CZECH EXPLANATORY DICTIONARY AND ITS COMPUTER IMPLEMENTATION

Václav Matoušek and Jana Michalicová and Roman Mouček

University of West Bohemia in Plzeň (Pilsen), Czech Republic
matousek | moucek@kiv.zcu.cz

Abstract: This article deals with the overview and splitting of the dictionaries used in several language engineering applications, analysis of the contemporary attempt to the creation of on-line explanatory language dictionaries, analysis of the content of keyword paragraphs of explanatory dialectic language dictionary, and keynote issues of the implementation of these kinds of dictionary.

1 Introduction

The objective of the Czech Explanatory Dictionary is to give full explanations of the meaning of general terms chosen for their importance and complexity from the point of merging Czech spontaneous speech. This requires a full description of the underlying concepts, going beyond a normal dictionary definition. Often linguistic barriers lead to problems in obtaining a common understanding of terminology at international level and between disciplines. The explanatory definitions should help to break down such barriers. Therefore the people and mainly the students of the Faculty of Applied Sciences of the University of West Bohemia in Pilsen made some experiments to create general on-line explanatory language dictionaries based on the analysis of the content of keyword paragraphs of explanatory dialectic language dictionary and keynote issues of the implementation of these kinds of dictionary.

Our first Explanatory Dictionary contains and explains more than 20,000 Czech words in this time, and it is intended both for a wide range of the readers, and for the linguists. In a word article the brief explanations of meaning of a word, examples of its use in language, its terminological and phraseological combinations are given. They are widely used in the dictionary as illustrations aphorisms, proverbs, and proverbial phrases; in necessary cases their meanings are explained. The basic grammatical forms of a word are presented, the stylistic marks are resulted which specify sphere of its use. The accent is also underlined. In a word article the derivative forms of word are completely given. The dictionary consists of two main sets of terms which will be explained below.

1.1 Dictionaries

A lot of kinds of particular dictionaries have been created and used in various fields of language science. Etymological dictionaries deal with the word origin, word variations, word formation methods and coherency of words within various languages. Dialectic dictionaries introduce dialect of specific areas. They cover the whole area lexicon or describe only words different from official language. Phraseological dictionaries explain the meaning of phrases and terminological dictionaries gather the vocabulary of the specific field of knowledge. Authorial dictionary introduces an author's vocabulary. Historical dictionaries describe an origin and evolution of specific vocabulary and they usually include also topographical-historical dictionaries containing lists of historical personalities and history of towns, castles etc. The other special kinds of dictionaries include e.g. the homonymic and antonymic dictionaries, slangy, frequency, orthographical, orthoepical, pictorial, normative, informative, and thesaurical dictionaries. This article is focused on the analysis, description and implementation of our On-line Explanatory Dialectic Dictionary.

1.2 Corpora

Language corpora as the wide collections of natural language records have a great significance for the computer processing of written or spoken natural language. The special corpora are focused on the specific domain according to defined criteria (they can cover e.g. dialect vocabulary). General corpus includes any possible natural language record. The spoken corpora are usually used in computerized dialogue systems as the training set for next recognition of user utterance. The written corpora then often serve as the sources for text mining methods and text classification.

2 Contemporary Approach to Creation of Explanatory Dialectic Dictionaries

2.1 Data Collection

The initial and crucial part during creation of explanatory dialectic dictionary is process of data collection [3]. Corpus of dictionary is usually obtained using following methods:

- a question-form (written corpus);
- an interview with people, who actively use collected corpora (finally written corpus);
- audio recording of spontaneous utterances of interlocutor (spoken corpus) etc.

Then the selection of language material is performed to keep the language authenticity and to ensure the representative sample for the next linguistic elaboration. The problems usually arise if the context specifying the selected language material is left out.

2.2 Creation of Keyword Paragraph

The keywords of collected corpus are then analyzed and the keyword paragraphs are created. The keyword paragraph usually contains: keyword; keyword identifier; lemma; grammatical information; linguistic information – region, style, time, imagery, etc.; syntactic complements; meaning definition; interpretation – standard form of keyword; compound terms; subkeyword; keyword frequency; derived keywords; cross references; synonyms and comments.

The processing of keyword paragraphs is connected with a wide range of linguistic problems, which have to be solved (e.g. lemmatization – keyword is completed with its variants). There is also necessity to divide the context from the meaning interpretation.

2.3. Creation of On-line Dictionaries

The contribution of on-line dictionaries includes e.g. inner hypertext connections, possibility of multimedia presentation, different access rights for its users and comfortable process of adding, editing and searching data. The other requirements on on-line dictionaries include also data protection, support of transactions, authorization support, quality of data presentation etc.

The number of existing professional (it means created by linguists) explanatory dialectic Czech on-line dictionaries is very low. Moreover, they suffer from simple and non-interactive computer implementation. On the other side, there is a large number of on-line explanatory dialectic dictionaries of high technical quality created by nonprofessionals in linguistic field of science. The examples of both categories mentioned above include e.g.:

- Thematic differential dictionary of Ratibořice dialect [10];
- Dictionary of dialects of Kobylí and its surroundings [11];

- The first internet Moravian-Czech dictionary [9];
- Dictionary of Pilsen surroundings [1].

3 Development of New On-line Explanatory Dialectic Dictionary

The proposal and implementation of new On-line Explanatory Dialectic Dictionary for Czech dialects [4] is one of the results of long-term work of Genius loci team prepared by a group of people from different field of science – members of University of West Bohemia, Pilsen, University of South Bohemia, České Budějovice, and State Scientific Library in Pilsen. The final result includes among other things many years of language material collection, analysis of keyword paragraph and finally design and implementation of on-line application.

3.1 Analysis of Keyword Paragraph

The content of keyword paragraph was initially determined on the base of keyword paragraph analysis of the textbook Doudlebské nářečí a slovník [2]. The content and organization of keyword paragraph was also compared to other explanatory dictionaries [4]. Because the organization of keyword paragraph is different in analysed dictionaries, they served as a wide basis for proposal of general keyword paragraph organization.

The most important difficulties achieving a general keyword paragraph organization are connected with polysemy. The structure of keyword paragraph is usually different for keywords with only one meaning and for keywords with more than one meaning. There is a question where to position the prospective second, third, etc. keyword meaning (and herewith its exemplification and its context information) within the keyword paragraph. Finally, we decided to position all the information referring to the keyword meaning (e.g. territory of occurrence, exemplification, context information, usage, audio record and information about interlocutor and its place of living) straight after meaning description. This solution restrains the text numbering and text amount. The sequence of keyword meanings depends on the keyword submitter. If the keyword can exist as more than one part of speech, then the parts of speech are numbered with the Roman numbers. The other information about keyword paragraph analysis you can find in the work of J. Michalicová [5].

3.2 Structure of Keyword Paragraph

The final structure of keyword paragraph in the form used in on-line application is presented in Fig. 1 (see the next page). The final structure of printed output does not include the parts audio, interlocutor and place of living. The typographical layout corresponds to the most used layouts of explanatory dialectic dictionaries. The page is divided into two columns. Keywords and corresponding lexicographical and grammatical information are structured into keyword paragraphs.

The typographical layout of keyword paragraph is described in detail in work of J. Michalicová [5]. Alphabetization of key words observes the Czech alphabet. Capitals and small letters are not distinguished. Phrases are included in the corresponding keyword paragraph. The sequence of the keyword meanings depends on the submitter of the key-word (the frequency sequence is considered).

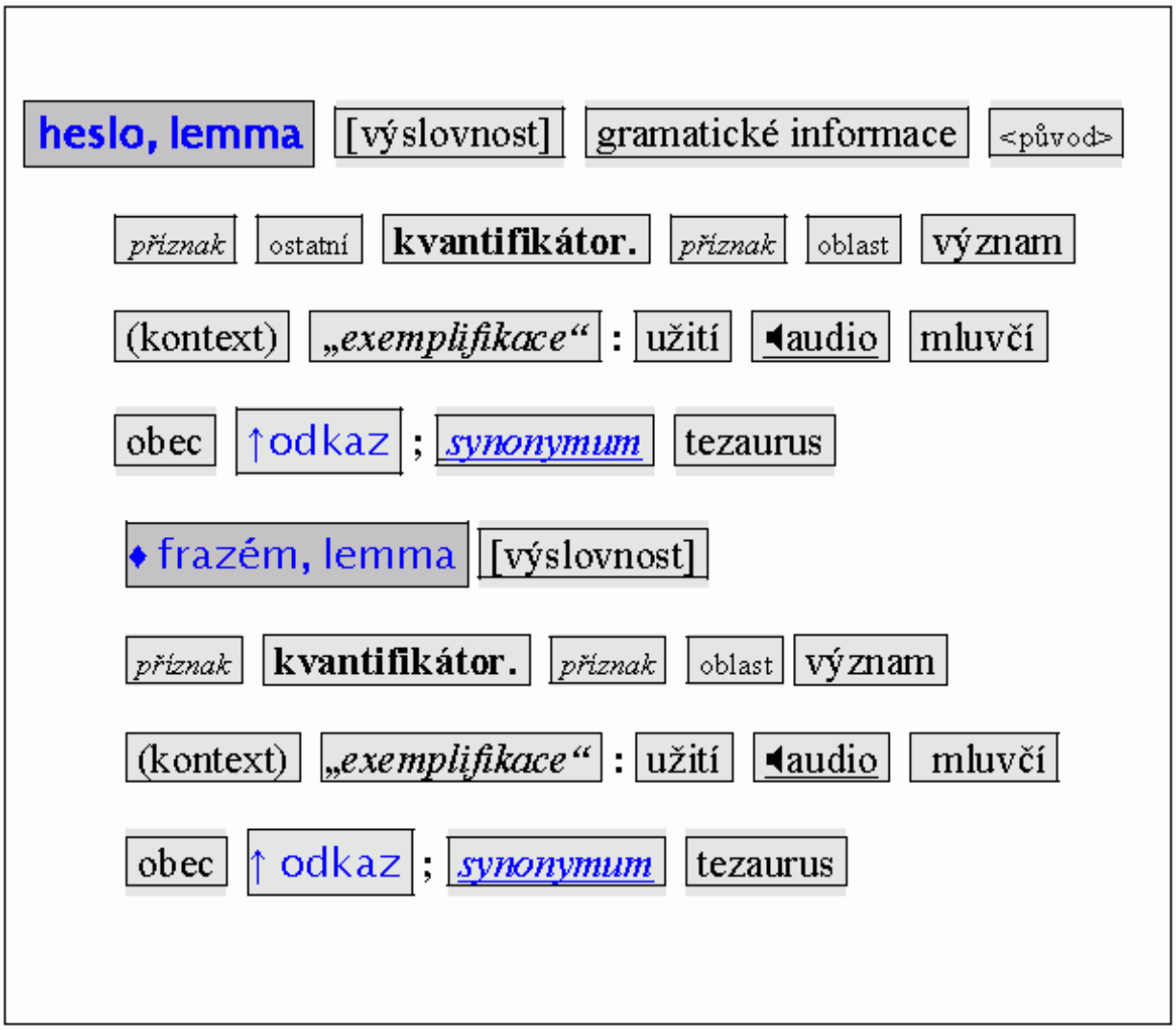

Figure 1 - Keyword paragraph in on-line dictionary

4 Explanatory Dialectic Dictionary and its Computer Model

The following ERA model (Fig. 2) represents all the information contained in On-line Explanatory Dialectic dictionary. Synonyms are stored as attributes of the keyword meaning. This approach involves difficulties with synonym search. On the other side, we avoid the necessity at first to add all the synonyms as the keywords to the database and then to link them mutually. Also the thesaurus is proposed as the attribute of keyword meaning and not as the attribute of the keyword itself.

The final computer application is based on modified standard 3-layer architecture. Fig. 3 introduces the main window of developed on-line application with several examples of keyword paragraphs. The software tools for this implementation were developed only experimentally with the aim to verify the functional properties of the developed dictionary computer model.

5 Conclusion

This paper has introduced a proposal of Explanatory Dialectic Dictionary organization and consequent on-line dictionary application. Presently the application is tested by linguists. We consider that the application becomes a comfortable and effective tool for building and storage of dialectic explanatory dictionaries.

The explanatory dictionary will be extended for a lot of new words in the next time. There are working several students of fourth and fifth courses and they test the developed dictionary on several practical tasks.

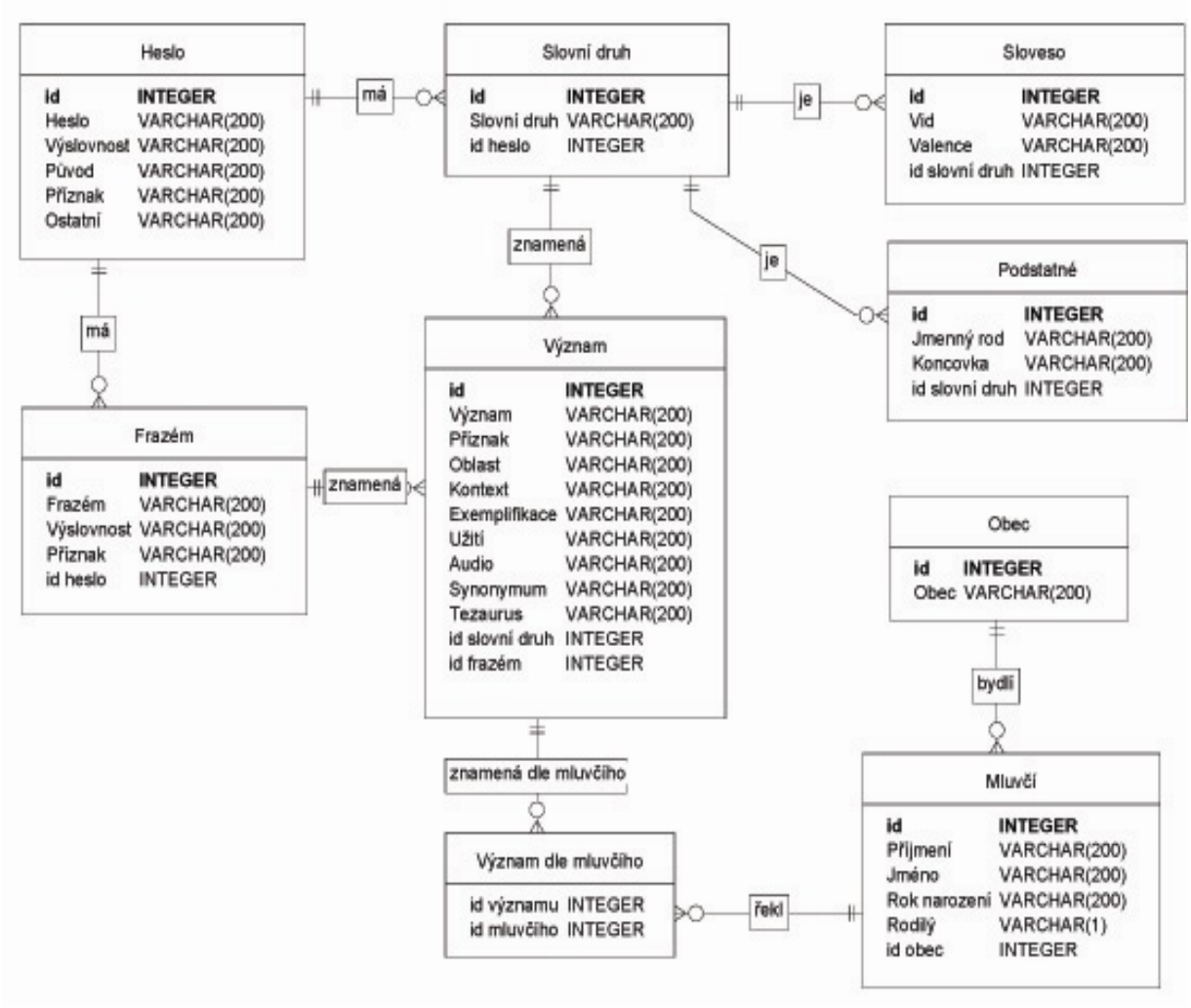

Figure 2 - ERA model of Explanatory Dialectic Dictionary

Figure 3 - On-line application of Explanatory Dialectic Dictionary

Acknowledgement

The presented work was developed within the framework of the solving of the National Research Task No. 2C06009 "Complex knowledge base tools for natural language communication with the semantic Web".

References

[1] Blail, L.: Tvorba multimediálního slovníku v prostředí databázového systému Oracle. Diploma thesis, KIV FAV Pilsen, 2005
[2] Holub, Z.: Doudlebské nářečí a slovník. Univerzita České Budějovice, 2004
[3] Holub, Z.: Lexicon nejjižnějšího úseku českých nářečí. Pelhřimov, 2003
[4] Martincová, O. et al.: Nová slova v češtině. Prague, 1998
[5] Michalicová, J.: Databázový model a www rozhraní výkladového slovníku. Diploma thesis, University of West Bohemia in Pilsen, 2006.
[6] Petráčková, V., Kraus, J. et al.: Akademický slovník cizích slov. Prague, 2001
[7] Rejžek, J.: Český etymologický slovník. Český Těšín, 2001
[8] http://dione.zcu.cz/~janamich/
[9] http://morce.slovniky.org
[10] http://www.etf.cuni.cz/~Emiksik/texty/dialekty.htm
[11] http://www.sepl.rulez.cz/slovniky/narkob/kobyli.htm