PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 106 OCTOBER 2016 205–213

Lexicographic Tools to Build New Encyclopaedia of the Czech Language

Aleš Horák, Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

Abstract

The first edition of the Encyclopaedia of the Czech Language was published in 2002 and since that time it has established as one of the basic reference books for the study of the Czech language and related linguistic disciplines. However, many new concepts and even new research areas have emerged since that publication. That is why a preparation of a complete new edition of the encyclopaedia started in 2011, rather than just re-printing the previous version with supplements. The new edition covers current research status in all concepts connected with the linguistic studies of (prevalently, but not solely) the Czech language. The project proceeded for five years and it has finished at the end of 2015, the printed edition is currently in preparation. An important innovation of the new encyclopaedia lies in the decision that the new edition will be published both as a printed book and as an electronic on-line encyclopaedia, utilizing the many advantages of electronic dictionaries.

In this paper, we describe the lexicographic platform used for the Encyclopaedia preparation and the process behind the work flow consisting of more than 3,000 pages written by nearly 200 authors from all over the world. The paper covers the process of managing entry submissions, the development of tools to convert word processor files to an XML database, tools to cross-check and connect bibliography references from free text to structured bibliography entries, and the preparation of data for the printed publication.

1. Introduction

The first edition of the Encyclopaedia of the Czech Language (Bachmannová et al., 2002) was published in 2002. Since that time it has been adopted as one of the basic reference books for the study of the Czech language and related linguistic disciplines

 Corresponding author: rambousek@fi.muni.cz
Cite as: Aleš Horák, Adam Rambousek. Lexicographic Tools to Build New Encyclopaedia of the Czech Language. The Prague Bulletin of Mathematical Linguistics No. 106, 2016, pp. 205–213. doi: 10.1515/pralin-2016-0019.

not only in the Czech Republic, but by Bohemists all over the world. However, many new concepts and even new research areas have emerged since that publication. The Encyclopaedia editorial board (led by Petr Karlík) decided to prepare a complete new edition of the encyclopaedia, rather than just a re-print with supplements. The new edition covers current research as well all the concepts of linguistic studies connected with the Czech language. The project is coordinated by a team at the Faculty of Arts, Masaryk University, it started in 2011 and has finished at the end of 2015. Currently (June 2016), the encyclopaedia data undergoes the final proofreading and the final processing phase before publishing. The final version of the New Encyclopaedia contains 1569 entries, spanning over 3,000 pages, written by 191 authors.

As an important innovation of the original encyclopaedia from 2002, the new edition is primarily organized as an electronic encyclopaedia, utilizing the advantages of the electronic dictionaries. The printed version will be published by a well-known Czech publishing house Nakladatelství Lidové noviny based on the preprocessed data exported from the electronic edition. This move to electronic publishing is in line with recent trends in dictionary publishing (Tarp, 2012; Verlinde and Peeters, 2012). The DEB platform was selected as the dictionary writing system for the preparation of the new edition.

2. The DEB Platform Overview

Based on the experience with several tens of dictionary projects, the team at the NLP Centre FI MU has designed and implemented a universal dictionary writing system that can be exploited in various lexicographic applications to build large lexical databases. The system has been named Dictionary Editor and Browser (Horák and Rambousek, 2007), shortly DEB,[1] and has been used in more than twenty lexicographic projects since 2005, e.g. the development of the Czech Lexical Database (Rangelova and Králík, 2007), or currently running projects of the Pattern Dictionary of English Verbs (Hanks, 2004), Family names in UK (Hanks et al., 2011), and highly multimedial Dictionary of Czech Sign Language (Rambousek and Horák, 2015).

The DEB platform is based on the client-server architecture, which brings along a lot of benefits. All the data are stored on the server side and a considerable part of the client-side functionality is also implemented on the server, thus the client application can be very lightweight. The DEB platform approach provides very good tools for team cooperation: all data modifications are immediately seen by all involved users. The server also provides well arranged authentication and authorization functions. Unlike other dictionary writing systems (both commercial, and open-source), the DEB platform is not limited to one type of language or knowledge resources. DEB supports requirements of many frequently used resource types, while most of the applications

[1] http://deb.fi.muni.cz

specialize only on one type of data. The DEB platform and related projects are covered in detail in Rambousek (2015).

3. The Encyclopaedia Editing Process

3.1. The Editing Team Management

The encyclopaedia text preparation team consists of 191 authors, supported by 15 other team members. Here, the DEB platform support for complex access rights is utilized – all the users are hierarchically organized as *entry authors*, *entry referees*, *area administrators*, *editing assistants*, and *encyclopaedia coordinators* with various levels of access to the dictionary data. For example, the entry authors may only edit the entries assigned to them by the area administrator. The system can also limit access to all entries (or a selected subset) for some users during various editing phases, eg. for batch update. The editors can compare several versions of each entry – the original document provided by the author(s), the XML file stored in the database, the HTML preview for checking, and the final electronic version.

The management system also provide reporting tools to track progress of individual entries and overall encyclopaedia, such as:

- the current *editing phase* of an entry (posted by an author, converted to XML database, proofread, electronic version verified, confirmed by area administrator, etc.),
- the *number of work-in-progress* and *finished entries*,
- the *number of entries* and *"normalized" pages*[2] written by each of the authors,
- the option to see and compare the *editing history* of each entry.

Apart from the full history of document changes, the system also provides daily backups of the database.

3.2. Entry Editing and Conversion

Since the source materials of the original edition were prepared as a set of word processing documents, and mainly because some of the authors could not use the online editing tools, it was decided by the editorial board that in the first stage of the editing process, the entries will be written in the form of a word processing documents. To allow the use of the new features for the electronic encyclopaedia, special markup tags inside the standard document text were introduced for the project:

- the entry headword and its variants,
- the entry category classification,
- the entry author(s),
- splitting the entry text to two parts – a standard part for public users, and an advanced part for researchers and experienced scholars,

[2] 1,800 characters of text per page

- the list of bibliographic references,
- a definition (form of a sub-entry) in the text,
- multimedia attachment files (images, audio recordings, video files), either included inline in the text, or on separate pages,
- cross-references to other entries, sub-entries or bibliographic references.

At the first step, documents provided by the authors in several word processing formats are unified by automatic conversion to the Open Document format (Brauer et al., 2011).

In the next step, the ODF documents are converted to an internal XML format. The word processor instructions and special markup tags are converted to semantic tags of the encyclopaedia XML format. Wherever possible, the included images are converted to vector graphic formats to provide the best image quality both in the electronic and the printed edition. During this phase, varying text formatting is unified to the common layout of the New Encyclopaedia.

All of the 1569 entries were regularly processed and converted during the Encyclopaedia editing. As more documents were provided by the authors, the conversion tools were continuously updated to handle various input document formats.

After the upload and conversion, the documents are stored in the DEB XML database and edited via the online tools. It is also possible to download the entries for offline editing and upload the updated version later.

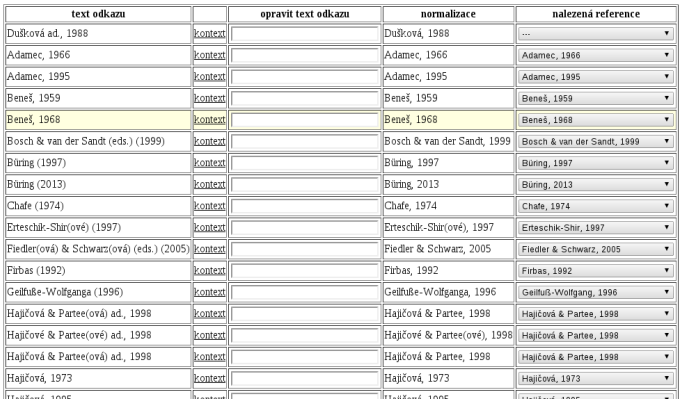
3.3. Cross-references Checks

Although the subject areas and entry lists were planned beforehand, many changes were introduced during the writing process. Sometimes, completely new entries emerged to describe the current state of linguistic research. In other cases, entries were split or merged for the best presentation of the concepts and spreading the length of entries more evenly. However, such changes could cause various errors in entries cross-referencing.

In the final phase of the Encyclopaedia preparation, all cross-references between entries were checked and verified. The lexicographic system tools scanned all the entries and their connections, reporting any inconsistencies or links to missing entries. The editing assistants then browsed through and fixed each of the errors, either with updating the cross-reference to another entry, creating new variant headword, or deleting the link. During this process, several entries were identified that were omitted during the writing phase and needed to be added.

3.4. Bibliographic References Processing

After the final form of all entries was delivered, the bibliography lists and all bibliographic references in the text were verified. Since the encyclopaedia texts are written in Czech, the bibliographic references within the entry texts may come in different



text odkazu		opravit text odkazu	normalizace	nalezená reference
Dušková ad., 1988	kontext		Dušková, 1988	---
Adamec, 1966	kontext		Adamec, 1966	Adamec, 1966
Adamec, 1995	kontext		Adamec, 1995	Adamec, 1995
Beneš, 1959	kontext		Beneš, 1959	Beneš, 1959
Beneš, 1968	kontext		Beneš, 1968	Beneš, 1968
Bosch & van der Sandt (eds.) (1999)	kontext		Bosch & van der Sandt, 1999	Bosch & van der Sandt, 1999
Büring (1997)	kontext		Büring, 1997	Büring, 1997
Büring (2013)	kontext		Büring, 2013	Büring, 2013
Chafe (1974)	kontext		Chafe, 1974	Chafe, 1974
Erteschik-Shir(ové) (1997)	kontext		Erteschik-Shir(ové), 1997	Erteschik-Shir, 1997
Fiedler(ová) & Schwarz(ová) (eds.) (2005)	kontext		Fiedler & Schwarz, 2005	Fiedler & Schwarz, 2005
Firbas (1992)	kontext		Firbas, 1992	Firbas, 1992
Geilfuße-Wolfganga (1996)	kontext		Geilfuße-Wolfganga, 1996	Geilfuß-Wolfgang, 1996
Hajičová & Partee(ová) ad., 1998	kontext		Hajičová & Partee, 1998	Hajičová & Partee, 1998
Hajičové & Partee(ové) ad., 1998	kontext		Hajičové & Partee(ové), 1998	Hajičová & Partee, 1998
Hajičová & Partee(ová) ad., 1998	kontext		Hajičová & Partee, 1998	Hajičová & Partee, 1998
Hajičová, 1973	kontext		Hajičová, 1973	Hajičová, 1973

Figure 1. Verification and matching of bibliographic references.

inflected forms (grammar cases, masculine/feminine name endings, etc.). As a first step, a uniform and unique representation of each item in the bibliography list was created. Although the authors followed the CSN ISO 690-2 standard (CSN690, 2011) for references, many items contained some sort of spelling or typing errors. All inconsistencies to the standard were reported and fixed.

In the next step, all references in the entry text were transformed to the same unified form and matched against the entry bibliography list. From the total of 16,252 bibliography links, 95 % were correctly matched using the uniform representation. See Figure 1 for an example of the bibliography checking form to verify and interlink the bibliographic references. The remaining cases consisted of following issues that were handled by the editing assistants:

- an unusual form or misspelled name, year or other part of the bibliography reference,
- a bibliography entry not following the standard form,
- a choice of two (or more) publications by the same author in the same year,
- a missing bibliographic entry,
- a misplaced reference in the entry text.

3.5. Final Proofreading by Authors

When the conversion, verification and finalization processes were successfully carried out, all the authors were asked to proofread their own entries before submitting the final data to the publisher.

For this purpose, an entry representation similar to the printed edition was created in the PDF format and prepared for download on personalized author checking web

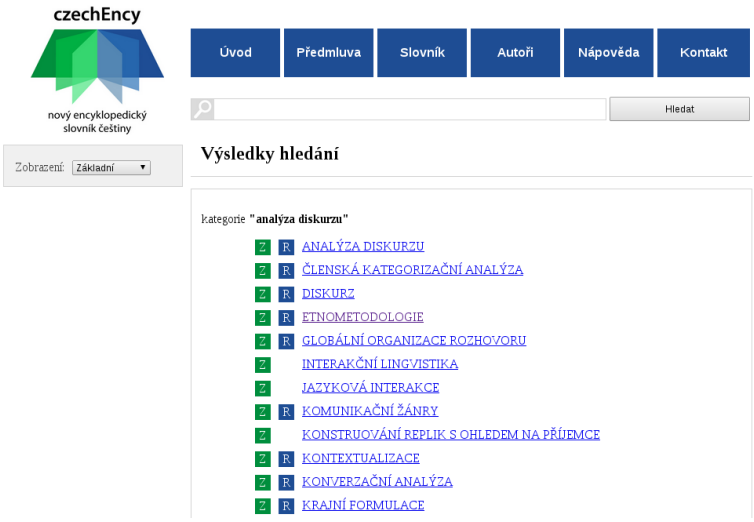


Figure 2. Search results displaying entries from a selected category. The green and blue squares indicate the proficiency level (standard/advanced) of the entry parts available.

pages. Subsequently, the authors were able to enter the proofreading comments and requests into a special web-based form. All identified issues[3] were then transferred to the database by the team of editing assistants. During the review, 110 entries written in an early stage of the project were (usually briefly) updated to reflect the current research. Because of the changes, it was required to verify some of the cross-references and bibliographic references again.

3.6. The Printed Edition

The printed edition of the Encyclopaedia is going to be published by one of the largest Czech publishing houses, Nakladatelství Lidové noviny. The lexicographic system contains tools to prepare the data in the format used by the publisher for typesetting:

- each entry is saved in a separate XML file,
- the metadata are updated (e.g. author abbreviation is changed to full name),
- the cross-references are updated to link correctly to the saved XML files,
- all included multimedia files are downloaded,
- all images are saved in all available formats to provide the best quality for typesetting.

[3]There were about 1,700, mostly small, corrections reported.

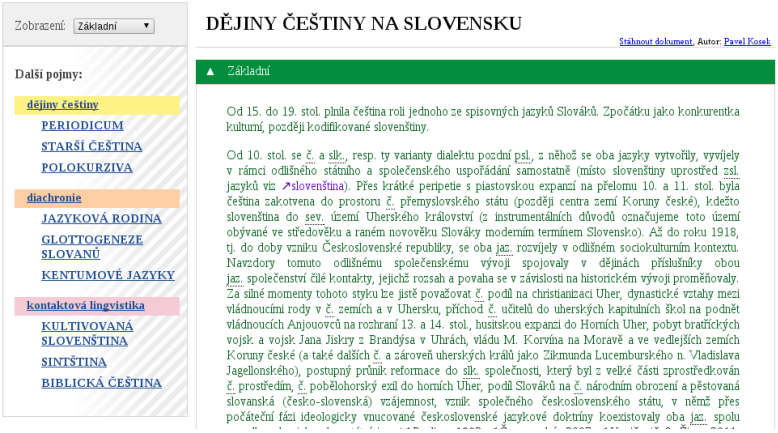


Figure 3. Preview of an entry, with links to more information in the same category.

Due to file conversions, cross-references checks, and various document updates, preparation of all 1569 entries for the publisher takes one hour. Without additional features, the complete export of the database takes less then 5 minutes.

3.7. The Electronic Edition

The New Encyclopaedia edition takes the advantage of the electronic and multimedia dictionary features to help users with navigation in the encyclopaedia, to obtain extra relevant information, and to better understand the concepts. See Figure 2 for an example of search results, and Figure 3 for an example of an entry preview. The DEB platform tools take care of properly encoding and providing all the following information:

- cross-references to other entries or entry parts,
- links to external websites,
- references to the bibliography, with the possibility to query external library resources,
- images, charts, diagrams etc.,
- sound recordings (e.g. samples of various dialects, see Figure 4),
- animations and video recordings (e.g. video recordings showing sign language gestures),
- explanations of the abbreviations used in the text or bibliography lists.

To make the encyclopaedia content comprehensible and useful for different reader groups, the entries can be described in two different levels of proficiency, i.e. the entry text can contain a standard part and an advanced part. Out of the total of 1,569 entries

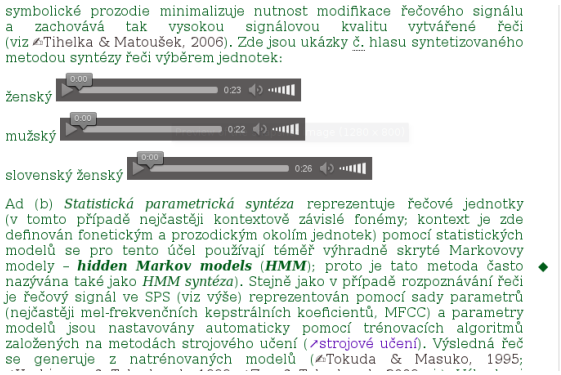


Figure 4. Example of an entry with inline sound recordings.

in the encyclopaedia, 1,093 entries contain just the standard part, 193 entries contain only the advanced part, and 283 entries have both descriptive parts.

On the encyclopaedia website, readers may choose their preference of the default description level. For example, readers may hide the advanced information and when they search for an entry, only the standard entries or descriptions are provided.

The system tracks the most often visited entries in each category and provides hints to extra information in related categories for readers interested in certain topics.

4. Conclusions

We have described the tools and processes utilized to build the New Encyclopaedia of Czech, the largest electronic encyclopaedia devoted to the Czech Language and related linguistic studies. The presented lexicographic tools successfully supported the team of more than 200 authors and assistants during creation of both printed and electronic version of one of the most important resource for the study of the Czech language and many connected areas of linguistic research.

Acknowledgements

This paper describes the Encyclopaedia created as a result of the Czech Science Foundation project P406/11/0294. This work was also partially supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and by the national COST-CZ project LD15066. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

Bibliography

Bachmannová, Jarmila, Petr Karlík, Marek Nekula, and Jana Pleskalová. *Encyklopedický slovník češtiny*. Lidové noviny, Praha, 2002.

Brauer, Michael, Patrick Durusau, Gary Edwards, David Faure, Tom Magliery, and Daniel Vogelheim. ISO/IEC 26300:2006: Open Document Format for Office Applications 1.2, 2011.

CSN690, 2011. *ČSN ISO 690 (01 0197) Informace a dokumentace - Pravidla pro bibliografické odkazy a citace informačních zdrojů*. Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Praha, 3rd edition, 2011.

Hanks, Patrick. Corpus Pattern Analysis. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France, 2004. Universite de Bretagne-Sud.

Hanks, Patrick, Richard Coates, and Peter McClure. Methods for Studying the Origins and History of Family Names in Britain. In *Facts and Findings on Personal Names: Some European Examples*, pages 37–58, Uppsala, 2011. Acta Academiae Regiae Scientiarum Upsaliensis.

Horák, Aleš and Adam Rambousek. DEB Platform Deployment – Current Applications. In *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*, pages 3–11, Brno, Czech Republic, 2007. Masaryk University.

Rambousek, Adam. *Creation and Management of Structured Language Resources*. PhD thesis, Faculty of Informatics, Masaryk University, 2015.

Rambousek, Adam and Aleš Horák. Management and Publishing of Multimedia Dictionary of the Czech Sign Language. In *Natural Language Processing and Information Systems - 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015. Proceedings*, Lecture Notes in Computer Science. Springer, 2015.

Rangelova, Albena and Jan Králík. Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century. In *Proceedings of the Computer Treatment of Slavic and East European Languages 2007*, pages 209–217, Bratislava, Slovakia, 2007.

Tarp, Sven. Theoretical challenges in the transition from lexicographical p-works to e-tools. In Granger, Sylviane and Magali Paquot, editors, *Electronic Lexicography*, pages 107–118. Oxford University Press, Oxford, 2012. ISBN 978-0-19-965486-4. doi: 10.1093/acprof:oso/9780199654864.001.0001.

Verlinde, Serge and Geert Peeters. Data access revisited: The Interactive Language Toolbox. In Granger, Sylviane and Magali Paquot, editors, *Electronic Lexicography*, pages 147–162. Oxford University Press, Oxford, 2012. ISBN 978-0-19-965486-4. doi: 10.1093/acprof:oso/9780199654864.001.0001.

Address for correspondence:
Adam Rambousek
rambousek@fi.muni.cz
Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic