

Two Decades of Baltian-Latgalian Electronic Dictionary

Ansis Ataols Bērziņš
Riga Technical College
Riga, Latvia
ansis@latnet.lv

Abstract — We would like to present the dictionary we have been compiling for two decades, as well as the electronic lexicographic system used to create it, which was specifically developed for this purpose twenty-three years ago. This is a good example of an almost-without-financing, long-term strategy for developing IT systems for language technology purposes, and it could be successfully used to create dictionaries for other endangered languages as well.

Keywords — dictionary, electronic, lexicography, language technology, endangered, less-resourced, Latvian, Latgalian, Baltian.

In 2003, the author created a new electronic system for the Baltic Latvian – Latgalian Latvian dictionary, which he began compiling in the late 1990s together with Armands Kociņš.

In slang, Baltic Latvian is simply called Latvian, and Latgalian Latvian is simply called Latgalian. However, this is not entirely correct, as both nations identified themselves as Latvians („latvieši“ in Baltic Latvian and „latviši“ in Latgalian Latvian). Therefore, we suggest using the terms Baltian and Latgalian instead when shorter forms are needed.

In 2004, the dictionary had 2,000 entries, in 2006 – 8,000, in 2011 – 14,000, in 2017 – 24,000, and now there are almost 28,000. In 2010, a public interface was created, and all these years it has been widely used by people interested in the Latgalian language. This entire job was done without any luck in financial support; we had just some small projects with budgets of a couple of hundred lats (Ls, LVL). The architecture of the dictionary from the very beginning was innovative – it was not just a translation dictionary, but a kind of integrated, universal dictionary that included information for the simultaneous development of translation, explanatory, terminological, dialectal, and other types of dictionaries, with the

possibility of their separate use afterwards. [7] In traditional „paper“ lexicography, the dream of a universal dictionary has been known at least for many decades.

Для теории лексикографии необходимо понятие максимальной лексической системы и понятие универсального словаря, в котором зафиксирована максимальная лексическая система.¹ [8]

It is recognised as merely a theoretical, idealistic and unrealisable construction. Objectively, due to the limitations of paper format – it is impossible to include all the information and maintain the readability of the text.

Универсальный словарь должен существовать в качестве лексикографического идеала в умах составителей словарей. Это цель, достижимая лишь в абстракции.² [4]

However, there are successful partial realisations of this concept, for example, The Big Chinese-Russian Dictionary:

БКРС занимает особое место среди двуязычных китайско-русских словарей. Возможно даже, что среди двуязычных словарей он вообще не имеет аналогов. Как лингвистический словарь он вообрал в себя различные аспекты китайского языка, сведения о которых обычно выносятся в отдельные словари — грамматические, толковые, фразеологические, диалектные, исторические, этимологические. Вместе с тем он расширил рамки лингвистического словаря до энциклопедического, став энциклопедией культуры, общественных отношений, духовного облика Китая и даже местного колорита.³ [9]

Electronic dictionaries are not limited in their filtering capabilities, which is the main problem for paper dictionaries when dealing with large volumes of information, especially across different types of it. Thus, it is possible to collect and include all information for all dictionary types covered by the „umbrella“ of universality,

¹ For the theory of lexicography, the concept of a maximal lexical system and the concept of a universal dictionary in which the maximal lexical system is fixed are necessary.

² A universal dictionary must exist as a lexicographic ideal in the minds of dictionary compilers. This is a goal achievable only in the abstract.

³ The BCRS occupies a special place among bilingual Chinese-Russian dictionaries. It may even be unique among all bilingual dictionaries. As a linguistic

dictionary, it encompasses various aspects of the Chinese language, information about which is usually confined to separate dictionaries – grammar, explanatory, phraseological, dialectal, historical, etymological. At the same time, it expanded the scope of the linguistic dictionary to become an encyclopaedic one, becoming an encyclopaedia of culture, social relations, the spiritual makeup of China, and even local flavor.

but display only partial information in a filtered format for a lexicographic subtype that best suits users' needs.

Now that our electronic dictionary has reached the required level and volume, it is time to edit and publish the

paper version too – this is the final step we plan to take using our electronic system in order to, we hope, preserve this edition of the dictionary for centuries.

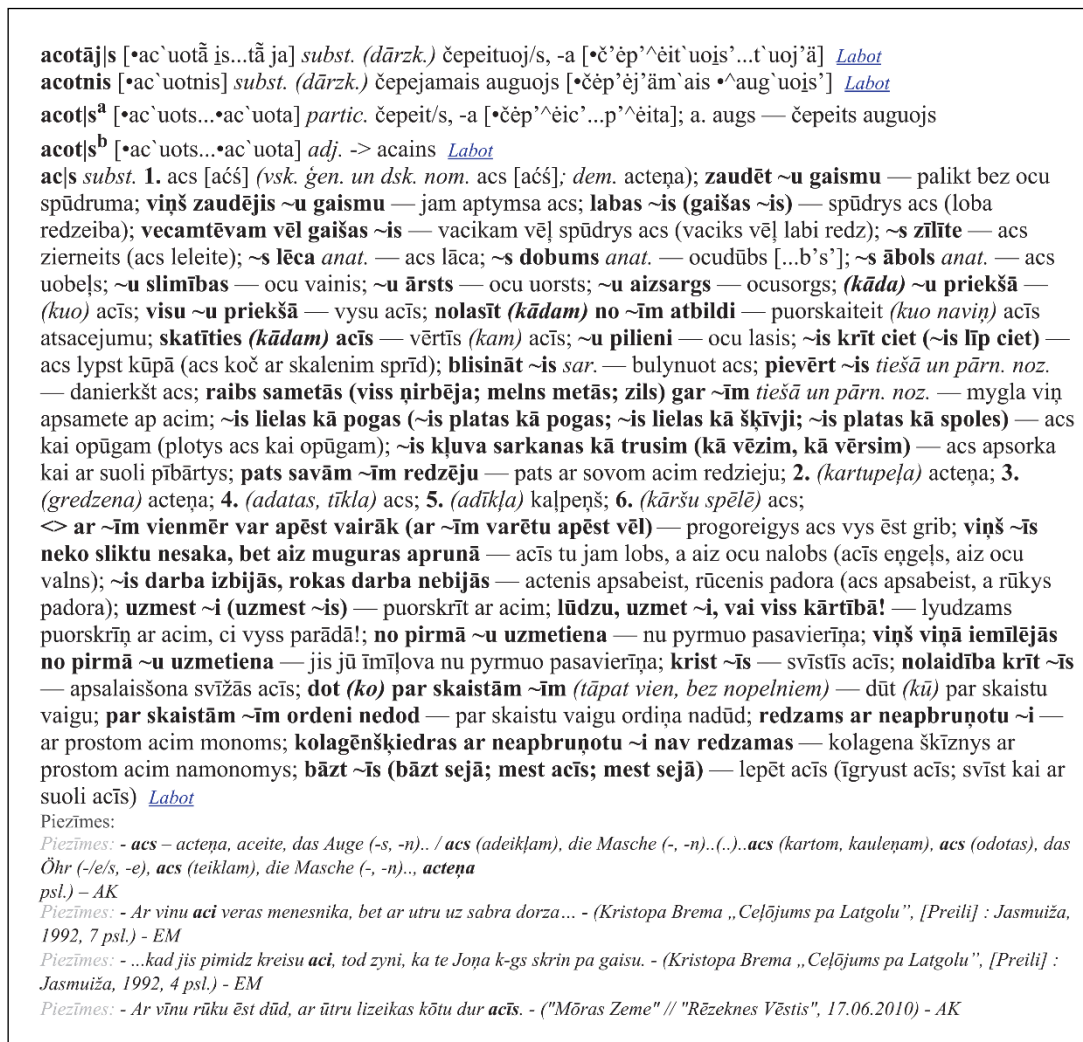


Fig. 1. An example of „as-paper-dictionary“ printout of the dictionary.

Electronic development of dictionaries opens up new possibilities for developers and makes non-topical classical lexicographic perceptions.

However, it adds risks of inappropriate quality too:

Очертания современной лексикографии существенно изменились в связи с внедрением компьютерных технологий: с одной стороны, появились реальные возможности осуществления в обозримые сроки масштабных лексикографических проектов, с другой — расширилось пространство для деятельности компиляторов.⁴ [5]

In our dictionary and development system we strive to keep all the values of paper dictionaries, and just to add advantages of electronic ones.

While dictionaries were developed „on paper“, the main question was „what to include and what not to include?“. The balance between wishes and possibilities, between interestingness and perceptibility of information, between knowledge and space was the main lexicographical challenge.

⁴ The contours of modern lexicography have changed significantly due to the introduction of computer technology: on the one hand, real opportunities have emerged for the implementation of large-scale lexicographic projects within a

foreseeable timeframe; on the other, the scope for the work of compilers has expanded.

Vārds: **ēna** (atrasts)

☒ Homonīms vai homogrāfs: Atdalītājs aiz burta Nr.: Saīsinājumi:

Nozīmju skaits: Izruna: Defise: Paskaidrojums: S.dz.sāk.: S.dz.gal.: S.dz.izr.:

1.

☒ Skatīt vārdu, homogrāfu, nozīmi: Saīsinājumi: Gram. forma:

Mainīt nozīmes numuru: Tulkojumu skaits: Piemēru skaits: Paskaidrojums: Konteksts:

☒ Tulkojums: Saīsinājumi: Locīšana: Defise: S.dz.sāk.: S.dz.gal.: S.dz.izr.:

Paskaidrojums pirms: Paskaidrojums aiz: Izruna:

☒ Tulkojums: Saīsinājumi: Locīšana: Defise: S.dz.sāk.: S.dz.gal.: S.dz.izr.:

Paskaidrojums pirms: Paskaidrojums aiz: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

2.

☒ Skatīt vārdu, homogrāfu, nozīmi: Saīsinājumi: Gram. forma:

Mainīt nozīmes numuru: Tulkojumu skaits: Piemēru skaits: Paskaidrojums: Konteksts:

☒ Tulkojums: Saīsinājumi: Locīšana: Defise: S.dz.sāk.: S.dz.gal.: S.dz.izr.:

Paskaidrojums pirms: Paskaidrojums aiz: Izruna:

☒ Tulkojums: Saīsinājumi: Locīšana: Defise: S.dz.sāk.: S.dz.gal.: S.dz.izr.:

Paskaidrojums pirms: Paskaidrojums aiz: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

☒ Piemērs pirms: Piemērs aiz: Saīsinājums: Paskaidrojumi: Izruna:

Piezīmes:

- ..pasargos nakti susatiwi (ciēn) nu kō nibiēš.., tyūlēn godoj un rur
 - Susatiwi parodiejot Sowu, kurs nikni iznika. - ("Inflantuziemies t:
 - Kae rīcās nēraukēt) (Susatiwe) - (tairu meikla) - ("Tairas dīcēm t:
 < >

Nosūtīt No jauna Homonīmu/homogrāfu skaits: Stāvoklis: Pārjaunot veidlapu

ēn|a subst. 1. (tumšs atspīdums) susātivs; susāts poēt.; **koka ē.** — kūka susātivs; **cilvēka ē.** — cylvāka susātivs; **mest ~u t. un pār.** — mest susātivi; **gulēt ~ā** — gulēt susātivi;
 < **palikt (kāda) ~ā** — palikt (kuo) susātivi; **baidīties no savas ~as** — beifis sova susātiva; **acu ~as** — ocu susātivi; **~u valstība** — susātivu kieneste
 2. (vieta, kur var patverties no saules) pakrieslis; krieslis poēt.; **koka ē.** — kūka pakrieslis; **no saules patverties ~ā** — nu saulis globuotīs pakrieslī; **vairums augu ~ā neaug** — lelums augu pakrieslī naaug

Fig. 2. An example of the interface for editing one entry of the dictionary..

Electronic systems give us the ability to filter in real time. This means we can collect all available information without any disturbance of further work in case of forward-looking design of the dictionary structure and the architecture of the development software too. This fundamentally changes the basic principles of dictionary creation and deserves extensive discussion and analysis. Moreover, this means that the main classification criteria used for „paper“ dictionaries become secondary when creating a „universal“ electronic dictionary: all of them can be used as filtering criteria (instead of creating criteria): the density of entries can cover both general and narrow, as well as special lexis (regional and social dialects, jargons and slangs, archaisms), monolingual dictionary can be combined with a bilingual (or even multilingual) one, lexical data – with encyclopaedic data, diachronic (dynamic) data – with synchronic (static), normative – with referential, data for general use – with specialised. Arrangements can be defined freely depending on users' needs, including alphabetical and semantic ones. Simultaneous development of translational, explanatory, terminological, dialectal, phraseological and other types of dictionaries with a possibility of separate usage of them afterwards from one side makes the process more rational, as it allows all the information to be stored in the right place at the same time. On the other hand, it makes the system more cumbersome, and the design process significantly more complex. How can such a „universal“ dictionary and its development system be created? What should be considered? Certainly, by universality we mean not only the type of dictionary content, but also its technical meaning: a single dictionary should meet the needs of both modern electronic (both offline and online) dictionaries and traditional „paper“ dictionaries—it should be capable of automated printing and proper publication. We felt the idea of universality and integrity of dictionary from the very beginning of our work on it, however it was formalised and formulated just in [7] – what it should contain, what should be included, what kinds of classic „paper“ dictionaries it should cover etc. Of course, there is nothing absolute in this world, and universality does not mean that it covers all possible dictionary variations, but only a fairly large part of them. For now, if we look at Shcherba's list of lexicographical contradictions – he determined the impossibility of combining the terms of both contradictions in a single dictionary – we will see that we have combined 5 of the 6 pairs (with the exception of the distinction between the standard and the ideological; however, if necessary, we can also add an option of ideological structuring too).

Одним из первых вопросов лексикографии является, конечно, вопрос о различных типах словарей. Он имеет непосредственное практическое значение и эмпирически всегда как-то решался и решается. Между тем в основе его лежит ряд теоретических противопоставлений, которые и необходимо вскрыть.

1. Противоположение первое: словарь академического типа — словарь-справочник ...
2. Противоположение второе: энциклопедический словарь — общий словарь ...
3. Противоположение третье: thesaurus — обычный (толковый или переводный) словарь ...
4. Противоположение четвертое: обычный (толковый или переводный) словарь — идеологический словарь ...
5. Противоположение пятое: толковый словарь — переводный словарь ...
6. Противоположение шестое: неисторический словарь — исторический словарь ...⁵ [3]

Two decades ago, we made such a development system, and all this time we've been slowly but steadily filling it with linguistic data (of all types suitable for lexicographical purposes and from all sources available for us).

We began explaining the idea behind our lexicographical system in [4] and, as we previously wrote, have been formulated a conceptual generalisation that can be used as a kind of manual for creating other integrated / universal dictionaries in [7]. We published several more examples of what entries in our dictionary look like in [1]. The dictionary is also available online. [6] We are currently continuing to provide updates on the progress of our dictionary development and are also considering integrating other natural language processing methods (for example, based in machine learning and deep learning) both for collecting language usage examples for the dictionary and for generating language resources (e.g., parallel corpora, machine translation models etc.) with the help of the dictionary.

Now it's time to take the next step and to prepare it for print. The system supports automatic output in a look of paper dictionary from the very beginning. This means that even the data structure was created specifically to cover almost any paper dictionary (we analysed a lot of them during the design stage). Therefore, the preparation for publication does not mean adaptation, but just a usual editorial work of fully preparing the content, editing, and correcting mistakes. In any case, this is too much of a workload for two researchers without organisational background and financial support. Thus, we are open to proposals from institutions interested in: 1) publishing the Baltian-Latgalian (and in future probably Latgalian-Baltian too) dictionary; 2) to prepare and publish other dictionaries.

REFERENCES

- [1] Bērziņš A.A. Latgaliešu terminoloģijas izstrāde // Latviešu terminoloģija simts gados. LNB zinātniskie raksti. 6. (XXVI) sējums. Rīga: Latvijas Nacionālā bibliotēka, 2020.
- [2] Берзінь А.У. Электронная сiстема распрацоўкі латышка-латгальскага слоўніка // Язык и межкультурные коммуникации. Минск: БГПУ, 2011.
3. Contrast three: thesaurus vs. standard (explanatory or translation) dictionary ...
4. Contrast four: standard (explanatory or translation) dictionary vs. ideological dictionary ...
5. Contrast five: explanatory dictionary vs. translation dictionary ...
6. Contrast six: non-historical dictionary vs. historical dictionary ...

⁵ One of the most important questions in lexicography is, of course, the question of the different types of dictionaries. It has immediate practical significance and has always been and continues to be addressed empirically. However, at its core lies a series of theoretical contrasts that need to be explored.

1. Contrast one: academic dictionary vs. reference dictionary ...

2. Contrast two: encyclopedic dictionary vs. general dictionary ...

- [3] Щерба Л.В. Языковая система и речевая деятельность. Москва: Наука, 1974.
- [4] Дубичинский В.В. Теоретическая и практическая лексикография. Вена-Харьков: Gesellschaft zur Förderung slawistischer Studien, 1998.
- [5] Козырев В.А., Черняк В.Д. Современные ориентации отечественной лексикографии // Вопросы лексикографии. № 1 (5). Томск: Томский государственный университет, 2014.
- [6] Bierzeņš A.A., Kūceņš A. Baļtīšu-latgaļīšu vuordineica. <http://vuordineica.lv/> [Accessed 16.IV.2026]
- [7] ბერზინი ა.ა. უნივერსალური ლექსიკონის კონცეფცია / Conception of the Integrated Dictionary // II საერთაშორისო სიმპოზიუმი ლექსიკოგრაფიაში / II International Symposium in Lexicography. ბათუმში / Batumi, 2012.
- [8] Денисов П.Н. Системность и связанность в лексике и система словарей // Проблематика определений терминов в словарях разных типов. Ленинград: Наука, 1976.
- [9] Стабуrowa Е.Ю. Большой китайско-русский словарь: к 40-летней годовщине // Российское китаеведение. № 3(4). Москва: Институт Китая и современной Азии РАН, 2023.