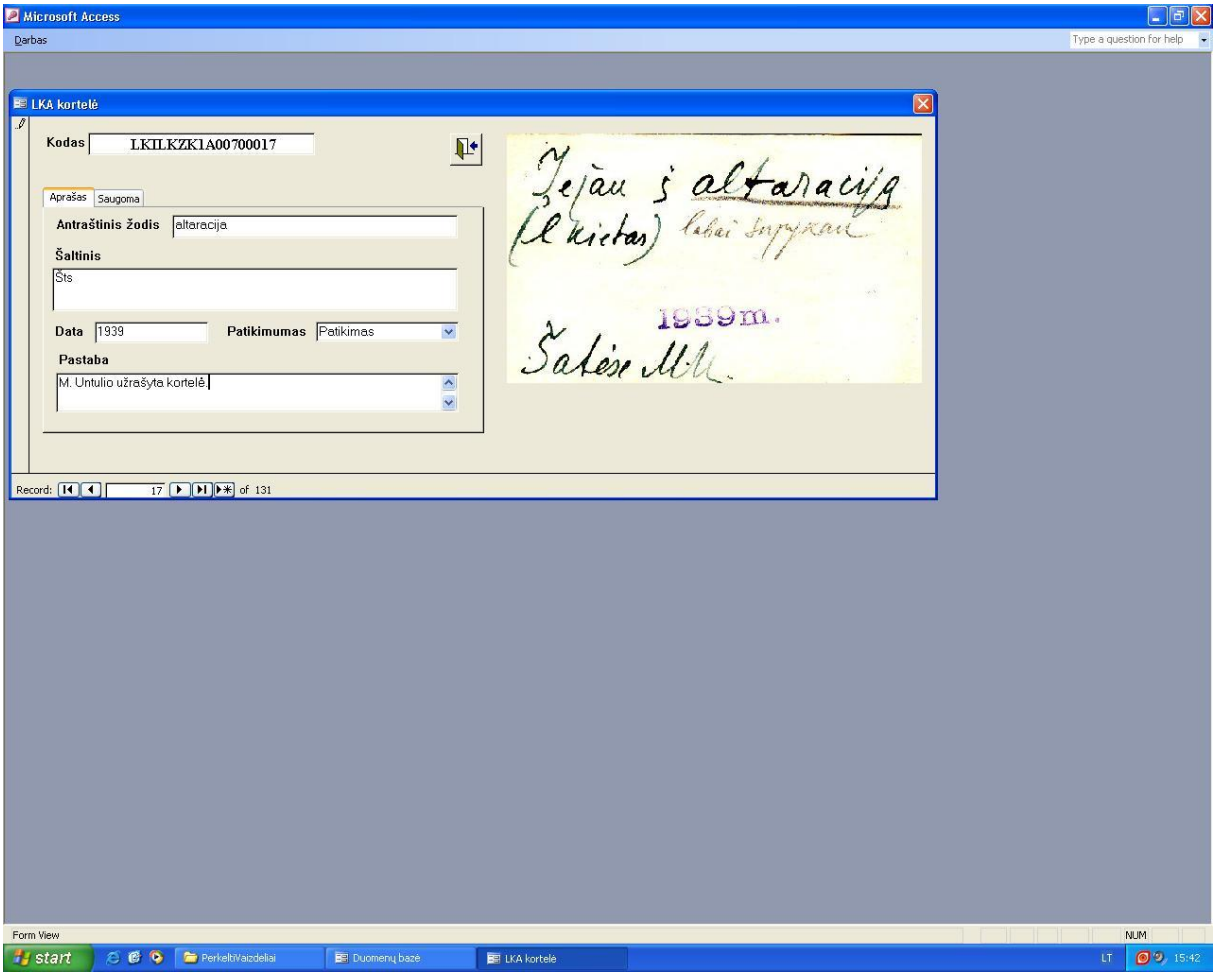

Figure 2. Database of the *Main* card index.

4. Concept of LKŽ Digitizing

Together with IT specialists, lexicographers of the Institute of the Lithuanian Language created the concept of LKŽ digitizing:

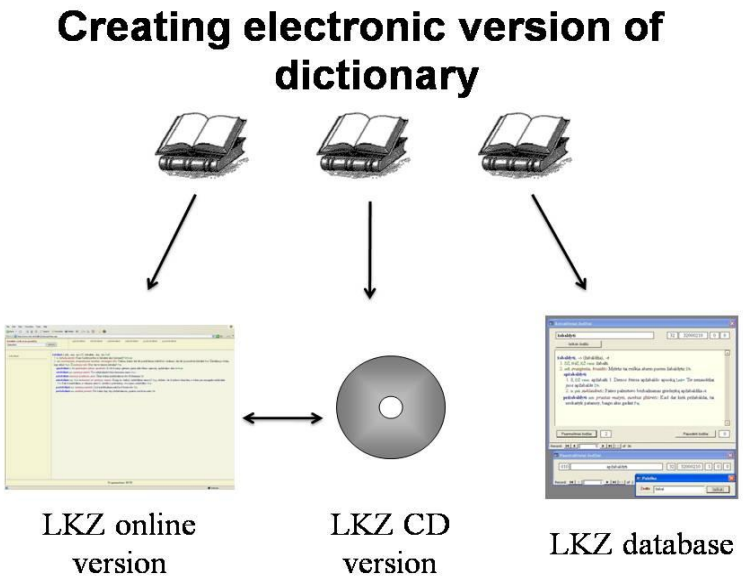

Figure 3. Concept of Dictionary digitizing.

5. Electronic version of the Dictionary of Lithuanian. Object, source, and instrument of science

5.1. Issues of adapting LKŽ text to digital environment

The first electronic version of LKŽ with a search engine of a headword was released in 2005. However, it was decided to present LKŽ to the public not only as a source of language history but as an active instrument of language and linguistics cognition. In the electronic version, language facts of the dialectal and old writings were amended based on the results of the latest linguistic researches. Entries of mistakenly transcribed words (the so called 'words absentees') were deleted and their examples were moved to the entries of the existing words (for example, entry of the 'word absentee' *džirbti* 'dial. to work' was deleted and its illustrative sentences of the South Dialect (*Dzūkai*) were moved to the entry of the word *dirbti* 'standart l. to work'.). Semantic differences, inequalities in the use of the main forms as well as accentuation errors were corrected.

As mentioned before, the Soviet authorities demanded of the dictionary edition to better portray the 'Soviet reality,' therefore the text of the dictionary was abundantly supplemented with various language examples from the Soviet writings. The Soviet authorities also required to explain certain word meanings on the basis of communistic ideology.

Once adaptation of LKŽ texts to the electronic version began, it appeared that life of the Soviet reality is no longer reflected after a certain part of inadequate illustrative examples abundantly used in the previous volumes of LKŽ is deleted or edited out. A new abbreviation *sov.* (Sovietism) was additionally introduced considering that reflection of the Soviet reality and ideology in a dictionary might be a research object of linguistics, history, cultural studies, and other humanities.

5.2. Technical Challenges of LKŽ Digitizing

With the support of the State Commission of the Lithuanian Language the entries of the Dictionary were digitized in 4 years (1999-2002). It took three years to edit, adapt and release the electronic version of the Dictionary with the financial support of the Government of the Republic of Lithuania (2002-2005).

IT specialists together with lexicographers had to face the issues of sophisticated Dictionary structure, variety of characters and symbols, and other problems (for information on technical solutions for creation of electronic version further refer to Zinkevičius (2004)).

The Dictionary is to be combined with other Lithuanian language resources. For example, in 2008 it was combined with the database of the Lithuanian Language Dialect Areal Atlas:

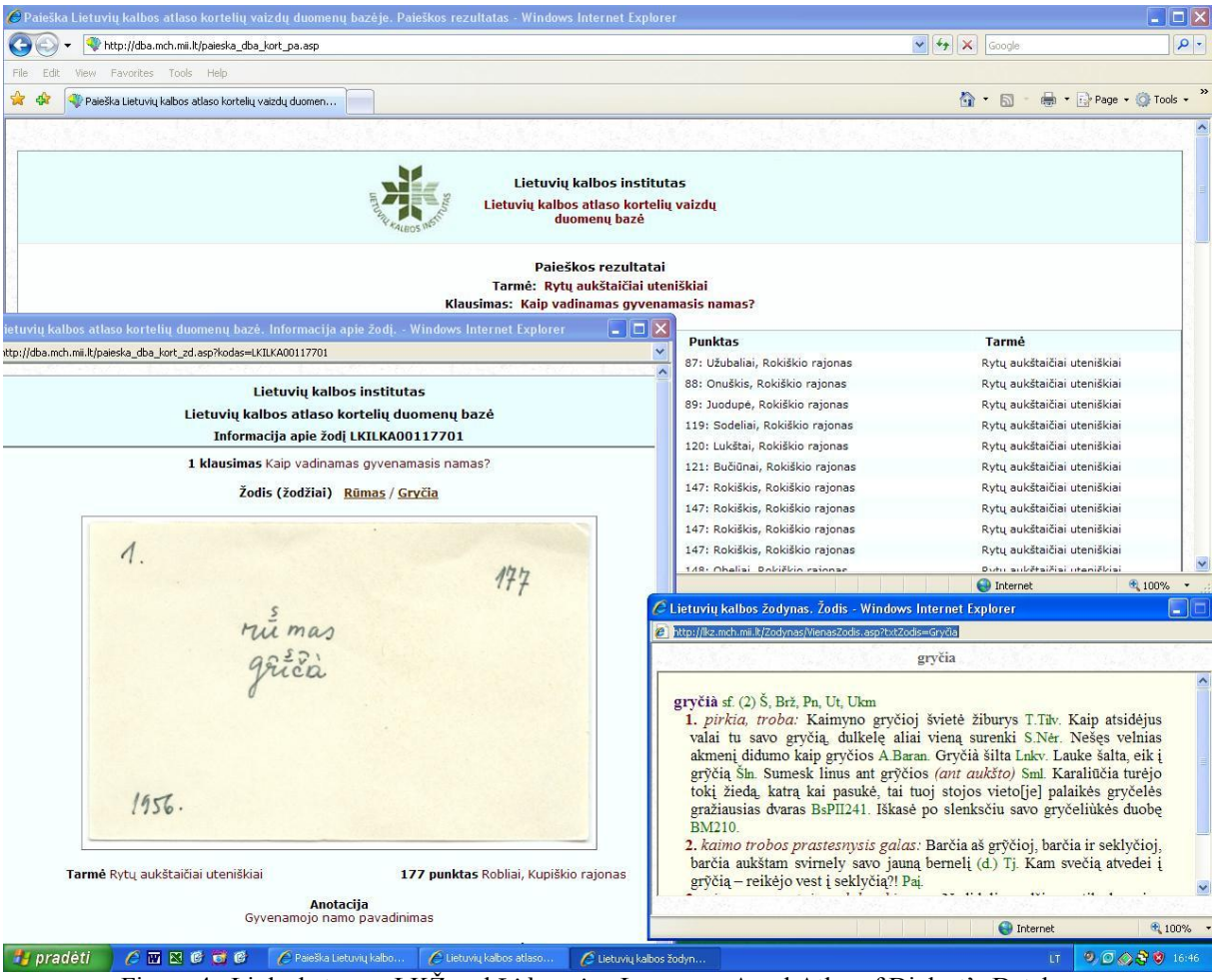

Figure 4. Links between LKŽ and Lithuanian Language Areal Atlas of Dialect's Database.

5.3. Website of LKŽ

For the convenience of the users there is a website of the Dictionary at www.lkz.lt (Administrator Gertrūda Naktinienė). The online Dictionary is constantly upgraded taking into consideration, to the extent possible, various abundant requests from the users (for further information on communication of lexicographers and users further refer to Zabarskaitė, Naktinienė (2007)). Responses from the online users show relatively high interest among the society in this freely accessible scholarly dictionary. According to the statistics, the Dictionary is accessed approximately 1200 times a day and about 10,000 entries a day are reviewed.

6. CD of LKŽ

In 2010 a second edition of LKŽe(2) was released in a CD prepared on the basis of LKŽe online version. A new Lithuanian Unicode font *Palemonas* was adapted to this version of the

Dictionary specifically intended to show Lithuanian stressed letters and other specific symbols that are necessary for complicated philological works such as LKŽe. More search options are available. By user's selection, all abbreviations are decoded:

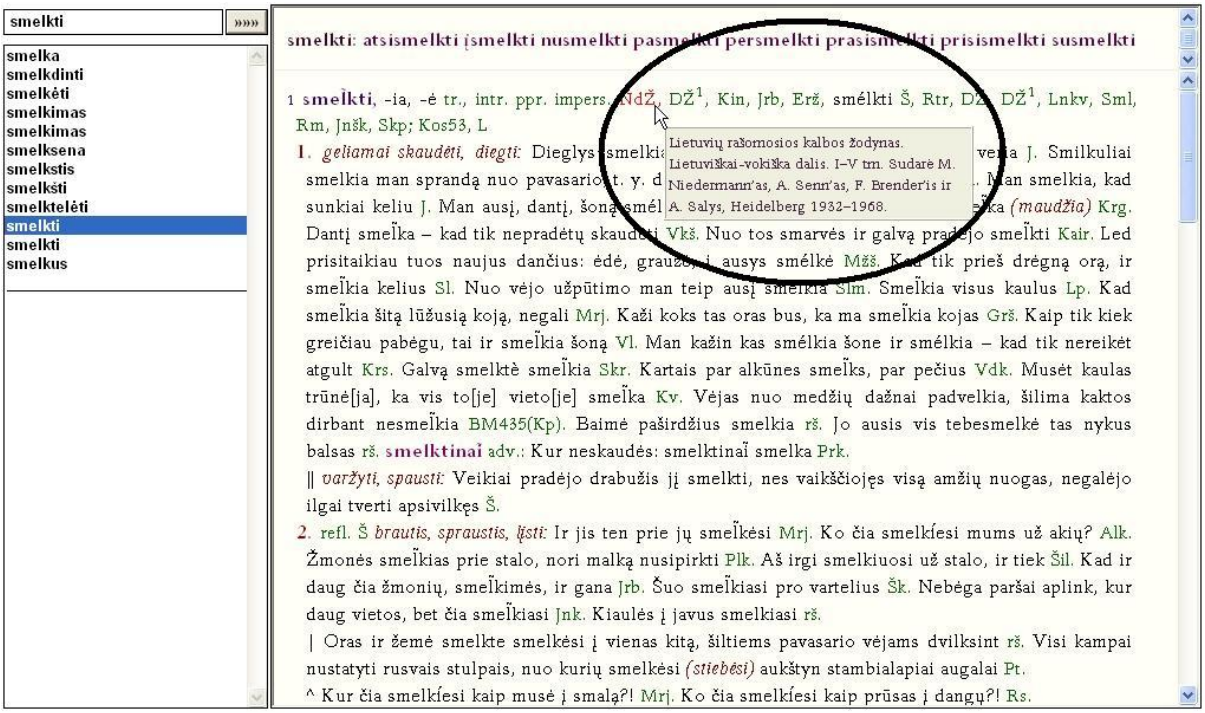

Figure 5. LKŽe(2).

7. Database of LKŽ

Further work on the Dictionary is related to the development of the database (Naktinienė, Zabarskaitė (2004); Zabarskaitė (2007)). A detailed search system of many boxes is under construction.

Tools enabling database search by abbreviations of locations, sources, and last name of the authors, by accentuations, by grammatical, stylistic, usage fields and other references are also intended to be created.

This database also has to serve as a tool for the preparation of amended and appended versions of the Dictionary. The first task is to add data of the card index of the *Supplements* to the new version of the Dictionary. Implementation of this task will start once software, the so called 'LKŽ lexicographer's workplace,' is created.

It is also planned to combine the text of LKŽ and its *Main* card index. Dictionary users would then be able to access data that is in the card of the card index only, phonetic transcription of dialect sentences, date of the record, names of the presenter and recorder, handwriting, names of small villages, and etc.

The formation of the LKŽ database will facilitate dictionary compilation, lexicographical research, machine translation, and etc.; it will also contribute to the preservation of the Lithuanian linguistic heritage under the conditions of global integration (Zabarskaitė,

Naktinienė, Šepetytė (2005)). Thesaurus of English key words will be created for the convenience of foreign scholars. The other great responsibility of the Institute of the Lithuanian Language is the placement of the Lithuanian databases on the Internet and its spread in the academic sphere of the European Union.

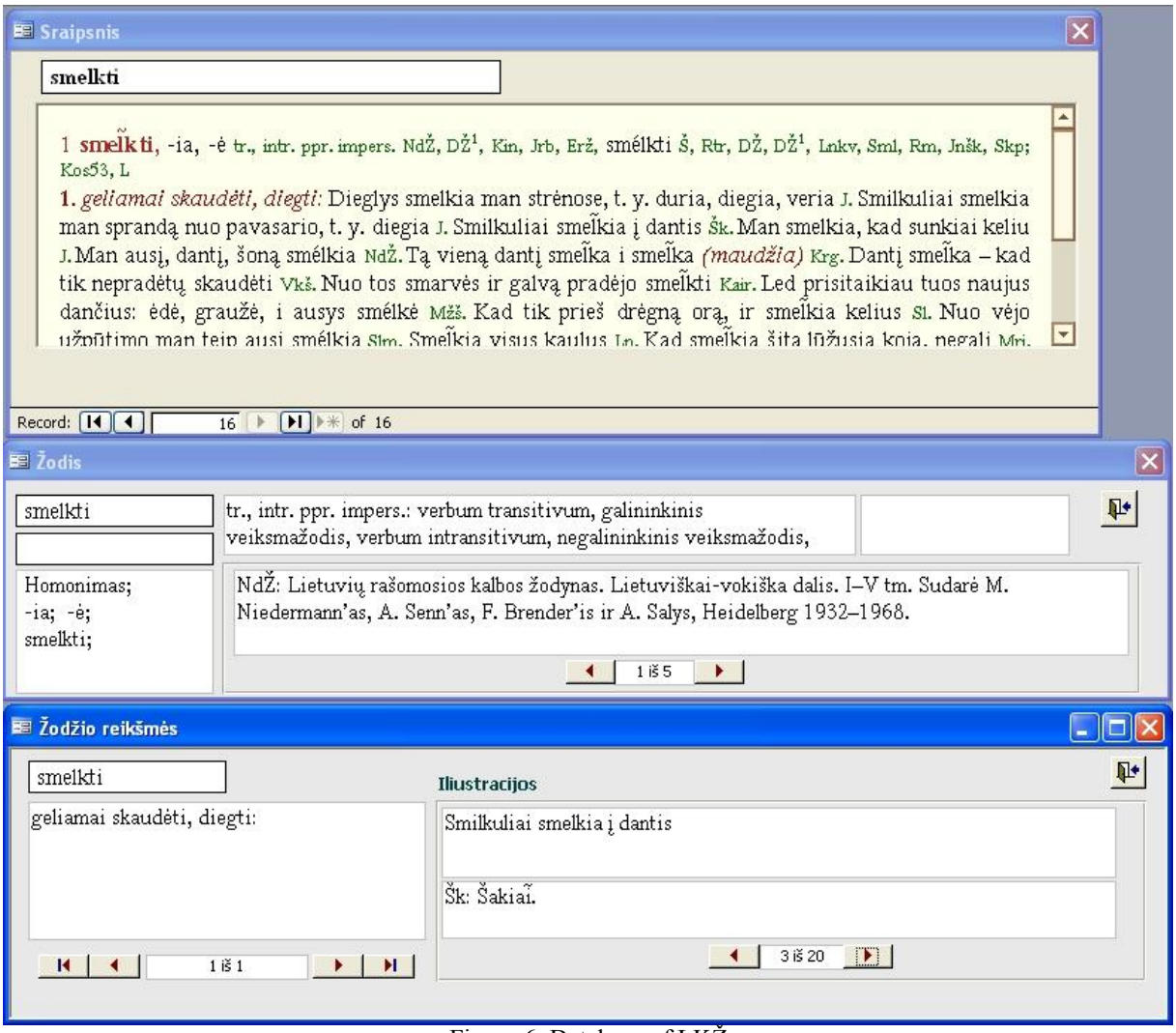

Figure 6. Database of LKŽ

References

Dictionaries

LKŽe – *Lietuvių kalbos žodynas* (t. I–XX, 1941–2002): elektroninis variantas / G. Naktinienė (vyr. red.), J. Paulauskas, R. Petrokienė, V. Vitkauskas, J. Zabarskaitė. Programuotojai: E. Ožeraitis, V. Zinkevičius. – Vilnius: Lietuvių kalbos institutas, 2005. – Atnaujinta versija 2008. – www.lkz.lt. [*The Dictionary of the Lithuanian Language* (Vol. 1-20, 1941–2002): electronic release [online], 2005. – Renewed version, 2008.]

LKŽe(2) – *Lietuvių kalbos žodynas* (t. I–XX, 1941–2002): elektroninis variantas. 2-as leidimas [kompaktinė plokštelė] / G. Naktinienė (vyr. red.), J. Paulauskas, R. Petrokienė, V. Vitkauskas, J. Zabarskaitė. Programuotojai: E. Ožeraitis, V. Zinkevičius. – Vilnius: Lietuvių kalbos institutas, 2010. [*The Dictionary of the Lithuanian Language* (Vol. 1-20, 1941–2002): electronic release [CD], 2010.]

Other References

Kažukauskaitė, O. (2002). 'Le grand dictionnaire d'une petite nation, une histoire de cent ans'. In *Cahiers lituaniens* 3. Strasbourg. 29–33.

Naktinienė, G. and Zabarskaitė, J. (2004). 'On the Linguistic Databases of the Institute of the Lithuanian Language'. In *The First Baltic Conference 'Human Language Technologies: the Baltic Perspective'*. Riga. 187–190.

Schmalstieg, W. R. (1996). 'Some Comments on new Volumes of the Lithuanian Academy Dictionary'. In *Lituanus* 42 (1). 18–23.

Топоров, В. Н. (2004). *К выходу в свет большого „Словаря литовского языка'. Балто-славянские исследования XVI*. Москва. 408–415.

Zabarskaitė, J. (2007). 'Prameny a data Institutu pro litevsky jazyk'. In *Europeica–Slavica–Baltica, Praha: Narodni knihova ČR*. 261–272.

Zabarskaitė, J., Naktinienė, G. (2007). 'Слоўнікі ў Інтэрнеце: дыялог мовазнаўцы і носьбіта мовы'. In *Слово и словарь. Vocabulum et vocabularium. Сборник научных трудов по лексикографии* / Отв. ред.: Л. В. Рычкова, В. Л. Воронович. Рецензенты: А. В. Никитевич, И. С. Лисовская. Гродно. 2007. 32–34.

Zabarskaitė, J., Naktinienė, G., Šepetytė, R. (2005). 'Современность и перспективы большого „Словаря литовского языка'' . In *Исторический путь литовской письменности. Сборник материалов конференции* / редколлегия сборника: Ю. Будрайтис, Й. Забарскайте, М. В. Завьялова, Ю. А. Лабынцев, Е. Л. Назарова, Л. Л. Щавинская; научный редактор С. Ю. Темчин. Вильнюс: Институт литовского языка, 2005. 340–355.

Zinkevičius ,V. (2004). 'Creating the Electronic Version of the Dictionary of Lithuanian'. In *The First Baltic Conference. Human Language Technologies. The Baltic Perspective*. Riga. 170-173.