A lexico-semantic database of Czech. An interim report

Ondřej Tichý — Zora Obstová — Aleš Klégr (Charles University, Prague)

ABSTRACT

The paper describes the intermediate stage of a lexicographical project, whose aim is to digitize and align two Czech onomasiological dictionaries (Haller 1969–77; Klégr 2007) in order to create an integrated digital multi-purpose lexico-semantic database of Czech. The two dictionaries are based on different categorization systems (Hallig and von Wartburg; Roget) and use different formats. Their content only partially overlaps, making them largely complementary. Their linkage is planned to be achieved through their structural elements (categories of their hierarchies) rather than by matching individual headwords. The four phases of the project are digitization, encoding, programming and testing. The digitization of both dictionaries and the encoding of one of them have been completed, and the preliminary steps in programming the platform are underway.

KEYWORDS

onomasiological lexicography, thesaurus, lexico-semantic database, digitization, Czech

DOI

https://doi.org/10.14712/18059635.2021.1.5

1 INTRODUCTION

The project is part of a drive to give a new lease of life to onomasiological lexicography and take it to the 21st century technologically. In general, this type of dictionary tends to be underrepresented and so relatively little attention is paid to it in the lexicographical literature (Reichmann 1990: 1063–1064). The historical tradition of onomasiological lexicography is outlined by Sterkenburg (2003) and more recently by Kay and Alexander (2016). The English tradition and its seminal culmination, Roget's *Thesaurus*, is covered by Hüllen (1999, 2004). He also pays attention to the European scene, including the contribution made by John Amos Comenius (*Janua Linguarum, Orbis Pictus*) in the mid-17th century. The various offshoots of Roget's *Thesaurus* in over 150 years are detailed in Klégr (2000). In fact, most thesauruses today can be traced either to Roget's approach or to the German systems (Dornseiff 1933; Hallig and von Wartburg 1952). Both Roget's and the German systems are reviewed by Fischer (2004).

Although a dictionary is nowadays generally associated with an alphabetical listing of words, dictionary-making in classical antiquity (and before) started with thematic compilations collected on a semantic basis. The one serious drawback of meaning-based dictionaries has always been the difficulty of locating a given word in them, so much so that thesauruses were routinely provided with an alphabetical index to facilitate the search. The onset of digital lexicography, allowing full-text search, has removed this obstacle, blurring the distinction between onomasiological and alphabetical dictionaries (Sierra and McNaught 2000: 265). If anything, com-

puter technology facilitated the compilation of new onomasiological dictionaries in the early 21st century, e.g. the Czech thesaurus (Klégr 2007), *WordNet* (Princeton 2010), *Historical Thesaurus of the OED* (Kay et al. 2009), two Italian dictionaries (Simone 2010; Feroldi and Dal Prà 2011), the Danish thesaurus (Nimb et al. 2014) and most recently two Slovene synonym dictionaries (Ahlin et al. 2016; https://viri.cjvt.si/sopomenke/slv/), although, so far, few of these works are digitized. E-dictionaries can be produced from scratch, or by building on digitized print dictionaries, of which the present project is a case in point.

2 AIMS & MOTIVATION

The project aims to create a digital lexico-semantic database of Czech by marrying two Czech onomasiological print dictionaries, Haller (1969–77) and Klégr (2007), conflating their data and interfacing their structures. Given the differences in the design, conception and focus of the two dictionaries and their entries, their merging poses a formidable technical and labour-intensive task. The outcome will be an integrated and multi-functional semantic network.

The database will be equally useful to both the NLP specialist and just about anyone working with Czech (writers, translators, journalists, etc.). The latter type of user will benefit from a data-and-information rich, highly searchable, and freely available online application (editable and updatable), while the former group will be able to use the database for lexico-semantic research by means of the online app, application programming interface (API), or to work with the exportable dataset. The database is also expected to help future lexicographers exploit the semantic relations encoded in the two dictionaries when describing lemmas or corpus linguists performing semantic tagging or disambiguation. The API will likewise allow third party applications to benefit from semantic fields, e.g. to increase search yields.

3 STATE OF THE ART

Considering the enormous success of Roget's *Thesaurus* (more than 32,000,000 copies sold since its first edition in 1852), surprisingly it has no easily accessible online version to date, unlike many comparably influential semasiological dictionaries. The fact that few onomasiological dictionaries are digitized is probably due to the full-text search available in the digital form, which gets around the distinction between differently organized dictionaries and the need for them, making the onomasiological dictionary redundant for some users.

This, however, is only partly true, because the entries of semasiological dictionaries may contain information not conveniently retrievable even using full-text function-focused search (Sierra and McNaught, 2000, p. 265). Moreover, retrievability by human users is only part of the intended functions of our project. Digitization of onomasiological resources is not done merely to improve access: these resources are highly useful in NLP (Cassidy 2000; Kwong 2001; Kennedy 2008; Jarmasz 2012; Nimb et al.

2014). But to be useful, they need to be digitized in a way that makes their structural information explicit, unambiguous and consistent enough to be computer tractable.

With hardly any digital onomasiological dictionaries available, the process of their digitization is largely unexplored and so is the standardization of encoding the onomasiological data (see Section 5.2 below). A number of ventures using *Roget's* are described in Jarmasz (2012: 10), but except for the datasets available in FACTOTUM, the Electronic Lexical Knowledge Base or the follow-up Open Roget's, none has led to a full-fledged online dictionary. Another notable exception outside English is the *Den Danske Begrebsordbog* project (Nimb et al. 2014), whose online version, however, is still in the planning stage.

Although most of our methodology cannot rely on previous projects, a few authors (e.g. Kwong 2001 and Sierra and McNaught 2000) did try to integrate two onomasiological resources. Unlike them, we have decided to link the resources primarily through their structural elements (categories of their hierarchies) rather than by matching each individual headword. Also, we will not attempt to enrich the structure by turning the implicit relations between headwords into explicit formalized structural elements. We expect these two decisions to make the process of digitization and integration much easier than suggested by these authors.

4 DESCRIPTION OF THE INPUT PRINT DICTIONARIES

The two input thesauruses, Haller's *Český slovník věcný a synonymický* (*ČSVS*, 1969–77) and Klégr's *Tezaurus jazyka českého* (*TJČ*, 2007), offer a comprehensive coverage of the Czech language, but due to differences in the time of their origin (they are separated by several decades), categorization approaches and the extent of completion (Haller remains unfinished), their word-stock is only partially overlapping. As they are largely complementary, their synthesis is mutually enriching. Especially relevant is the fact that *TJČ* is based on Roget, while *ČSVS* uses Hallig and von Wartburg's (1952) conceptual system. In this respect Kay and Alexander's (2016: 379) claim that "[t]here is no record of [their conceptual system] … having much effect on actual thesaurus-making" has to be corrected.

ČSVS was initiated by the Translators' Section of the Czechoslovak Writers' Union in the late sixties and conceived as a reference book for translators, writers and journalists as well as teachers and the general public. In addition it was to serve linguistic purposes and aimed to collect and categorize the entire Czech lexicon. Unfortunately, due to Hallig and von Wartburg's overabundant conceptual system and Haller's untimely death, the ambitious project stopped halfway. Its three volumes came out in 1969, 1974 and 1977, the index volume only in 1986. The dictionary carries 3193 entries (sections) over 1594 pages, comprising in excess of 400,000 lemmas. Of the three main parts of Hallig and von Wartburg's *Begriffssystem*, The Universe (Sky, Earth, Plants, Animals), Man (Physical Aspects, Soul and Intellect, Social Aspects and Social Organisation) and Man and the Universe (The Basics, Science and Technology), only the first and half of the second part (its first two sections) are covered. The choice of Hallig and von Wartburg's *Begriffssystem* was a risky decision; their sophisticated universalist system, made up purely of concepts independent of words, does

not specify the microstructure or the grammatical form of the words representing particular concepts. It was meant to be tested by dictionaries composed for particular languages and this is what Haller attempted. He (and his collaborators) had to design the microstructure, the format of the entries, and other features. Their entries are often of uneven length; apart from co-ordinate and synonymous terms, they may contain additional information of various types (illustrative, explanatory, stylistic, collocational, etc., according to the character of the lemma) at the expense of uniformity, which complicates their digitization.

TJČ, on the other hand, largely follows the well-tested format of *Roget's*, namely its abridged pocket version (Carney and Waite 1985) which implicitly preserves the original hierarchical system (Dutch's 1962 revision) of six classes (Abstract Relations, Space, Matter, Intellect, Volition, Emotion), their respective sections and heads. It preserves the progression of the heads, the principle of each head being followed by a semantically opposite (contrasting) one, and it maintains the microstructure of the head, presenting the respective concept in the form of a noun, adjective, verb and (loosely) adverb(ial). Carney and Waite's 882 heads were expanded by two original heads. None of the Czech entries are translations of the English ones; the head concepts were used as a springboard to the search for Czech words representing the concept independently of English. The 885 heads in the dictionary part take up 499 pages and the rest of the 1189 pages comprise an alphabetical index. The estimated number of lexical items in *TJČ* is some 200,000. Like *Roget's*, it is intended as a practical reference book for active use.

For a detailed comparison of *ČSVS* and *TJČ* see Obstová (2021).

5 PLAN OF IMPLEMENTATION

Our implementation plan consists of four major phases: 1. digitization; 2. encoding; 3. programming; and 4. testing.

5.1 DIGITIZATION

The plan is to scan the paper dictionaries and perform an automatic optical character recognition (OCR) using the ABBYY FineReader program set to recognize standard Czech. Since especially Haller's dictionary contains special characters and complex typography, some effort must be put into pattern training provided by the software to improve the accuracy. Previous experience indicates that a great many manual corrections will still be necessary because automatic OCR is never completely reliable, even with the recognition patterns trained specifically for the data. However, since all the digital data need to be manually processed in the encoding phase in any case, OCR corrections will be left to that stage.

5.2 ENCODING

In the encoding phase, the digitized data will be converted into TEI-XML format. XML is an (extensible) markup language that employs tags to express structures and their properties (such as dictionary micro- and macrostructure) and stores the ac-

tual content as open text inside those tags. The Text Encoding Initiative (TEI) maintains guidelines and schemata that set a standard of usage of the XML encoding for a variety of fields mainly in digital humanities, including lexicography. This format is chosen because it is both open and popular. It is also relatively well defined. Although the current TEI Dictionary specifications were not developed and are not a very close fit for onomasiological dictionaries, and in its permissive interpretation are too pliable to be really useful, the DARIAH supported TEI-Lex0 initiative's specifications (Bański et al. 2017) that are currently in development seem the best choice for possible interoperability of our data as well as for its long-term preservation. Based on this specification, a subset schema needs to be developed that will define where each element and structure can appear. The schema should be as restrictive as possible, yet allow encoding to fully capture the actual structure of the original dictionaries. The restrictiveness of the schema helps to make the encoding process more precise and also prevent errors in manual editing of the data.

When planning the project, we had the intention to process the digitized texts using a series of custom-made scripts that would take cues from the original typography of the dictionaries and, based on typography, to tag the data according to the TEI standard. While this is feasible and has been done in the past, experience shows that with relatively complex and largely inconsistent data like that of Haller's dictionary, transforming the data using custom-made scripts is a laborious process of trial and error with often mixed results necessitating large-scale manual corrections. With this in mind, we have explored other options for tagging the data, focusing on machine-learning methodologies that might potentially save time and effort with both designing custom scripts and with manual corrections, if better efficiency could be reached. In the end, the GROBID-Dictionaries project was chosen. According to Khemakhem et al. (2017), "GROBID (GeneRation Of Bibliographic Data) is a machine learning system for parsing and extracting bibliographic metadata from scholarly articles, mainly text documents in PDF format. It relies on Conditional Random Fields (CRF, Lavergne et al. 2010) to perform a multi-level sequence labelling of text blocks in a cascade fashion; the text blocks are then being extracted and encoded in TEI elements." GROBID-Dictionaries is a derived project that uses similar technology but focuses on parsing lexicographical data. It is beyond the scope of this paper to describe the inner workings of GROBID-Dictionaries; suffice it to say that the tool allows users to define a subset of a TEI schema for a specific dictionary, manually tag a relatively small set of sample data and then train the model on the data, which can be applied to tag the rest of the dictionary, and if need be repeated, until a sufficient precision of encoding is reached (for further details on GROBID-Dictionaries, see Khemakhem et al. 2017 and Khemakhem et al. 2018). While the software is also not specifically designed for onomasiological dictionaries, its main researcher has joined our team to accommodate his tools to our specific purposes.

While the precision of data encoded with machine learning technology is expected to be higher than that of the data encoded with custom scripting, it is nevertheless necessary to go through the whole dataset manually and check both for the precision of the tagging and the quality of the initial OCR and the plan is to train student assistants for the task using an XML editor that will ensure compliance with our schema.

This phase will conclude with manual pairing of the dictionary hierarchies (as noted above).

5.3 PROGRAMMING

The online platform will be realized in JavaScript programming language both on the server as well as on the client side. The server side will be running in a Node.js environment separated into four modules based on Express.js. These will be contained in virtual spaces using Docker. Thanks to this, all modules will be easy to update and deploy under all kinds of technical scenarios.

The Thesaurus module will form the backbone of the platform. It will communicate with the database using xQuery and process the data in XML. The Transformation module will transform the data from XML into other formats, especially into HTML for the web interface and JSON for the API. The Users module will be an auxiliary tool for authentication, user and client management. The Gateway module will provide flexible communication between the modules, the user and the application interfaces.

The client-side interface will be based on a Nuxt.js framework, which in turn is based on a Vue.js framework. The main advantages of this solution are that, thanks to its server-side compilation, it is available to indexing tools and that it supports the creation of progressive cross-platform web apps. The progressive web apps will make the platform interface comparable to any native applications providing access to local storage and therefore allow some basic offline functionality.

Since the basic data format will be the TEI-XML standard, it will be stored in an ExistDB database. ExistDB should allow for all necessary operations, but in cases where this proves problematic or too demanding in terms of computational resources, the data may at times need to be transformed into, or from, a more common relational database such as MySQL.

5.4 TESTING

The testing will be carried out in several phases, starting with automated testing using the Jest framework, followed by human tests with the project members, student assistants, specialists affiliated with the project and finally the public.

6 CURRENT STATE OF IMPLEMENTATION

The project has already successfully completed the digitization stage and the encoding phase for the *TJČ*. The encoding of the *ČSVS* proved to be structurally more complex and less consistent than expected. The machine learning technologies have been adopted for its encoding, which is currently under way. Concurrently, the early stages of programming of the platform have also been initiated.

FUNDING

This work was supported by Technologická agentura České republiky [TL02000041]; and by the European Regional Development Fund project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (reg. no.: CZ.02.1.01/0.0/0.0/16_019/0000734).

REFERENCES

Ahlin, M. et al. (2016) *Sinonimni slovar slovenskega jezika.* Ljubljana: Založba ZRC.

Bański, P., J. Bowers and T. Erjavec (2017) TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In: Kosem I. et al. (eds) *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference, Sep 2017, Leiden, Netherlands*, 485–494. Brno, Czech Republic: Lexical Computing CZ s.r.o.

Carney, F. and M. Waite (eds) (1986) *Pocket English Thesaurus*. London: Penguin.

Cassidy, P. (2000) An Investigation of the Semantic Relations in the Roget's Thesaurus: Preliminary Results. *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics — CICLing 2000*, 181–204.

Dornseiff, F., U. Quasthoff and H. E. Wiegand (2004, 1st ed. 1933) *Der deutsche Wortschatz nach Sachgruppen*. Berlin, Boston: De Gruyter.

Dutch, R. A. (ed) (1962) *Roget's Thesaurus of English Words and Phrases*. London: Longman.

Feroldi, D. and E. Dal Prà (2011) *Dizionario analogico della lingua italiana*. Bologna: Zanichelli.

Fischer, A. (2004) The Notional Structure of Thesauruses. In: Kay C. and J. Smith (eds) *Categorization in the History of English*, 41–58. Amsterdam: Benjamins.

Haller, J. (1969–77) *Český slovník věcný a synonymický* 1–3 [The Czech Thematic and Synonym Dictionary]. Praha: Státní pedagogické nakladatelství.

Hallig, R. and W. von Wartburg (1952) *Begriffssytem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas*. Berlin: Akademie-Verlag.

Hüllen, W. (1999) *English Dictionaries 800–1700. The Topical Tradition*. Oxford: Oxford University Press.

Hüllen, W. (2004) *A History of Roget's* Thesaurus*: Origins, Development, and Design*. Oxford: Oxford University Press.

Jarmasz, M. (2012) *Roget's Thesaurus as a Lexical Resource for Natural Language Processing*, *CoRR, abs/1204.0*. Available at http://arxiv.org/abs/1204.0140 (last accessed 12 April 2020).

Kay, C. and M. Alexander (2016) Diachronic and Synchronic Thesauruses. In: Durkin P. (ed) *The Oxford Handbook of Lexicography*, 367–380. Oxford: Oxford University Press.

Kennedy, A. and S. Szpakowicz (2008) Evaluating Roget's Thesauri. *Proceedings of Acl-08: HLT*: 416–424.

Khemakhem, M., L. Foppiano and L. Romary (2017) Automatic extraction of TEI structures in digitized lexical resources using conditional random fields. *Electronic Lexicography, eLex 2017*, Leiden, Netherlands, hal-01508868v2.

Khemakhem, M., A. Herold and L. Romary (2018) Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX workshop at LREC 2018*, Miyazaki, Japan, hal-01708137v2.

Klégr, A. (2000) Rogetův Thesaurus a onomaziologická lexikografie [Roget's Thesaurus and onomasiological lexicography]. *Časopis pro moderní filologii* 82/2, 65–84.

Klégr, A. (2007) *Tezaurus jazyka českého. Slovník českých slov a frází souznačných, blízkých a příbuzných* [Thesaurus of the Czech Language. A Dictionary of Synonymous, Similar and Related Words and Phrases]. Prague: Nakladatelství Lidové noviny.

Kwong, O. (2001) Forming an Integrated Lexical Resource for Word Sense Disambiguation. In: *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation.* Hong Kong: City University of Hong Kong.

Lavergne, T., O. Cappé and F. Yvon (2010) Practical very large scale crfs. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 504–513. Association for Computational Linguistics.

Nimb, S., L. Trap-Jensen and H. Lorentzen (2014) The Danish Thesaurus: Problems and Perspectives. In: *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15–19.

Obstová, Z. (2021) Zwei onomasiologische Wörterbücher als Basis für eine

lexikalisch-semantische Datenbank des Tschechischen. In: Kloudová V. et al. (ed), *Spielräume der modernen linguistischen Forschung*, 92–111. Praha: Karolinum.
Princeton University (2010) About WordNet. *WordNet*. Princeton University.
Sierra, G. and J. McNaught (2000) Extracting Semantic Clusters from MRDs for an Onomasiological Search Dictionary. *International Journal of Lexicography*, *13*(4), 264–286.
Reichmann, O. (1990) Das onomasiologische Wörterbuch: Ein Überblick. In: Hausmann, F. J. (ed) *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Berlin: De Gruyter.
Simone, R. (2010) *Grande dizionario analogico della lingua italiana*. Torino: UTET.
Sterkenburg., P. van (2003) Onomasiological Specifications and a Concise History of Onomasiological Dictionaries. In: Sterkenburg, P. van (ed) *A Practical Guide to Lexicography*, 127–153. Amsterdam: Benjamins.

LIST OF ABBREVIATIONS

API: Application Programming Interface
CRF: Conditional Random Fields
ČSVS: Český slovník věcný a synonymický
DARIAH: Digital Research Infrastructure for the Arts and Humanities
GROBID: GeneRation Of Bibliographic Data
HTML: Hypertext Markup Language
JSON: JavaScript Object Notation
(My)SQL: Structured Query Language
NLP: Natural Language Processing
OCR: Optical Character Recognition
OED: Oxford English Dictionary
TEI: Text Encoding Initiative
TEI-XML: Text Encoding Initiative — Extensible Markup Language
TJČ: Tezaurus jazyka českého
XML: Extensible Markup Language

Ondřej Tichý
Faculty of Arts, Charles University, nám. Jana Palacha 1/2, 116 38, Praha 1
ORCID ID: 0000-0001-5088-3129
ondrej.tichy@ff.cuni.cz

Zora Obstová
Faculty of Arts, Charles University, nám. Jana Palacha 1/2, 116 38, Praha 1
ORCID ID: 0000-0002-1678-6947
zora.obstova@ff.cuni.cz

Aleš Klégr
Faculty of Arts, Charles University, nám. Jana Palacha 1/2, 116 38, Praha 1
ORCID ID: 0000-0001-7760-6631
ales.klegr@ff.cuni.cz