Management and Publishing of Multimedia Dictionary of the Czech Sign Language

Adam Rambousek, Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{rambousek,hales}@fi.muni.cz

Abstract. This paper describes the development of a multimedia dictionary writing system for the Czech Sign Language dictionary, prepared in cooperation of several institutions dedicated to the sign language research and study. The presented dictionary system takes the advantage of electronic format and strongly relies on the use of multimedia evidences. The dictionary system has to deal with large amount of video recordings of both a human narrator and digital avatar. Since the dictionary is prepared by several remotely located groups, the dictionary system provides support for complex publishing processes. The dictionary writing system is used in active preparation of the Czech Sign Language dictionary and already serves as a publisher for the dictionary data to the general public.
Keywords: dictionary writing system, sign language, multimedia dictionary, DEB platform

1 Introduction

Preparation of a new dictionary of the Czech Sign Language is one of the key activities of the *Network of Expert Centres Providing Inclusion in Tertiary Education*[1] project. The dictionary is being created in cooperation of five Czech universities and organizations dedicated to the sign language issues. The whole project is co-ordinated by the Teiresiás Centre, Masaryk University. At the current stage of the process, the aim is to create an extensive dictionary of the Czech sign language and the explanatory dictionary of the Czech language, both interconnected to serve also as a bilingual dictionary. More languages (both sign and spoken) may be added later to form a multilingual dictionary.

2 DEB Platform

Utilizing the experience from several lexicographic projects, we have designed and implemented a universal dictionary writing system that can be exploited in various lexicographic applications to build large lexical databases. The system

[1] http://www2.teiresias.muni.cz/expin/en

is called Dictionary Editor and Browser, or the DEB platform [1]. Since 2005 the DEB platform was applied in more than 10 large international research projects, see for example [2,3,4,5]. The DEB platform is based on the client-server architecture, which brings along a lot of benefits. All the data are stored on a server and a considerable part of the functionality is also implemented on the server-side, while the client application can be very lightweight. This approach provides very good tools for editor team cooperation; data modifications are immediately seen by all the users. Server also provides authentication and authorization tools.

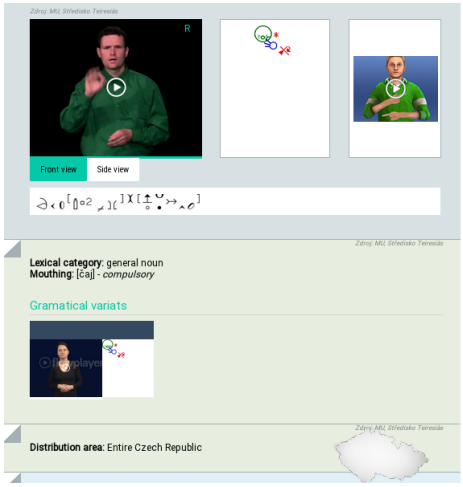

Fig. 1. Entry in the Czech Sign Language dictionary.

3 Entry Representation

Creating the dictionaries of sign languages has always been challenging in the printed form, this may change dramatically with electronic dictionaries. Kristoffersen and Troelsgard [6] have described their experience in building the Danish Sign Language Dictionary: *The overriding challenge in sign language lexicography is how to render signs in the absence of a written language. ... the use of video recordings has predominated. This approach is obviously limited to electronic dictionaries.*

The Czech Sign Language dictionary relies heavily on video recordings, they are used not just for the sign representation (front view and side view), but also for meaning explanations or usage examples. See Figure 1 for the example of an entry in the Czech Sign Language dictionary. However, the video is not the only way to present the sign to the dictionary users. The dictionary also

provides tools to enter and display two methods of formal sign transcription: *SignWriting*[7,8] (iconic script system which encodes hand shapes and movements with image symbols) and *Hamburg Notation System (HamNoSys)* [9] (transcription system similar to IPA [10]).

4 Access Control and Validations

The dictionary is produced by several teams at different institutions, each covering different parts of a dictionary entry or different domain. Users may get various roles (e.g. editor, reviewer, supervisor, ...) and the DEB platform access management module combines user role with the group, resulting in complex access control system. Furthermore, each part of the dictionary entry (e.g. grammatical information, definitions, translations, ...) have to be approved by the supervisor before it may be published for public presentation.

Validations are also needed to check the completeness and correctness of input information for each dictionary entry. First stage of validations is provided during the editing, for example the correctness of the references between the sign language and the Czech language. More validations are run after the batch of entries is handed over by a group of authors to the reviewers. For example, the reviewer may check for entries that contain video recording of the lemma and the definition, but do not contain a usage example or translations.

5 Data Import

Each institution or department involved in the project specializes on one task. For most of the entries in the initial phase, the following process was exploited. One department recorded the videos for a batch of signs. The batch was then handed over to the lexicographic team that added grammatical information, translations, etc. Meanwhile, another team provides SignWriting transcriptions. To speed up the creation of new entries and updates of existing entries, the dictionary system supports complex tools for batch import of video recordings.

The import system scans the batch of video recordings and detects the type of video (sign, definition, usage example) and checks whether the file is an update of an existing video, completely new entry, or an additional video to an existing entry. The system is able to detect whether to add a definition video to a certain entry meaning. Translations to the Czech language are proposed, linking with existing Czech entries or creating new Czech equivalents.

As for the Czech language dictionary in the project, the goal was not to create a completely new dictionary, but re-use and update existing resources. For that reason, the three large monolingual dictionaries were included into the system – see [11,12,13]. Each of the dictionaries follows different entry structure. During the import, all the information had to be normalised to the common entry structure while not losing the information. To enrich the dictionary entries for the Czech language with "real world" usage examples, the dictionary

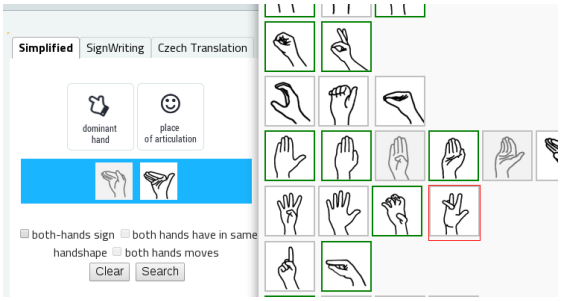

Fig. 2. Example of hand shape selection in "iconic search".

writing system provides examples extracted from the CzTenTen12 corpus [14], rated by the GDEX tool [15] to get the most suitable examples.

6 Searching

Searching in the sign language dictionary is a challenging issue. For the Czech Sign Language dictionary, several methods are provided and users may choose their preferred method. Thanks to the cross-references between sign language and spoken language dictionaries, users may search for a word in its written form and the sign language translations are displayed. Experienced users may enter the SignWriting or HamNoSys transcription.

The most convenient search method is the "iconic search". Users may select hand shapes or positions from the set of images. The selection is internally converted to the SignWriting representation. Users are able to enter even very complex queries just by clicking on a few icons. The graphical interface for this search method is inspired by the Dutch Sign Language Dictionary [6], but provides new options to exactly specify the signs (e.g. sign symmetricity, or hand positions). See the shape selection in Figure 2.

7 Current Results and Future Work

The dictionary writing system was successfully implemented and tested. The system is used for the dictionary editing on a daily basis. Already approved parts of the dictionary are publicly accessible at the following link `http://www.dictio.info`. Currently, the dictionary contains 11,477 entries for the Czech Sign Language, with 26,040 video recordings, and 121,354 entries for Czech. After the final public release of the Czech dictionary, multiple languages will be included to the dictionary.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Horák, A., Rambousek, A.: DEB Platform Deployment – Current Applications. In: RASLAN 2007: Recent Advances in Slavonic Natural Language Processing, Brno, Czech Republic, Masaryk University (2007) 3–11
2. Horák, A., Rambousek, A.: PRALED – A New Kind of Lexicographic Workstation. In Przepiórkowski, A., Piasecki, M., Jassem, K., Fuglewicz, P., eds.: Computational Linguistics: Applications. Springer (2013) 131–141
3. Horák, A., Vossen, P., Rambousek, A.: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In: Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing, Haifa, Israel, Springer-Verlag (2008) 1–15
4. Maarouf, I.E., Bradbury, J., Baisa, V., Hanks, P.: Disambiguating verbs by collocation: Corpus lexicography meets natural language processing. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014)
5. Hanks, P., Coates, R., McClure, P.: Methods for Studying the Origins and History of Family Names in Britain. In: Facts and Findings on Personal Names: Some European Examples, Uppsala, Acta Academiae Regiae Scientiarum Upsaliensis (2011) 37–58
6. Kristoffersen, J.H., Troelsgård, T.: The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In Granger, S., Paquot, M., eds.: Electronic Lexicography. Oxford University Press, Oxford (2012) 293–315
7. Sutton, V.: SignWriting Basics. Center for Sutton Movement Writing, Incorporated (2009)
8. Kato, M.: A Study of Notation and Sign Writing Systems for the Deaf. Intercultural Communication Studies **17**(4) (2008) 97–114
9. Hanke, T.: HamNoSys-representing sign language data in language resources and language processing contexts. In Streiter, O., Vettori, C., eds.: LREC 2004 Workshop proceedings: Representation and processing of sign languages, Paris, ELRA (2004) 1–6
10. International Phonetic Association: Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
11. Filipec, J., et al.: Slovník spisovné češtiny (SSČ). 1^{st} edn. Academia, Praha (1995) elektronická verze, LEDA, Praha.
12. Martincová, O., et al.: Nová slova v češtině 1. Academia, Praha (2004)
13. Kraus, J., Petráčková, V., et al.: Akademický slovník cizích slov (SCS). Academia, Praha (1999) elektronická verze, LEDA, Praha.
14. Suchomel, V.: Recent Czech Web Corpora. In Horák, A., Rychlý, P., eds.: 6th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2012) 77–83
15. Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M., McAdam, K.: GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the XIII EURALEX International Congress, Barcelona, Institut Universitari de Lingüística Aplicada (2008) 425–432