Slovene Lexical Database

Polona Gantar[1] and Simon Krek[2]

[1] Scientific Research Centre of the Slovenian Academy of Sciences and Arts
[2] Amebis, d.o.o., Kamnik; Jožef Stefan Institute, Slovenia

Abstract. The paper describes the concept of the new Slovene lexical database which is compiled within the "Communication in Slovene" project. The database has a twofold goal: it is intended as the basis for the future compilation of different dictionaries of Slovene, both monolingual and bilingual, and as such its concept is biased towards lexicography. Secondly, it will be used for the enhancement of natural language processing tools for Slovene. The database is organized in six hierarchical levels with lexico-grammatical information which spans from simple morphological data on the top level to semantic, syntactic and collocational data on subordinate levels, with corpus examples at the bottom. Sketch Engine tool with word sketch, tickbox lexicography and GDEX modules is used to enable faster and more efficient extraction of corpus data from the 620-million word FidaPLUS corpus which is used as the source for the data in the database.

1 The "Communication in Slovene" project

Slovene Lexical Database is one of the results of the "Communication in Slovene"[1] project which started in 2008 in will end in December 2013. Other results include: (a) natural language processing tools and resources for Slovene: a statistical tagger and parser with a training corpus and an extensive lexicon with information about word inflection and derivation; (b) language data resources: a billion word written corpus and a million word spoken corpus; (c) a study on language teaching practices in Slovene schools which includes the compilation of a corpus of school essays with teachers' revisions and the analysis of common problems in text production, and (d) language description resources which include a manual of style for writers, a pedagogical corpus-based grammar and the lexical database, all of them freely available in an interactive web portal.

A new web format for language data is considered which will incorporate traditional dictionary information on words and word combinations (senses, collocations, examples, grammatical information etc.), visualization of corpus data and semantic ontologies, real-time exploration of web data, question-answering system etc. The portal is intended both for school population and for general use. Information from

[1] The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational programme Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013. Project web page: http://www.slovenscina.eu/.

the lexical database will be used in two different contexts: (a) together with the lexicon and other resources it will be integrated in the portal for automatic generation of answers to questions such as "how is this word declined/conjugated; what does it mean; how do I spell it", and (b) it will be used as a lexico-grammatical resource to be used in natural language tools for Slovene.

2 The structure of data in the lexical database

With regard to different user needs, there are two types of information in the Slovene lexical database. First, lexico-grammatical information will be used for different functions of the portal and intended for human end users, such as sense descriptions in the s.c. semantic frames, representing the starting point for whole sentence definitions, collocations attributed to particular senses of the lemma, and typical examples from the corpus. Second, different types of information are designed for natural language processing tools. These are encoded in a more complex way and – in addition to their immediate use in NLP tools – need an expert to process or interpret them. Among them are the formal encoding of syntactic patterns on the phrasal and clause level as well as the formal encoding of semantic arguments and their types. The database is conceptualized as a network of interrelated lexico-grammatical information on six hierarchical levels with the semantic level functioning as the organizing level for the subordinate levels.

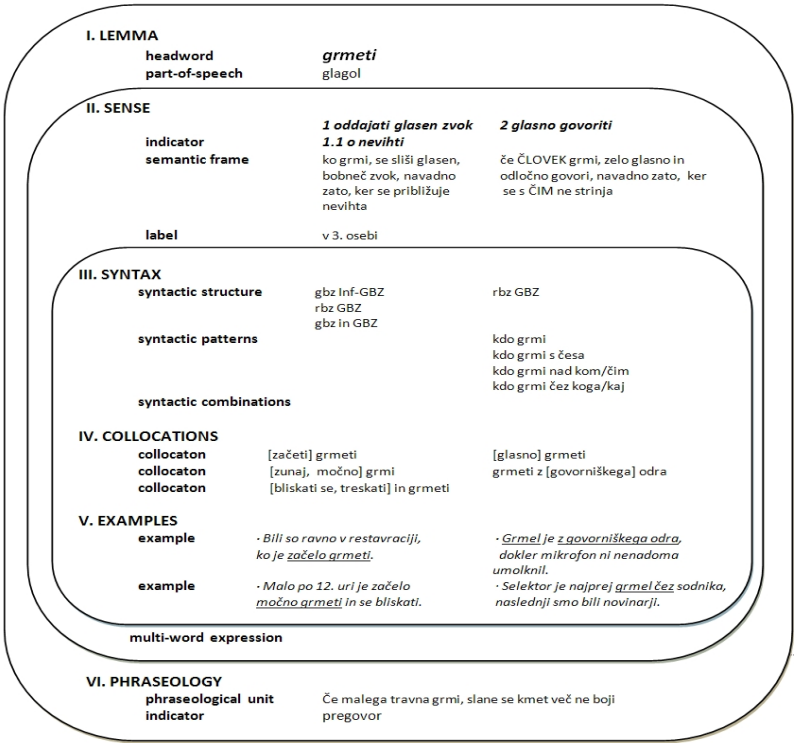

Fig.1. Structure of data in the lexical database

2.1 Lemma

Lemma – or the headword – represents the top hierarchical level and functions as the umbrella for all lexical units placed under it: senses and subsenses, multi-word expressions and phraseological units. Multi-word expressions are recorded only for nouns and adjectives and placed within particular senses and subsenses. Phraseological units and recorded in a separate section outside the sense or subsense structure. Both multi-word expressions and phraseological units can be given headword status if they show complex semantic structure and/or high frequency in the corpus. On the lemma level each headword is classified as pertaining to one of four parts-of-speech: noun, verb, adjective or adverb. Function words classes are not recorded in the lexical database as entries. Lemma is considered as linked to its inflectional paradigm in the lexicon, therefore word class conversions (e.g. noun-adjective or vice versa) are analyzed on the sense level and not as new entries in the database.

2.2 Sense/subsense

On the sense level, senses and subsenses of the lemma are specified. Therefore, a two-level hierarchy is allowed for with the possible role of the upper level to function as an empty category subsuming the subsenses pertaining to a common semantic field, as in the case of the lemma “grmeti” (to thunder) in Figure 2. All senses and subsenses are labelled with semantic indicators whose primary function is to form a sense menu intended for easy navigation within a polysemic entry structure.

Fig.2. Sense menu of the verb “grmeti” (to thunder)

Another kind of information recorded on the sense level are semantic frames which are conceptually close to frames in the FrameNet project [1] [2] and to prototypical syntagmatic patterns in the Corpus Pattern Analysis system [3]. With verbs, as well as some nouns and adjectives, semantic frames are used to record argument structure and semantic types found in a particular sense or subsense. Therefore, semantic frames provide a link between a particular sense of the headword made explicit by semantic indicators, and syntactic conditions for its realization. At the same time, they represent the starting point for creating the whole-sentence definitions similar to the ones found in Cobuild dictionaries [4] [5].

Fig.3. Semantic frames from a verb, noun and adjective entry

Whole-sentence definitions in the form of if-clauses include information about typical syntactic patterns (lemma “aktiven” – active: /predicative use/ a PERSON is active in a FIELD or an ACTIVITY if he/she is participating in it on a regular basis; /attributive use/ an active DEVICE or COMPUTER PROGRAM is ready to function immediately), reflexivity (lemma “briti” – to shave: if a PERSON shaves his/her HAIR |or| if a PERSON shaves another PERSON's HAIR, he/she removes them with a DEVICE), pragmatic aspects of headword usage (lemma “lahkomiseln” – ≈loose: a WOMAN is considered loose if she likes to be in the company of different men or frequently changes partners), or grammatical limitations (lemma “grmeti” – to thunder: /only in 3rd pers. sing./ if it thunders in an AREA or if WEAPONS thunder, shooting can be heard). Semantic types are linked to other kinds of information on subordinate levels thus enabling the user to access data at different level of abstraction, from natural and explicit corpus contexts to implicit semantic types.

2.3 Multi-word expressions

Multi-word expressions are registered only in the entries with noun or adjective headwords, either within a particular sense/subsense or after all registered senses and subsenses if semantic relation cannot be established between the MWE and one of the sub/senses. Multi-word expressions must demonstrate a non-compositional idiosyncratic sense, again described by a semantic indicator, mostly identifying the rather broad semantic field or domain. MWUs can show variant forms which are listed under the same MWU entry section and can have their own collocations.

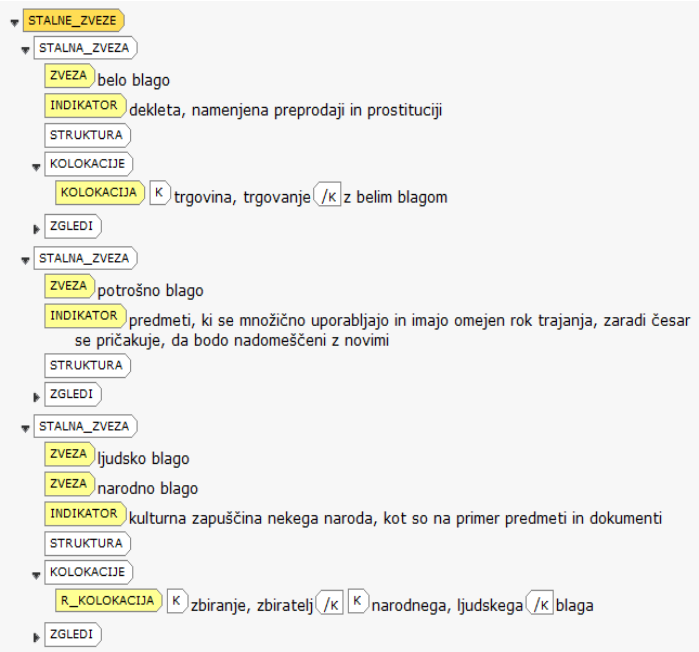

Fig.4. Multi-word expressions with other kinds of information in the DPS Entry Editor software

2.4 Syntactic structures

Clause patterns: a degree of syntactic information in the form of patterns is already present in the clause structure within semantic frames, e.g. for the verb “sesti” (to sit) in the sense “to agree with, to suit” the pattern “kaj ustreza komu” (sth agrees with sb) is registered, for the noun “nota” (a note) in the sense of “characteristics” the pattern “kaj daje noto čemu” (sth gives a particular note to sth) is registered, and for the adjective “pozoren” in the sense “attentive, tender, caring” the pattern “kdo je pozoren do koga” (sb is attentive to sb) is registered. On syntactic level in LBS, clause patterns are registered systematically with pronouns (sth, sb) in place of semantic arguments to account for alternations of the prototypical pattern in the manner of the theory of norms and exploitations [6]. Alternations include cases where particular arguments are realized by different syntactic possibilities (prepositional phrases, subordinate clauses etc.) or not at all (as in inherent arguments, e.g. in the case of the verb “dihati” – to breathe, “air” is the inherent argument since is it rarely expressed as the object: to breathe air). Clause patterns represent useful information for grammar writing and for teaching of Slovene as foreign language, and they enable automated transition from prototypical patterns in semantic frames to typical alternations of these patterns.

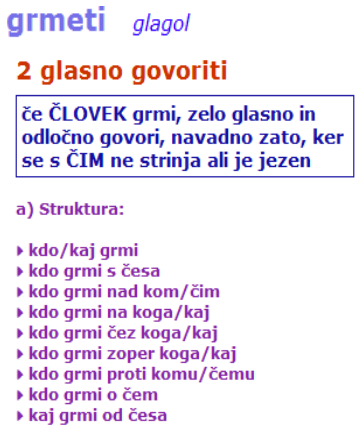

Fig.5. Clause patterns of the verb “grmeti” in the sense “to talk loudly”

Syntactic structures and collocations: syntactic structures represent a formalization of typical patterns on the clause and phrasal level and are primarily intended for natural language processing tools. They are registered in the form of sytagmatic combinations of words and phrases, and are composed of a part-of-speech label plus the information on grammatical case. Formally, labels are conformant with morpho-syntactic tags used in the FidaPLUS/Gigafida corpus [7] in the Sketch Engine tool [8]. Listed types of syntactic structures are partly dependant on the part-of-speech of the headword. Within a particular structure, the position of the headword is indicated by the capitalization of the label. Where typical collocates realizing a collocation for a particular syntactic structure exist, they are registered under that structure. With verbs, syntactic structures include verbal phrases with infinitives, adverbs or coordinating structures among others, as shown in Figure 6.

Fig.6. Syntactic structures and collocations of the verb “grmeti” in the sense “to talk loudly”

The source for extracting syntactic structures and collocations from the corpus are s.c. word sketches in the Sketch Engine tool described below. The number of syntactic structures is finite – at the time of writing almost 300 structures are recorded, however, not all of them exhibit collocations, as in the case of the structure /pbz0 SBZ0/ (adjectival phrase + nominal phrase, both with the whole inflectional paradigm) shown in Figure 6.

2.5 Syntactic combinations

Syntactic combinations represent an intermediate level between collocations and multi-word expressions. The most typical members of the class are prepositional phrases and other multi-word combinations which extend beyond the binary syntactic structures, but are on the other hand compositionally fixed and have at least one invariable lexical element, together with another lexically variable but syntactically obligatory element. Syntactic combinations also include elements with numerical expressions, comparisons and coordinate structures. Contrary to multi-word expressions, syntactic combinations do not need explanation in semantic terms (their meaning is compositional) and therefore no indicators or semantic frames are provided. Consequently, they cannot be given headword status.

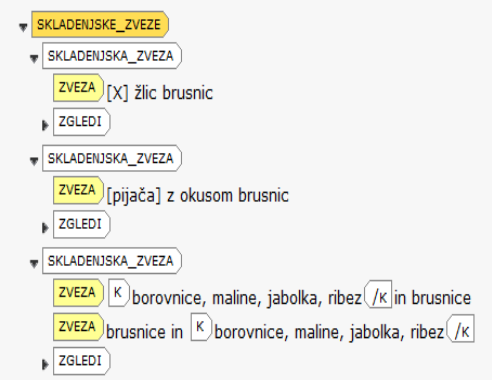

Fig.7. Syntactic combinations under the headword “brusnica” (cranberry) in the “fruit” sense

2.5 Collocations and examples

On the collocation level, patterns and structures are verified by recording typical collocates of the headword realized in the anticipated syntactic positions. At the same time, higher levels in the hierarchy are taken into account and the same collocation can be recorded under different syntactic structures or senses if it shows semantic diversity. On the last level in the hierarchy, collocations and also all parent levels (patterns, structures and frames with semantic types) are verified by recording corpus examples using GDEX [9] and TBL tools [10] in the Sketch Engine.

3 Corpus data and tools

3.1 The corpus

At the time of writing, corpus used for the compilation of the lexical database is the FidaPLUS corpus [11] with 620-million words and containing text from 1990–2006. In 2011 it will be replaced by another generation of the corpus line which started in 2000 with the 100-million word FIDA corpus with restricted access, which was upgraded and made publicly available as FidaPLUS corpus in 2006. The new corpus is called Gigafida and contains 1.1 billion words from texts of different genres spanning from 1990–2010. Its composition and characteristics are described in [7]. Together with the new text data, a new web interface was developed with a particular focus on user friendliness and ease of access to the data for non-expert users. However, for the purpose of lexical database compilation, the FidaPLUS corpus was put into the Sketch Engine tool and this version of the corpus is used by the lexicographers, making use of advanced corpus query features provided by the tool.

3.2 Sketch Engine

Sketch Engine represents the basic lexicographic corpus data extraction tool used by the lexicographers compiling Slovene lexical database. Together with the standard use of a concordances with advanced options such as the use of corpus query language (CQL) and similar, two additional features are used which enable faster compilation of the database. The first is word sketches module which provides one-page automatic summaries of a word's grammatical and collocational behaviour. Word sketches are based on s.c. sketch grammar where grammatical relations are defined as regular expression over POS-tags. Slovene sketch grammar currently contains 32 grammatical relations or gramrels which basically reflect the 300 recorded syntactic structures. The other feature are the combined Tickbox lexicography and GDEX modules which provide a faster way to select good dictionary examples recorded under each structure and collocation in the database. The module described in [9] and [10] was upgraded for its use with the Slovene language.

4 Conclusions

The concept of the Slovene lexical database is biased towards lexicography but its intended use is also the enhancement of natural language processing tools for Slovene, such as taggers and parsers. Along with the use of data within the language portal for human end users, we believe that syntactic structures and patters recorded in the database will make contribution to the better quality of the parser which is also under development in the same project. This assumption will be tested at the end of the compilation process when automated extraction of data from the corpus will be tested. Further research is foreseen also with the analysis of semantic type recorded in semantic frames where an ontology could be constructed and linked to the FrameNet and/or WordNet data. Lastly, we expect that the database will be used for the first automatic word sense disambiguation experiments for the Slovene language.

References

[1] Fillmore, Ch. J., Atkins B.T.S. (1992). Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In: Lehrer, A., Feder K.E. (eds.) *Frames, Fields, and Contrasts.* Lawrence Erlbaum Associates, 75-102.

[2] Baker, C.F., Fillmore, C.J., Cronin, B. (2003). The Structure of the Framenet Database. International Journal of Lexicography, 16(3), 281-296.

[3] Hanks, P. (2004). Corpus Pattern Analysis. In: Williams, G., Vessier, S. (eds.) EURALEX 2004. *Proceedings*. Lorient: Université de Bretagne-Sud.

[4] Barnbrook, G., Sinclair, J. (1994). Parsing Cobuild Entries. In: Sinclair, J., Hoelter, M., Peters, C. (eds.) *The Languages of Definition: The Formalisation of Dictionary Definitions for Natural Language Processing.* Luxembourg. European Commission. 13-58.

[5] Barnbrook, G. (2002). *Defining Language: A Local Grammar of Definition Sentences.* Studies in Corpus Linguistics: John Benjamins Publishing Company.

[6] Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers need Prototype Theory and Vice Versa. In: F. Kiefer, G. Kiss, and J. Pajzs (eds.) *Papers in Computational Lexicography: Complex '94.* Research Institute for Linguistics, Hungarian Academy of Sciences.

[7] Logar Berginc, N., Krek, S. (2010) New Slovene corpora within the Communication in Slovene project. International Conference SLAVICORP. Corpora of Slavic Languages. 22-24 November 2010. *Abstract*, 8.

[8] Kilgariff, A., Tugwell, D., (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proceedings of the AVL workshop on COLLOCATION: Computational Extraction, analysis and Exploitation.* Toulouse. 32-28.

[9] Kilgariff, A., Husak, M., McAdam, K., Rundell, M., Rychly, P., (2008). Automatically finding good dictionary examples in a corpus. In: Bernal, E., DeCesaris, J. (eds.) *Proceedings of the XIII EURALEX International Congress. Barcelona, 15-19 July 2008.* Barcelona: Documenta Universitaria: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra, 425-432.

[10] Kilgarriff, A., Kovar, V. & Rychlý, P. (2009). Tickbox Lexicography. In: Granger, S. & Paquot, M. (eds.) *Proceedings of eLex2009: eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve, Belgium. 411-418.

[11] Arhar, Š., Gorjanc, V., Krek, S. (2007). FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools. In: *Proceedings of the Corpus Linguistics Conference, CL2007, University of Birmingham, UK, 27-30 July 2007*. Birmingham, 2007.