THE SLOVENE LEXICAL DATABASE: THE ORGANIZING PRINCIPLES OF THE ARGUMENT STRUCTURE

Polonca Kocjančič
Amebis, d. o. o., Kamnik
Petra Zaranšek
Trojina, Institute for Applied Slovene Studies

Abstract

The creation of the Slovene Lexical Database is one of the central goals of a major Slovene lexicographic and human language technologies project (Communication in Slovene: http://www.slovenscina.eu). The database should, on the one hand, serve the needs of posterior compilation of monolingual and bilingual dictionaries, and, on the other hand, be robust enough to serve further human language technologies needs.

The initial design phase of the database has been aimed at defining the guidelines and principles of the database compilation. In this period, several similar lexical database projects have been scrutinized. The paper presents a review of their features and then focuses on the semantic level of the description of the lexical unit. This level has two subcategories, namely the semantic indicator and the argument structure. The organizing principles of these two categories, especially the latter, are presented along with sample materials.

Keywords: corpus, lexical database, semantics, semantic indicator, argument structure, corpus evidence.

I. INTRODUCTION

This paper begins with a brief overview of the scope of the project within which it has been developed and then moves on to its main focus –to describe certain aspects regarding one of the larger tasks within the project, the Slovene Lexical Database.

The Communication in Slovene (*www.slovenscina.eu*) Project is a five-year lexicographic and human language technologies project which started in June 2008. There are five institutions in the project consortium, with several teams currently working on the central module, comprising the reference corpus of Slovene and the Slovene Lexical Database, together with a grammatical annotation tool. The project will also develop two other modules: (a) new didactic methods for Slovene language teaching, and (b) a pedagogical corpus-based grammar and style manual. The reference corpus will contain several million words and will also include a transcribed and linguistically annotated spoken corpus of one million words. Furthermore, manual annotation with morphosyntactic and syntactic information is underway in order to obtain a training corpus, which will serve to improve the existing tagging and parsing tools for Slovene. The nature and funding of the project require that all the results be publicly accessible without restrictions and free of charge.

II. THE SLOVENE LEXICAL DATABASE (SLD)

II.1. Basic information, timeline, and source material

The creation of the Slovene Lexical Database (hereinafter SLD) is one of the central goals of the project. Its objective is to serve the needs of subsequent compilation of monolingual dictionaries, bilingual dictionaries, and other reference works, and be robust enough to serve further human language technologies needs.

The main principles governing the creation of the SLD can be narrowed down to two points: (a) the database will be a monolingual corpus-driven lexical resource, (b) the core vocabulary of the Slovene language will be presented and described, together with semantic, syntactic, collocational, and phraseological information, supported by illustrative corpus examples.

The first phase of work on the SLD, from August to December 2008, consisted of a thorough review of various lexical databases and other lexical resources for most European languages. In November 2008 intensive work started on defining the corpus analysis procedures and establishing the standards for the compilation of different kinds of lexical units within the database[1]. The preparation of the system, style guide, and sample material is due to be completed by June 2009, and the compilation of the SLD by July2012[2].

The SLD team members are presently using the existing 620-million-word FidaPlus reference corpus of Slovene, which is available over the web through its purpose-built concordancer (*www.fidaplus.net*) as well as by means of the SketchEngine corpus query tool (*www.sketchengine.co.uk*), while the new corpus that is currently being compiled will be available in the course of the project soon enough to adequately support the core activities.

Since this paper will further focus on the treatment of semantic information and the argument structure of the entries, only those details regarding the nature and structure of the database will be presented that are considered necessary for this subject, while overall presentation of the SLD structure is a matter of other publications.

II.2. Preliminary work

During the initial phase several lexical database projects were analyzed (GENELEX, LE PAROLE, SIMPLE, ACQUILEX I, and II, ILC-DELIS, as well as language-specific Elexico, CLIPS, CORNETTO, DAFLES, ALFALEX, STO, ADESSE, Grial, CEGLEX, SPRÅKBANKEN, PRALED, etc.). This period helped the team deepen the knowledge of the various approaches used and check our ideas against those developed elsewhere, as well as to see the utility of their experience for our needs. Some of these were further reviewed in more detail, namely FrameNet, Corpus Pattern Analysis, STO, ADESSE, and Grial (Braasch, 2002; Fillmore et al., 2003; García-Miguel et al., 2005; Hanks, 2004; Vázquez et al., 2006).

Both FrameNet and Corpus Pattern Analysis have been of fundamental importance in creating a system that would suit the needs of the SLD and be adaptable to the resources currently available for corpus analysis of Slovene. For example, the way the lexical unit definition is formed in FrameNet (*http://framenet.icsi.berkeley.edu/*) is very appealing from the point of view of the description of the frame elements and the scenario. However, we did not start by creating semantic frames and subsequently ascribing lexical units into appropriate frames. Our purpose is not to create ontologies by grouping related lexical units into semantic frames, but rather to take as a starting point individual lexical units and describe different meanings of each one of them by means of FrameNet-type scenarios which include defining the range of semantic and syntactic combinatory possibilities (valencies of each word in each of its senses). The reason for this is to avoid the accumulation of possible, but not real and

corpus-proven patterns. The idea has rather been to define a set of semantic-syntactic patterns through the analysis of corpus material.

The compilation of the SLD comes close to the Corpus Pattern Analysis system (*http://nlp.fi.muni.cz/projekty/cpa/*) with regard to the following: 1) The analysis of a large number of concordance lines, which is more detailed than just a quick overview since we ascribe senses to concordance lines and group related ones together; 2) Further analysis of concordance lines and senses from the perspective of argument structure and scenario –with regard to this it must be added that we do not ascribe a detailed argument structure and scenario to each individual concordance line, but we construct a summarized argument structure and scenario description which can include a larger number of concordance lines than is the case for the Corpus Pattern Analysis system. As a result, in the SLD the argument structure and description of the scenario are also shorter and more general.

The Spanish projects ADESSE (*http://webs.uvigo.es/adesse/index.html*) and Grial (*http://grial.uab.es/index.php*), both essentially syntax-oriented, have proven helpful in clearing up issues regarding the organization of syntactic patterns, the presentation of patterns, corpus materials, and the future end results of the analysis. Nevertheless, recent work of the SLD team has been oriented also towards testing the opposite viewpoint: first creating a list of potential grammatical patterns and subsequently on evaluating the utility of each one of them based on the corpus data that they attract. This has also been instrumental in the improvement of the grammatical relations file for the SketchEngine corpus query tool (*http://www.sketchengine.co.uk/*), which is used to support the process of transforming corpus data into a concrete SLD entry.

III. THEORETICAL BACKGROUND AND ENTRY STRUCTURE

Meaning categories, sense divisions, and the concept of the cognitive perception of meaning itself have indeed long been a burning theoretical and –more importantly to lexicographers and users– practical issue: "Firstly, any working lexicographer is well aware that, every day, they are making decisions on whether to lump or split senses that are inevitably subjective: frequently, the alternative decision would have been equally valid." (Kilgarriff, 1997)

Several other linguists, such as John Sinclair, Patrick Hanks, Sue Atkins, Michael Rundell, to name just a few, have also given much thought to this subject and have urged the linguistic community to rethink these categories; they have also shared their self-reflections about the work they have been dedicated to, claiming, for example, that "the numbered lists of definitions found in dictionaries have helped to create a false picture of what really happens when language is used" (Hanks, 2008: 125).

"The corpus citations will be clustered into senses according to the purposes of whoever or whatever does the clustering. In the absence of such purposes, verb senses do not exist." (Kilgarriff, 1997) In the SLD analysis, this process can be described as an attempt to identify lexico-syntactic regularities that appear in raw corpus material. The semantic indicator and the argument structure are those elements that are meant to describe the semantic aspects that a lexical unit exhibits. In order to obtain information for the construction of these two entry levels and group similar information regarding a lexical unit, raw concordance lines are analyzed according to the process described below.

For the compilation of the SLD, we consider the principles of lexicogrammar as the most helpful starting point in the process of organizing the structure of the entries. "Drawing a clear-cut distinction between meaning and grammar is not an easy task, because the two are so intimately interwoven /…/ ultimately, the only purpose of grammar is to serve the conveyance of meaning." (Cruse, 1986: 1-2) Clearly, the compilers' wish is to present the database entries as being subject to an intertwining between semantics and syntax; in other

words, we agree with Atkins and Rundell in that "a significant shift in meaning may be encoded in an apparently trivial change in grammar" (Atkins and Rundell, 2008: 300).

With regard to the above-mentioned issues, data in the SLD is organized into various levels:

- *lexical entry level* - containing the headword and part-of-speech information,
- *semantic level* - containing the semantic indicator and the argument structure,
- *syntactic level* - containing the syntactic patterns,
- *collocational level* - containing sets of collocations,
- *corpus examples level* - containing sentence examples from the corpus, and
- *phraseology level* - containing phraseological units, corresponding semantic indicators, and corpus examples.

IV. SEMANTIC INDICATOR AND THE ARGUMENT STRUCTURE

The principles guiding the construction of the semantic indicator and the argument structure should now be described. The semantic indicator is considered to provide guidance on the meaning of a lexical unit, similar to the short definitions in the top menu of the Macmillan English Dictionary for Advanced Learners (Rundell, 2007) or those in the Longman Dictionary of Contemporary English (*http://www.ldoceonline.com/*), hereinafter MEDAL and LDOCE, respectively. In the case of semantically complex lexical units, semantic indicators are used to draw the division lines between meanings or meaning tendencies, or also to collect semantically similar concordance lines according to the surrounding context. For example, in the verb *pasti* ("to fall")[3]:

znižati se na lestvici	**to decrease on a scale**
o količini ali vrednosti	of quantity or value
o intenzivnosti	of intensity
o statusu, kakovosti	of status, quality
o socialnem položaju	of social situation

The argument structure is the category describing the scenario, i.e. the main activity, and various circumstances that affect or modify the central deed. It is a short definition containing the lexical unit in question, the arguments, the circumstances, and pragmatic information when it is necessary to draw meaning distinctions; in these cases, it is treated as part of the scenario. In real use sentences, the arguments are either lexically or syntactically expressed or tacit from the perspective of sentence boundaries. Block capitals are used to describe the obligatory arguments and the rest of the information appears in lower case; for example, in the verb *podariti* ("to give, to donate"):

ČLOVEK podari drugemu ČLOVEKU DARILO za kaj

Examples:
*Največ mi pomeni, če mi narišejo risbo in mi jo **podarijo**, ker vem, koliko truda so vložili vanjo.*
*Kaj pa ji boš **podaril** za zlato poroko?*
*Znesek je **podarila** za obnovo vrtca v Bovcu.*

What means the most to me is when they draw a picture and **give** it to me, because I know how much effort they put into it.
What are you going to **give** her for her golden wedding anniversary?
She **donated** the money for the renovation of the Bovec kindergarten.

This particular argument structure can be paraphrased as “a HUMAN gives another HUMAN a GIFT for something”. The arguments listed refer to the following: ČLOVEK (literally “human”) is the person that gives (the giver); drugemu ČLOVEKU (literally "to another human") is the person that receives (the recipient); DARILO (gift) is the object or property that is transferred from the giver to the recipient (the object given); za kaj (literally “for what”), which is optional, describes the purpose of giving, either describing an occasion or a designated future use of the object that is transferred from one possessor to another.

These two concepts will be dealt with in more detail in the following chapter, introduced by a section describing analyses of concordance lines.

V. ANALYSIS OF CORPUS MATERIAL

In the setup of the SLD system, the compilers tested various approaches to corpus analysis in order to see as clearly as possible the connection between the individual processes and the results obtained. The team’s most obvious approach has been to depart from corpus data rather than to construct a rigid system without practical grounds. The compilation criteria thus follow the general outline described above, and are being constantly refined in order to be well established when the setup phase concludes. Content-wise, the first period was dedicated to verbs, while the remaining parts-of-speech are presently being examined.

V.1. Concordance lines

First, several very frequent and lexically complex verbs were analyzed by individual compilers using the method of reading through 300 random concordance lines for each of them and setting the corresponding semantic indicators and argument structures for each of them, while subsequent analyses of data were mainly based on the use of the information obtained from the statistical calculations by the SketchEngine corpus query tool. As for the number of concordance lines that should be subject to analysis, it was decided that lexical units with the highest frequency of occurrence will have to be analyzed using a higher number of concordance lines, while less frequent and less semantically complex ones will quite probably allow for a lower number of concordance lines. Therefore, a scale ranging from 100 concordance lines for less semantically complex lexical units, to 250-300 for the bulk of the more complex ones, and up to 400 or even 500 concordance lines for semantically most complex lexical units could be used to gain a balanced picture of lexical units.

V.2. Semantic indicators

In order to show various approaches when the group was initially dealing with the composition of the semantic indicator, we will look at one of the meanings of the verb *stisniti* (“to squeeze, to press”). This lexical unit was analyzed such that each member of the group went through the same process without previous group consultation. Later the results were merged and compared. One section is shown in the Table 1 below, summing up a set of 22 concordance lines with the meaning “to press something such as a liquid out of something”.

member 1	member 2	member 3	member 4	member 5	member 6
to squash	to emit/a	pressure	to press liquid out of something	to gain by squeezing	to cause

and press liquid	to emit/b		to press liquid into something	to gain (by squeezing)	reduction by force
	to add somewhere		to press	to extract	
				to squash	
				to convert*	

Table 1: Various approaches to defining the semantic indicators in the SLD research phase

Some of the sample source concordance lines for the analysis above are presented in Table 2. In the actual analysis, more context was viewed.

Iz grozdja so	**stisnili**	70 litrov vina ...
V čaj	**stisnimo**	sok limone.
Olupimo česen in ga	**stisnemo.**	
	Stisnite	si limono, pomarančo, grenivko ...
... grozdje, ki ga bodo	**stisnili**	v mošt ...

Table 2: Sample concordance lines to illustrate the verb *stisniti* ("to squeeze") meaning "to press something such as a liquid out of something".

In about two thirds of concordance lines, the meaning discrimination, reflected in the form of semantic indicators, was common to all team members –a sign that these can be considered core meaning categories– while the rest depended more on individual criteria for discerning meaning.

Consequently and following further testing, the following general guidelines were accepted:

- a semantic indicator defines the semantic field of a certain meaning of a lexical unit
- in semantically complex lexical units, the semantic indicator should emphasize in what ways several meanings differ, rather than strive to define in detail the meaning components
- a semantic indicator describes the meaning conceptually; presenting it with a synonym should be avoided
- a semantic indicator should be short but not too general as it would then become too polysemic, leading to examples that are too different from each other to fall into the same category (which is reflected in much vacillation in the use of inclusion/exclusion criteria for concordance lines among various team members)

V.3. The verb: argument structures

The role of the argument structures in an SLD entry is the identification of the scenario and of the arguments that are required for a verbal act to take place –or, to put it in a much simplified

manner, it has to define what is happening to whom when a certain verb is used. At this stage, the formalization of the general arguments was consciously put to the side in order to give enough space to expressions dictated by the verbal act, which is in line with the general rule described above.

Some of the argument structures:

ČLOVEK stisne, lahko s PRSTI ali s PRIPOMOČKOM, DVA ali VEČ DELOV skupaj, da se primeta/-jo

Collocations: stisniti [robove, testo]

> a HUMAN squeezes TWO or MORE PARTS together with his/her FINGERS or an APPLIANCE so that they are joined.
>
> Collocations: to squeeze [the edges, the dough]

OSEBA ali INŠTITUCIJA pobira DENAR od OSEB ali INŠTITUCIJ v zameno ZA UPORABO ČESA ali ZA STORITEV

Collocations: pobirati [najemnino, parkirnino]

> a PERSON or an INSTITUTION collects MONEY from PEOPLE or INSTITUTIONS in exchange FOR THE USE OF SOMETHING or FOR A SERVICE
>
> Collocations: to collect [a rent, parking fee]

KDO stisne PREDMET ali DEL TELESA z DLANJO ali s PRSTI tako, da ga zadrži v položaju

Collocations: stisniti [roko, prste, ročaj]

> SOMEBODY squeezes an OBJECT or a BODY PART with his/her HAND or FINGERS such that he/she keeps it in position
>
> Collocations: to squeeze [a hand, fingers, handle]

KOLIČINA, VREDNOST ali PREDMET, ki ima neko vrednost, pade na lestvici za določeno KOLIČINO ali VREDNOST
Collocations: [delnica, temperatura] pade; pasti za [pet stopinj]

> an AMOUNT, a VALUE, or an OBJECT which has some value decreases on a scale by a certain AMOUNT or VALUE
> Collocations: [share, temperature] decreases; to decrease for [five degrees]

V.4. Arguments in the argument structure

In the revision of the results stemming from the process of argument structure formation, 334 argument structures were checked in order to analyze the mechanisms of their creation, and

above all, the arguments used in argument structure formulations by different authors. Some conclusions are described below.

Since the group has so far consciously avoided predefined formalizations of argument structures and their individual arguments, some seemingly inconsistent reactions concerning the use of arguments have appeared, e.g. with regard to the use of the arguments a PERSON vs. a HUMAN, PEOPLE vs. a GROUP, or a THING vs. an OBJECT, to list just a few. However, in the first phase of the project such dichotomies are not seen as problematic, but rather are considered worthy of further exploration and consideration, such that final conclusions in this respect are made on the basis of real data. All currently existing arguments will be subject to possible inclusion in the list as it continues to be developed and refined. It has also been decided that in the preparatory phase of the project the same concept need not necessarily always be expressed in the same generic way, since it is thought that giving authors the possibility to brainstorm in different directions brings about a range of possible solutions from which the most optimal ones can be chosen.

The analysis of the argument structures used also showed that certain arguments are very generic, such as OSEBA (person), ČLOVEK (human), LJUDJE (people), SKUPINA (group), ŽIVAL (animal), DOGODEK (event), STVAR (thing), SNOV (substance), DEL TELESA (body part). On the other end of the scale, several arguments are more concrete, though still being representative of a group, such as OBLAČILO (garment), DREVO (tree), INŠTITUCIJA (institution), HRANA (food), PIJAČA (drink). This issue is related to the question of how much content should be allowed for arguments. In other words, should they be very general or can they also be very specific and concrete and thus sometimes overlap with collocations? The argument structure has its own position in the hierarchical structure of a lexical unit, which additionally consists of the indicator, grammatical patterns, and collocations, and the general rule is to not bend the argument structure too much towards either of these groups of information but to rather formulate it as some kind of a summary of them. Yet again, the description of a scenario has priority over a strict set of possible arguments, which is why, for example, it is acceptable to use the general expression DEL TELESA (body part) to describe an arm, a hand, or a neck in one argument structure, and to use VRAT (neck) in another one if a specific aspect is to be highlighted, despite the fact that VRAT (neck) appears on the level of collocations as well.

VI. SAMPLE ENTRY: *IGRATI*

To illustrate the semantic indicators and the argument structures in the SLD, the verb *igrati* ("to play") is presented after an analysis of 200 random concordance lines from the FidaPLUS corpus, with the data being sorted according to frequency:

No.	Semantic indicator (Slovene)	Semantic indicator (English)	Frequency
1	ukvarjati se s športom	to take part in a sporting activity	80
2	biti v igralski zasedbi	to be part of a cast	55
3	ustvarjati glasbo	to make music	26
4	početi kaj sproščujočega	to do something relaxing	20
5	delovati pod pretvezo	to pretend	6
Phrase: igrati vlogo		to play a role [phraseology]	8
	xxx (napake)	xxx (errors)	5

Interestingly enough, the following comparison of the semantic indicators of the Slovene verb *igrati* with the information from MEDAL and LDOCE for the verb "to play", shows a very similar picture:

MEDAL:	LDOCE:
1 take part in sport/game	1 CHILDREN
2 make music/sound	2 SPORTS/GAMES
3 have part in play, etc.	3 MUSIC
4 when children have fun	4 RADIO/CD ETC
5 when light moves	5 THEATRE/FILM
+ phrases	6 play a part/role
+ phrasal verbs	7 play ball
	8 PRETEND
	9 BEHAVE
	etc.

Below, the third meaning, i.e., "to make music", with 26 examples in the 200-concordance line sample, is presented as a whole (the English translations, however, are not part of the database); due to the various scenarios presented by the argument structures, it is internally structured. Further down the entry, lexical data should prove what is defined in the argument structure. The first semantic indicator and argument structure are used when they are so general that they can be valid for the dependent, specific pairs (called "submeanings" provisionally):

Semantic indicator 1: ***izvajati glasbo***
to play music

Argument structure 1: ***ČLOVEK igra GLASBO***
a PERSON plays MUSIC (3 out of 26)

Grammatical patterns: *kdo igra* somebody plays

Corpus examples:

*Če ne bi mogla **igrati**, bi se ukvarjala s čim drugim v glasbeni industriji.*
If I could not **play** (music), I would be involved in the music industry in some other way.

*Kdo **igra** in vabi k plesu, če povsod v umazanih jarkih teče deževnica?*
Who is **playing** music and inviting people to dance when rain water is flowing down all the dirty ditches?

*Zadnji album z naslovom Aaliyah je ravnokar izšel. **Igral** bom, dokler bom živ ...*
The last album, entitled Aaliyah, has jut been released. I will **play** as long as I live …

Semantic indicator 1.1: ***izvajati glasbo pred skupino ljudi*** to play music in front of a group of people

Argument structure 1.1: ***ČLOVEK igra na DOGODKU***

a PERSON plays at an EVENT (12 out of 26)

Grammatical patterns[4]:

kdo igra na čem	literally: somebody plays at something reference: somebody plays at some event
kdo igra v čem	literally: somebody plays in something reference: somebody plays in some place
kdo igra po čem	literally: somebody plays at/in something reference: somebody plays at some location
kdo igra komu	somebody plays to somebody

Corpus examples:

*Med drugo svetovno vojno se je bojeval v Evropi in **igral** vojakom.*
During the Second World War he fought in Europe and **played** music to soldiers.

*... v Piano Baru Gabbiano, kjer ob večerih **igrajo** živo glasbo.*
… in the Gabbiano Piano Bar, where they **play** live music some evenings.

*Dolga leta sem **igral** po hotelih, a sem pustil profesionalno glasbo in šel za navadnega delavca.*
For many years I **played** in hotels but later I gave up professional music and became an ordinary worker.

*Godba je leta 1901 **igrala** tudi na svečanem odprtju šole na Hardeku.*
In 1901 a brass band also **played** at the opening of the school in Hardek.

Semantic indicator 1.2: ***izvabljati zvoke iz glasbila***
to make music with an instrument

Argument structure 1.2: ***ČLOVEK igra GLASBILO***
a PERSON plays a MUSICAL INSTRUMENT
(7 out of 26)

Grammatical patterns:

kdo igra kaj	literally: somebody plays something
kdo igra na kaj	somebody plays on something

Corpus examples:

*Čas si je krajšal s petjem, naučil pa se je **igrati** tudi kontrabas. Zdaj ima v svojem repertoarju že okoli 700 pesmi.*

He liked to spend time singing and he also learnt to **play** the double bass. At the moment his repertoire already includes around 700 songs.

*Tudi inštrumente, na katere **igrajo** vaši priljubljeni glasbeniki, gotovo poznate.*
You certainly also know the instruments which your favourite musicians **play**.

Semantic indicator 1.3: ***izvajati zvrst glasbe***
to make music in a specific genre of music
Argument structure 1.3: ***ČLOVEK igra GLASBENO ZVRST***
a PERSON plays a KIND OF MUSIC
(2 out of 26)

Grammatical patterns:

kdo igra kaj someone plays something

Corpus examples:

*V Sloveniji sem pop igrala le dvakrat, sicer pa ves čas **igram** klasiko.*
I have played pop music only twice in Slovenia, otherwise I always **play** classical music.

Semantic indicator 1.4: ***izvajati glasbo skupaj z drugimi***
to make music together with other people
Argument structure 1.4: ***ČLOVEK igra v SKUPINI***
a PERSON plays in a GROUP (2 out of 26)

Grammatical patterns:

kdo igra v čem
literally: somebody plays in something
reference: somebody plays/is in a musical group

kdo igra pri čem
literally: someone plays at something
reference: somebody plays/is in a musical group

Corpus examples:

*V Carini **igrata** še Zdravko Štamcar in Borut Mehle.*
Zdravko Štamcar and Borut Mehle are **playing** in Carina as well.

*To vzdušje je soustvaril predhodni nastop harmonikarja Marka Hatlaka, ki bo avgusta začel **igrati** pri izjemni zasedbi Terra Folk.*
Such an atmosphere was co-created by the preceding performance of the accordion player Marko Hatlak, who will start **playing** in the excellent Terra Folk in August.

VII. CONCLUSIONS

As stated in the introduction, the main topic of this paper is the creation of the Slovene Lexical Database, which is a constituent part of the Communication in Slovene Project, the main focus of which is the treatment of semantic information therein. As mentioned above, most of the conclusions presented refer to the treatment of verbs, with regard to which the SLD structure includes two types of semantic information for each lexical unit –a semantic indicator and an argument structure. The two have been designed to complement each other by containing two different types of semantic information: the indicator provides guidance regarding the particular meaning of the lexical unit in question, while the argument structure describes the scenario and arguments necessary for a certain activity to take place.

In order to extract and create semantic information in the compilation of the SLD entries, a combination of manual and statistical approaches are used, namely an analysis of a random sample of concordance lines and an analysis of the information obtained from SketchEngine in the form of a WordSketch for each individual entry. Such an approach has proved to be the most successful in attempting to obtain an appropriate balance between the options available to the corpus lexicography of the 21st century, and results in relatively complete information on a certain entry.

In order to support and explain the above-mentioned decisions and guidelines, the paper contains a presentation of the work processes of the project, since its inception in June 2008, as well as further plans and objectives, together with the relevant theoretical and practical background. Furthermore, decisions and guidelines are illustrated by examples of actual solutions from some entries that have already been compiled, and also by a more complete treatment of one meaning of a sample entry –the Slovene verb *igrati* ("to play").

This having been said, it must also be pointed out that the compilation of the SLD is still in its initial phase, meaning that the team working on it is still preparing the style guide, and thus far has done a relatively complete analysis of only some verbs. Thus, we expect to reach some further conclusions and decisions as we move along and research other parts of speech, and that some of the conclusions and decisions already arrived at will be subject to change on the basis of future results and discoveries.

Some of the open issues regarding the extraction of corpus data and definition of semantic information are the following: the unification of certain arguments that appear in the argument structure where there are several possibilities, e.g. PERSON vs. HUMAN; a possible compilation of a list of such arguments; the question of how abstract vs. concrete the argument structure should be; how much and when it is acceptable for the argument structure to overlap with collocations and the like.

References

Atkins, B. T. S. and M. Rundell (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Braasch, A. (2002). Current developments of STO the Danish Lexicon Project for NLP and HLT Applications. *Third International Conference on Language Resources and Evaluation*, Proceedings. Las Palmas de Gran Canaria. Vol. III. 986-993.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Fillmore, C., C. R. Johnson, and R. L. Petruck, (2003). Background to Framenet. *International Journal of Lexicography*. Oxford: Oxford University Press 16/3. 235-250.

García-Miguel, J. M., L. Costas and S. Martínez (2005). "Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE". In G. Wotjak & J. C. Otal (eds.), *Entre semántica léxica, teoría del léxico y sintaxis*. Frankfurt am Main: Peter Lang, pp. 373-384.

Hanks, P. (2008). "Do Word Meanings Exist?". In T. Fontenelle (ed.), *Practical Lexicography: A Reader*. Oxford: Oxford University Press, 125-134.

Hanks, P. (2004). Corpus Pattern Analysis. In Williams, G. and S. Vessier (eds.), *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress* (Euralex 2004). Volume I. Lorient: Université de Bretagne Sud, pp. 87-97.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities* 31:2, 91-113.

Longman Dictionary of Contemporary English, http://www.ldoceonline.com/

Rundell, M. (2007). *Macmillan English Dictionary for Advanced Users*. 2nd edition, Oxford: Macmillan.

Vázquez, G., L. Alonso, J. A. Capilla, I. Castellón, A. Fernández (2006). SenSem: sentidos verbales, semántica oracional y anotación de corpus. In *Procesamiento del Lenguaje Natural*, 37, p. 113-120.

[1] At this stage, the team consists of the following members: Simon Krek, Polona Gantar, Mojca Šorli, Olga Pobirk, Simon Šuster, Katja Grabnar, and both authors. The findings presented in the article should be regarded as collective work.

[2] Owing to the fact that at the time when this presentation was being prepared the system was being intensively developed, some of the arguments presented in the paper may still have the status of a work in progress.

[3] The material in the SLD is monolingual, while the examples in the paper are translated into English in order to aid understanding.

[4] With regard to some of the following grammatical patterns, it must be pointed out that since this level of language analysis is language specific it was impossible to adequately translate all patterns and at the same time preserve the information about valency features in the Slovene language. Thus, we first listed literal translations of Slovene patterns and, when necessary, we defined their reference in suitable terms in English.