A Procedure for the Semiautomatic Building of Consistent Dictionary Definitions

Karel Pala and Eva Žáčková

NLP Laboratory, Faculty of Informatics, Masaryk University
Botanická 68, CZ-602 00 Brno, Czech Republic
E-mail: {pala,glum}@fi.muni.cz

1 Introduction

In this contribution we are going to explore both syntactic and semantic structures of the dictionary definitions as they can be found in standard printed or machine dictionaries (as e.g. SSJČ [11] or NODE [7]). Particularly, we would like to pay attention to the usual ways of building dictionary definitions, and to examine whether and to what extent they reliably and systematically follow the standard principles of the formal description of language data. The results following from the presented analysis should be applied to building dictionary definitions (glosses) for Czech WordNet (being further developed in the frame of Balkanet Project [1]) and also the new Czech Lexical Database.

When working with various standard dictionaries (such as SSJČ for Czech, or NODE for English) one can see that the lexicographers follow several general but rather pragmatic principles in building the dictionary definitions. In other words, the techniques applied in building the dictionary definitions are based on the selected general principles but we can hardly say that they form a consistent and complete theory. In this respect, it can be observed that there are considerable differences between semantic theories (see e.g. [4, 6] or Cruse, 2000) and the lexicographer's practice. However, the dictionaries are still almost the only resources of the lexical data for NLP, thus we have to pay attention to them at the first place. Though lexicographers follow the mentioned general principles and use well established techniques it is no secret that many objections can be raised with regard to the consistency of the dictionary definitions, both from the syntactic and semantic point of view. The considerable number of dictionary definitions are quite often just the examples (though recently selected carefully from corpora, see for example Cobuild [3]).

In building dictionary definitions the following techniques have been certainly regarded as the standard ones:

- definitions using genus proximum (GP) and the distinguishers (typical for nouns),
- definitions using semantic components or features (primitives) (with verbs: *kill* = cause to die),
- definitions based on the relation of troponymy (verbs: e.g. *talk* — *whisper*),

– definitions using synonymical explanations or just one word synonyms (typical for adjectives, *clever – bright*),

– definitions based on collocational determination of the sense of the entry (typical for adjectives: *good student, good mile*).

– definitions exploiting various kinds of ad hoc descriptions or explanations (any POS),

– definitions based on the descriptions of events or situations (see e.g. definition of table:2 *if you ask for a* **table** *in a restaurant, you want to have a meal there* (Cobuild [3], p.1697).

It is correct to say that the mentioned lexicographic techniques are not strictly related to any of the standard semantic theories, such as presented by Leech [4] or Lyons [6] and recently also by others (Cruse, 2000). For example, the spontaneous lexicographer's use of semantic components can be certainly linked rather freely to the theory of the semantic features (primitives) as we can find it in [4] (p. 89-122) but no standard dictionary exploits them systematically. Similarly, the meaning relations like synonymy, antonymy, homonymy and hyponymy as they are introduced, say, in [6] (p. 270-301), are used in standard dictionaries but rather spontaneously and without the necessary consistency. The obvious exception is, however, WordNet 1.5., or thesauri like Roget but they can hardly be labelled as a "standard" dictionaries; unfortunately in its "dictionary part" containing glosses Wordnet 1.5 suffers from many inconsistencies as any other "standard" dictionary (this is shown in [9]).

Moreover, the lexicographer's above mentioned and widely used techniques based on GP and distinguishers do not represent, as far as we know, an integrated part of any well established semantic theory though they are used frequently in many dictionaries. The only lexical resource in which the reasonably well defined hypero/hyponymy (H/H) relations can be found and are systematically exploited seems to be WordNet and lexical databases of this sort (according to our knowledge). If we look for other theoretical works that could serve as the theoretical basis for building formally more consistent dictionaries perhaps *The Generative Lexicon* by Pustejovsky [10] seems to offer a reasonably firm ground on which this could be attempted. Project named Corelex [2] seems to point in this direction as well but it can be hardly considered a standard dictionary either.

It can be pointed out that the need for a reasonable "dictionary theory" is obvious and the progress in the area of the lexical resources strongly depends on building such theory that will allow us to prepare the consistent and formally well constructed lexical databases or machine dictionaries from which also the classical dictionaries suitable for human user can be later and reasonably derived.

2 The Correspondencies between Syntactic and Semantic Structures of Dictionary Definitions

Here we will try to explore the main types of correspondencies between syntactic and semantic structures of dictionary definitions with respect to nouns as they can be found in Czech dictionary SSJČ.

2.1 The Syntactic Structures of Dictionary Definitions (in Czech)

It is our strong belief that if we want to make the building of the (semantically) consistent dictionary definitions more formal we have to know more about their formal structure. In other words, we should examine their syntactic structures and see what types can be found in this respect and how they can be related to the semantic organization of dictionary definitions indicated above. For this purpose we have chosen the representative Dictionary of Written Czech (SSJČ [11]) that now exists also in the machine readable form [8].

To examine larger data from SSJČ we need a tool that would enable us to analyse the syntactic structures of dictionary definitions from SSJČ. For this purpose we have chosen a partial parser for Czech [13] which after a modification of the used rules can be used to parse dictionary definitions for nouns (and for verbs, adjectives and adverbs as well but we will deal only with nouns). For the present purpose we have prepared a sample containing 10 000 dictionary definitions (with the following distribution: 40 % nouns, 30 % verbs, 15 % adjectives, 15 % adverbs). The goal is to find the main types of syntactic patterns appearing in the noun dictionary definitions as they occur in SSJČ and to make the possible generalizations.

Preliminary examination shows that the dictionary definitions display some regular syntactic structures which quite systematically correspond to the semantic organization of the dictionary definitions and indicate their individual parts.

2.2 Noun Definions in SSJČ

We are interested here in the semantic organization of the noun definitions, or in other words, in their "deep structure" and how it is related to their "surface" syntactic structures. It can be seen that they mostly "behave" in a standard way, typically, these definitions follow the classical dictionary definition pattern, i.e. first part of the definition consists of the *genus proximum (GP)* and the second one represents the distinguishers (d_1, $\mathrm{d}_2, \ldots$, d_n).

We can observe here a good parallelism between GP and the first noun group in the dictionary definitions What also can be done is to check semiautomatically the heads of these noun groups against the corresponding nouns in Czech WordNet and to see how regularly they contain the hyperonymical expressions (such as furniture in our example group of selected furniture nouns) – this can be done by comparing them with the corresponding H/H trees in WordNet (see below).

At the first glance we can see that the distinguishers are expressed in several ways: as noun groups, relative sentences, adjectival phrases with complements or as prepositional groups. The closer examination, however, shows that the picture is more complicated and the corresponding surface syntactic structures are much richer (see below). Thus the parsing the dictionary definitions should discover the inventory of the syntactic structures that may correspond to the GP + d_1, $d_2, \ldots, d_n$ scheme. We give some typical examples from SSJČ together with their English equivalents from NODE. Angle brackets in Czech descriptions mark out the particular groups (and their cases in which they may occur).

stůl: *<kus>*$_{ng1}$ *<nábytku>*$_{ng2}$ *<tvořený>*$_{ap}$ *<(vodorovnou) deskou>*$_{ng7}$ *<na nohách>*$_{png6}$ *<nebo>*$_{conj}$ *<na podstavci>*$_{png6}$
table: a piece of furniture *with a flat top and one or more legs, providing a level surface on which objects may be placed, and which can be used for such purpose as eating, writing, working or playing games*

židle: *<přenosný kus>*$_{ng1}$ *<nábytku>*$_{ng2}$ *<(s opěradlem)>*$_{ng7}$ *<k sezení>*$_{png3}$ *<pro jednu osobu>*$_{png4}$
chair: a separate seat *for one person, typically with a back and four legs*

křeslo: *<pohodlné sedadlo>*$_{ng1}$ *<s opěradly>*$_{ng7}$
armchair: a large, comfortable chair *with side supports for a person's arm*

skříň: *<vyšší kus>*$_{ng1}$ *<nábytku>*$_{ng2}$ *<na ukládání různých předmětů>*$_{png4}$ *<nebo>*$_{conj}$ *<na věšení šatstva>*$_{png4}$
cupboard: a piece of furniture *with a door and usually shelves, used for storage, wardrobe*

blbec, blb: *<velký hlupák>*$_{ng1}$, *<pitomec>*$_{ng1}$, *<idiot>*$_{ng1}$
idiot: *a stupid* person

letadlo: *<zařízení>*$_{ng1}$ *<schopné>*$_{ap}$ *<létání>*$_{ng2}$ *<k dopravě>*$_{png3}$ *<vzduchem>*$_{ng7}$
aeroplane: a powered flying vehicle *with fixed wings and weight greater than that of the air it displaces*

abeceda: *<soubor>*$_{ng1}$ *<všech písmen>*$_{ng2}$ *<zpravidla>*$_{adg}$ *<latinského písma>*$_{ng2}$, *<užívaných>*$_{ap}$ *<v některém jazyce>*$_{png6}$, *<v ustáleném pořádku>*$_{png6}$
alphabet: a set of letters or symbols *in a fixed order used to represent the basic set of speech sounds of a language, especially the set of letters from A to Z*

student: *<posluchač>*$_{ng1}$ *<vysoké školy>*$_{ng2}$ *<nebo>*$_{conj}$ *<žák>*$_{ng1}$ *<střední školy>*$_{ng2}$
student: a person *who is studying at a university or other place of higher education*

3 Parsing Syntactic Structures of Dictionary Definitions (in Czech – SSJČ)

The frequencies of the different definiton types in SSJČ are shown in Table 1. When we tried to extend the recall of the parser rules the time necessary for the analysis has increased enormously since the mechanism of the partial analysis is obviously not suitable for the exhaustive parsing. (Processing of approx. 3700 entries took about 3 days.)

Table 1. Frequencies of the Different Deffinition Types.

	1^{st} file	2^{nd} file	syntactic structure
# entries	3672	5935	
not processed (%)	21.0	20.3	
processed (%)	79.0	79.7	
from processed:			
def1 (%)	13.0	11.6	`E = one-word-synonym`
def2 (%)	62.7	63.2	`E = ( Ng | Pg )+`
def3 (%)	18.0	18.5	`E = ( Ng | Pg )+ Ap ( Ng | Pg )+`
def4 (%)	2.3	2.0	`E = Ng Sr`
def5 (%)	0.1	0.1	`E = ( Ng | Pg )+ Ap ( Ng | Pg )+ Sr`
def6 (%)	3.4	4.2	`E = [kdo|co|někdo|něco] .*`
def7 (%)	0.4	0.4	`E = [schopnost|neschopnost] .*`

`E` in the examples of syntactic structures stands for the defined word (it is simple nominal group) in the definition. `Ng` can represent both simple nominal groups but also quite complicated recursive nominal groups (including the groups with coordination and genitive groups). `Pg` represents prepositional noun and pronoun groups (usually are used separate nonterminals `Png` and `Ppg`). Usual conventions of context-free syntax are used: (`+` sign represents one or more repetition, `*` represents zero or more repetition, `|` stays for alternative etc.).

It can be observed that the type of definition labelled def5 is not very frequent, but the question is whether the definitions of this type can be classified under def4/def3, ev. to which category they belong. def6 a def7 represent very specific types of the definitions established only on the base of the specific key word like *kdo (who)* that do not belong into any of the previous category.

It can be said the definitions of the entries for whose no structure has been found usually can be intuitively classified as belonging to some of the groups 1–5. However, they display very complicated structures (e.g. very complicated attributive noun groups), that prevent the parser (the particular rules in it) from recognizing them. There are only few entries that do not belong into any of the introduced groups/categories, for example *názor, že... (the opinion that...)*.

The examples of the analyses of the individual definitions is given in Table 2. Angle brackets mark out the particular groups forming definition body. Labels used here are more specific then the ones used in the formal desriptions of the

Table 2. Examples of the Different Deffinition Types.

type	example
def1	***drożdí*** = <*kvasnice*>$_{ng1}$
def2	***afix*** = <*společné označení*>$_{ng1}$ <*pro předpony a přípony*>$_{png4}$
def3	***balneografie*** = <*obor*>$_{ng1}$ <*lékařství*>$_{ng2}$ <*zabývající se*>$_{ap}$ <*vědeckým popisem*>$_{ng3}$ <*lázní*>$_{ng2}$ <*a*>$_{conj}$ <*léčivých vod*>$_{ng2}$
def4	***akumulátor*** = <*zásobník*>$_{ng1}$ <*elektrické energie*>$_{ng2}$ <*, která v něm byla dříve nahromaděna chemicky*>$_{sr}$
def5	***konstantan*** = <*slitina*>$_{ng1}$ <*mědi*>$_{ng2}$ <*a*>$_{conj}$ <*niklu*>$_{ng2}$ <*vyznačující se*>$_{ap}$ <*velikým elektrickým odporem*>$_{ng3}$ <*, který se málo mění teplotou*>$_{sr}$
def6	***falšovatel*** = <*kdo*>$_{keyw}$ *něco falšuje*
def7	***výrazovost*** = <*schopnost*>$_{keyw}$ *něco vyjádřit, vyjadřovat*

particular definitions (but it is easy to transform them to the respective general form). English equivalents of defined words are *yeast, affix, balneology, accumulator, constantan (*a kind of copper-nickel based alloy*), forger, expressiveness.*

If we take the parsed syntactic structures of the processed dictionary definitions and extract their head noun groups representing (according to the parser) the GP pattern we obtain a list of expressions that are hyperonyms of the headwords in the dictionary definitions. The first part of this list is given below and it contains 30 most frequent (Czech) hyperonyms (sorted according to their frequency) from our sample of dictionary definitions. To confirm that they are hyperonyms we compared them with the corresponding expressions from Czech and English WordNet (the first number indicates the frequency in the Czech sample, then there is Czech literal with its sense number since the literals can be associated with more senses, and its English equivalent with its sense number as well). The results of the comparison show that all the expressions extracted from the dictionary definitions are hyperonymical, thus in this way confirming our starting assumption that GP patterns can be processed and obtained from the dictionary definitions automatically. Then it is also possible to check for their consistency. The next goal is to try to recognize the distinguishers at least in a semi-automatic way though we are aware that this task is not going to be as easy as the former one.

173: kdo (who)
91: přístroj:1 (apparatus:1)
72: druh:1 (sort:2, kind:1)
63: zařízení:1 (installation:2)
56: část:1 (part:3)
52: člověk:1 (human:1, person:1)
42: souhrn:1 (aggregate:1, sum:1)
42: místo:1 (place:10)
37: nástroj:2 (instrument:2)
36: obor:2 (discipline:5)
35: nauka:1 (doctrine:1)

34: látka:2 (matter:1, substance:1)
28: příslušník:1 (member:4)
27: skupina:1 (group:1)
26: způsob:2 (means:1, way:1)
22: jednotka:3 (unit:8)
21: něco (something)
21: činnost:1 (activity:1)
20: stav:1 (state:1)
20: součást:2 (component:1)
19: vlastnost:1 (quality:1)
18: místnost:1 (room:1)
18: hornina:1 (rock:1, stone:1)
17: stroj:1 (machine:2)
16: útvar:2 (formation:5)
16: sloučenina:1 (compound:4)
16: schopnost:3 (ability:1)
16: pracovník:1 (worker:2)
14: oddělení:2 (department:1)
14: nedostatek:1 (deficiency:1)
14: názor:1 (opinion:1)

4 Conclusions

We can conclude that this sort of analysis appears to be useful and helps in testing the noun dictionary definitions for their syntactic and semantic consistency. Its results can be, in our opinion, applied later to semi-automatic preparation of the dictionary definitions thus saving a considerable amount of manual work. Therefore the following steps will be taken in building dictionary definitions (glosses) for the developed Czech WordNet and prepared Czech Lexical Database as well:

- to use the different types of definitions for the different parts of speech in a systematic way, i.e. to use scheme GP + $d_1, d_2, \dots, d_n$ mostly for nouns, the definitions based on semantic components and troponymy relations for verbs and synonymical explanations combined with collocational examples for adjectives and adverbs,
- to use the semantic classification of Czech verbs and integrate it appropriately into the dictionary definitions (glosses),
- to examine in a more detailed way the definitions with GP + $d_1, d_2, \dots, d_n$ pattern for nouns and to check whether the distinguishers can be inherited systematically between H/H trees, eventually to explore further how the properties represented by the distinguishers can be related and systematically organized,

– to examine whether the distinguishers can also capture the relation of meronymy/holonymy and in the positive case to find out how frequent it is,
– to explore more systematically the semantic structure the collocational examples using corpus data and integrate them systematically into the adjective dictionary definitions (glosses),
– the ultimate goal of the mentioned steps is to obtain the systematic, formal and consistent dictionary definitions (glosses) for the entries (possibly taking form of synsets) in the Lexical Database and Czech WordNet.

References

1. Balkanet Project, 2001.
2. Buitelaar, P.: Corelex. Systematic Polysemy and Underspecification. Ph.D. Thesis, Computer Science, Brandeis University, 1998.
3. Collins Cobuild English Dictionary, ed. by J. Sinclair. Cobuild. London, Harper Collins Publishers, 1995.
4. Leech, G., N.: Semantics. Harmondsworth, Penguin, 1974.
5. Levin, B.: English Verb Classes and Alternations. The University of Chicago, Chicago Press, Chicago, 1993.
6. Lyons, J.: Semantics. Cambridge University Press, Cambridge, 1977.
7. New Oxford Dictionary of English. edited by P. Hanks. Oxford University Press, Oxford, 1998.
8. Pala, K. and Smrž, P.: Slovko 2001.
9. Pala, K. and Smrž, P.: Glosses in WordNet 1.5 and their Standardization, accepted at LREC, 2002.
10. Pustejovsky, J.: The Generative Lexicon, A Bradford Book, MIT Press, Cambridge, 1995.
11. Havránek B. et al.: Slovník spisovného jazyka českého (SSJČ, Dictionary of Written Czech), Academia, Praha, 1960.
12. Vossen, P.: EuroWordNet 1, 2, Final Report, University of Amsterdam, CD ROM, 1999.
13. Žáčková, E.: Partial Parsing (of Czech), Ph.D. Thesis, Masaryk University, Brno, 2002.