Overview and Future of Czech Wordnet

Adam Rambousek, Karel Pala, Sandra Tukačová

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanick 68a, 602 00 Brno, Czech Republic
rambousek@fi.muni.cz, pala@fi.muni.cz, 309607@mail.muni.cz

Abstract. Czech Wordnet represents one of the national wordnets created during the EuroWordNet and Balkanet projects. However, the data contains various issues that affects the use of Czech Wordnet in NLP applications. Due to lack of resources, it was not possible to update Czech Wordnet thoroughly since the publication of the first version. In 2017, we have started a project to evaluate and update Czech Wordnet, followed by the connection to Collaborative Interlingual Index. This paper provides overview of various updates and extensions of the Czech Wordnet data, and presents the roadmap to publish revised version of Czech Wordnet under open license.

Keywords: EuroWordnet, Balkanet, wordnet, Czech Wordnet, DEB-VisDic

1 Introduction and history of Czech Wordnet

After its publication, Princeton WordNet [1] proved its usability as a lexical resource, both for users and various NLP tasks. Wordnet also inspired many projects aiming either to create semantic networks in other languages, or extend wordnet with new features. The first attempt to build localized wordnets was the EuroWordNet [2] project in 1996, coordinated by Piek Vossen at the University of Amsterdam. In the first phase, EuroWordNet included Dutch, Italian, Spanish, and English wordnets. In the next phase, German, French, Estonian, and Czech wordnets were added.

EuroWordNet introduced two new features that were necessary for language compatibility. With the aim to build semantic networks in several languages that share the same language core, list of Base Concepts was described. The list includes 1310 synsets shared amongst all EuroWordNet languages and represents the part of wordnet that should be encoded first. Another purpose of Base Concepts were linguistic studies of language differences.

Because the national wordnets reflect various languages with specific hierarchy, EuroWordNet project established Interlingual Index (ILI). The index contained language independent ontology. Each wordnet connects synsets to ILI, thus enabling multi-lingual links. The features and processes developed during the EuroWordNet project were later re-used during building of other national wordnets.

One of such projects was the Balkanet [3] project in 2001-2004, aiming to expand the number of national wordnets for European languages. Balkanet project covered Bulgarian, Greek, Romanian, Serbian, and Turkish wordnets. Together with newly developed wordnets, verb synsets in Czech wordnet were extended.

As mentioned, Czech wordnet was created in EuroWordNet and Balkanet projects by the Natural Language Processing Centre at the Faculty of Informatics, Masaryk University (NLPC). However, it was published through ELRA under closed and paid license. Although it is possible to get research license, Czech Wordnet data are still not available in an open form, and that issue hampers many attempts to include synset data into any 3rd party tool.

Since 2004, no thorough work was possible to extend, fix, or update Czech Wordnet data. In 2017, we have started a project to evaluate and update Czech Wordnet. Followed by connection to Collaborative Interlingual Index and open publication of the wordnet.

2 Available versions of Czech Wordnet

2.1 Original Czech Wordnet

The original version of the Czech Wordnet [4] is available for licensing from ELRA. This is the version created during EuroWordNet and Balkanet projects, and contains 28,201 synsets with 43,958 literals. All the synsets are linked to their counterpart in Princeton Wordnet 2.0. Part of verb synsets (824) were also enriched with verb frames.

Primary method for the wordnet creation was the top-down approach (proposed in the EuroWordNet project). Lexicographers consulted several resources, available at the time in electronic form – Czech explanatory dictionary [5], English-Czech dictionary, Czech synonymy dictionary, and the DESAM corpus. Although the explanatory dictionary contained information about hypernym for some headwords, this information was not entered systematically. This led to the solution that most of the hyperonymical relations were directly transferred from the Princeton Wordnet. Information on Czech synonyms was more extensive, however not covering all concepts needed. As a result, many synsets were exact translations of synsets from Princeton Wordnet.

This approach caused various issues with the data. Most notable example are the synsets containing words that are not exactly synonyms, or only rare in the Czech language, but present in the Czech Wordnet because of the translation from English. For example, English synset *cabriolet:1, cab:2* has the equivalent Czech synset *kabriolet:2, dvoukolový jednospřežní povoz:1, koňská drožka:1* (cabriolet, two-wheeled one horse cart, horse-drawn carriage). Although the translation is correct, this sense of *kabriolet* in Czech is very archaic, in current language the only sense used in spoken language is the convertible car. Another problem is the inclusion of multiword expressions in the synset, which may be justified in some cases, these are not fixed lexical units in the Czech language.

2.2 2009 Edited version

To deal with some of the issues mentioned above, core synsets of the Czech Wordnet were edited by lexicographers in 2009. In total, 2,400 synsets from the Base Concept set were edited. Updates included synonyms revision and definition editing. Total number of synsets is the same (28,201). This version of Czech wordnet was not published publicly, but is available for research.

2.3 Extended with Bilingual dictionary

To increase coverage of the Czech Wordnet, semi-automatic method was proposed in 2011 [6]. We acquired machine-readable data from the largest one-volume English-Czech dictionary ever published. It contains more than 100,000 headwords and sub-headwords, more than 200,000 words and phrases and roughly 400,000 equivalents. We used the following algorithm to add new words and synsets:

- Extract translation pairs from the dictionary.
- Keep only pairs in which English literals are monosemous.
- If desired, keep only pairs with unique source literals (one-to-one translations).
- Match English literals with monosemous PWN literals.

The extended version of the Czech Wordnet contains 83,769 literals (growth by 76 %) organized into 40,621 synsets (growth by 43 %). Out of the synsets, 27,658 are noun synsets (increase by 6640, or 31.6 %), 5852 are verb synsets (increase by 690, or 13.3 %), 5651 are adjective synsets (increase by 3522, or 165.4 %) and 1457 are adverb synsets (increase by 1291, or 877.7 %).

Because of unsupervised nature of the extension, the newly produced Czech Wordnet data need to be inspected manually. We have checked a sample of 600 synsets, with the results that 30 % of the synsets contain wrong or unwanted synonyms, and 20 % of the newly created synsets are connected to an incorrect hypernym. For this reason, extended Czech Wordnet will not published publicly before the thorough editing, but is available for research.

2.4 Connection to Verbalex

VerbaLex [7] is a large lexical database of Czech verb valency frames and has been under development at NLPC since 2005. The organization of lexical data in VerbaLex is derived from the wordnet structure and entries follows the form of synsets. The current version of VerbaLex contains 6,360 synsets, 21,193 verb senses, 10,482 verb lemmata and 19,556 valency frames. When possible, the synset from VerbaLex is linked to its equivalent in Princeton Wordnet. Out of the total number, 3,725 synsets have English equivalent, remaining 2,635 are verbs specific for the Czech language.

2.5 Added definitions

Because a lot of synsets in the Czech Wordnet is missing definitions, students of the linguistics course at the Faculty of Arts were asked to update the missing parts. Czech definitions were written for 5,676 synsets from the Base Concepts set, consulting both Princeton Wordnet definitions and Czech explanatory dictionaries. These revisions are currently only saved in text files and were not inserted into Czech Wordnet.

3 DEBVisDic integration with Open Multilingual Wordnet

Since the Balkanet project, NLPC is developing browser and editor for wordnet-like lexical databases – VisDic, later reimplemented as DEBVisDic [8]. The editor is storing wordnet data in the XML format, thus making the wordnet-like databases more standard and exchangeable. Current DEBVisDic version is based on the DEB platform, general lexicographic platform, based on client-server architecture and adaptable for wide range of dictionary projects.

DEBVisDic is available as a web application and offers various features for wordnet browsing and editing. Users may work with several wordnets at once, utilizing linking and referencing between dictionaries. The application allows any user to create a new wordnet, without any complicated set-up, and start editing in a few minutes [9]. To promote wordnet sharing, DEBVisDic supports export to the Wordnet-LMF [10] format.

As the part of preparation of new version of Czech Wordnet, DEBVisDic editor will be updated to integrate better with Open Multilingual Wordnet [11] repository. Users will be able to easily connect synsets to the Collaborative Interlingual Index [12] and upload data to OMW repository directly from the DEBVisDic.

4 Open Czech Wordnet

Main impulse to speed up the creation of new version of Czech Wordnet was the proposal of integrating all available wordnets in the Global WordNet Association repository with Collaborative Interlingual Index. However, current Czech Wordnet is not published under open license. Another important motivation is the need to fix various linguistic issues that make it harder to use Czech Wordnet data in NLP applications.

We have decided to evaluate and combine all the available updates and extensions to Czech Wordnet. NLPC team has compiled the following roadmap that will lead to the publication of Open Czech Wordnet:

- Start with 2009 Edited version and combine it with definitions created for Base Concepts.

– Check synonyms present in synsets, remove unnecessary synonyms and add missing words.
– Revise or create definitions where missing. Join or split synsets to follow word senses used in Czech language, where necessary.
– Verify all types of relations between synsets semi-automatically and fix broken relations.
– Link Czech synsets to their equivalents in Princeton Wordnet 3.1 and to Collaborative Interlingual Index.

We plan to include extensions from the semi-automatically translated Czech Wordnet, but the data have to be evaluated by lexicographers first. Evaluation is planned at the end of year 2017.

It was not yet decided, in which way to include VerbaLex data. However, the best option for the wordnet composition is to create new synsets based on the VerbaLex entries, including only the synonyms and definition to the wordnet data and linking to the VerbaLex for full verb valency information. VerbaLex does not contain relations between synsets, thus hyperonymy and troponymy relations have to be set in the wordnet.

We are building the tool to allow any user of Open Czech Wordnet to submit suggestions and comments regarding wordnet data. All suggestions will be reviewed by linguists and if approved, the data will be automatically updated. We believe this tool will help to improve the quality of Open Czech Wordnet data.

5 Conclusions and Future Work

Editing work is already under way with the plan to finish the first phase in summer 2017. We plan to release the version of Czech Wordnet linked to Collaborative Interlingual Index in 2018 under open license and then continue with the evaluation of translated data. Depending on the funding and resources available, we plan to expand Czech Wordnet and reach the coverage of Princeton Wordnet.

Acknowledgments

This work has been partly supported by the Grant Agency of CR within the project 15-13277S by the Ministry of Education of CR within the national COST-CZ project LD15066.

References

1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
2. Vossen, P., ed.: EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer (1998)
3. Christodoulakis, D.: Balkanet Final Report, University of Patras, DBLAB (2004) No. IST-2000-29388.

4. Pala, K., Smrž, P.: Building Czech Wordnet. Romanian Journal of Information Science and Technology **7**(1-2) (2004) 79–88
5. Filipec, J., et al.: Slovnk spisovn etiny (SS). 1^{st} edn. Academia, Praha (1995) elektronick verze, LEDA, Praha.
6. Blahuš, M., Pala, K.: Extending Czech WordNet using a bilingual dictionary. In Fellbaum, C., Vossen, P., eds.: 6th International Global Wordnet Conference Proceedings, Matsue, Japan, Toyohashi University of Technology (2012) 50–55
7. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Proceedings of the Slovko Conference, Bratislava, Slovakia (2005)
8. Rambousek, A., Hrušo, T.: Web Application for Semantic Network Editing. In: RASLAN 2013: Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Czech Republic, Tribun EU (2013) 13–19
9. Rambousek, A., Horák, A.: DEBVisDic: Instant Wordnet Building. In Barbu Mititelu, V., Forascu, C., Fellbaum, C., Vossen, P., eds.: Proceedings of the Eighth Global WordNet Conference, Bucharest, Romania, Romanian Academy (2016) 317–321
10. Soria, C., Monachini, M., Vossen, P.: Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In: Proceedings of IWIC2009, New York, ACM Press (2009)
11. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: ACL (1), The Association for Computer Linguistics (2013) 1352–1362
12. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: CILI: the Collaborative Interlingual Index. In Barbu Mititelu, V., Forascu, C., Fellbaum, C., Vossen, P., eds.: Proceedings of the Eighth Global WordNet Conference, Romanian Academy (2016) 50–57