ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 79–88

Building Czech Wordnet

Karel PALA, Pavel SMRŽ
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {pala,smrz}@fi.muni.cz

Abstract. This paper describes the process of building Czech wordnet. We give the enumeration of the resources and tools used for this purpose and characterize so far obtained results. There are some problems with Czech as a synthetic language, with its rich inflectional morphology and word derivation. They are mentioned below and some solutions are suggested. The necessary resources for building Czech wordnet were worked out almost from scratch and required considerable adaptation.

1. Introduction

Czech has a special position in the BalkaNet project – the core of Czech wordnet was developed in EuroWordNet 2, thus Czech wordnet is now mainly being completed and considerably improved. In BalkaNet our experience has been exploited as a starting point. Having the necessary know-how in the field we succeeded in developing and implementing a tool for editing and browsing wordnets. VisDic become a standard for working with wordnets and that is also used outside of the project. Development of VisDic is closely related to the new way of representing wordnets, particularly to the XML representation of the wordnet data which is now used both for representing languages included in EuroWordNet 1, 2 and BalkaNet as well.

2. Selection of the Core of Czech Wordnet

The first set Czech synsets containing approximately 1 000 items has been obtained from Slovník spisovné češtiny (Dictionary of Literary Czech, DLC, Academia, Prague, 1960), which by this time was the only usable electronic source for Czech. Because of its inconsistent structure of entries and incomplete description of senses it could have been exploited just in part to obtain the core of Czech wordnet. We have sorted out all the entries from DLC that contained genus proximum definitions (hyperonyms).

This was done with a simple parser that had retrieved the definitions from DLC and yielded the list of the items (about 3 000 candidates) from which the set of the first Czech synsets has been finally compiled (using frequency criteria).

In the next step, however, DLC was not used because of its insufficiencies and we had to solve the problem what resource to use. We decided to take advantage of Czech-English and English-Czech electronic dictionaries. We linked the created set of Czech literals with the corresponding English equivalents. The result has been then semi-automatically compared with Princeton WordNet. The Czech synset candidates have been manually checked and the final version of the Czech wordnet core has been compiled.

This cycle has been repeated about three times. Most of the effort was spent on sense discrimination since this is the weakest point of DLC: in fact, it had to be done almost completely anew: the existing resources gave us not more than 30% of necessary information.

Though it was more time consuming, we concentrated our effort on the entries with more senses:

1. for noun synsets we have processed about 2 550 polysemous (4–5 senses in average) dictionary entries and 7 050 with one sense (these could be added semi-automatically with manual checking),

2. for verb synsets about 3 000 polysemous (more than 5 senses in average) dictionary entries have been retrieved and approximately 2 500 with one sense.

This gave us about 15 000 synsets but after manual checking the resulting number of synsets dropped to approximately 13 000 items. In this way the core of Czech basic word stock was compiled into Czech wordnet and it represented a sound base for further completing and enlargement.

3. Tools and Resources Employed

The following tools have been used in the course of building the core of the Czech wordnet:

- special sorting program developed for various types of dictionary processing (on the Linux platform);

- simple parsing program able to analyze dictionary entries in Dictionary of Literary Czech and Dictionary of Written Czech and to select entries containing genus proximum definitions (with hyperonyms), implemented in C under Linux;

- simple translating program able to process a bilingual dictionary (Czech-English-Czech): it uses a simple pattern matching and associates Czech entries with their English equivalents and then tries to link them to Princeton WordNet, this program helped us to enlarge the number of the synsets considerably in the beginning (its error rate was about 25%), the program was written in C under Linux;

- special program for creating ILRs. It worked in 3 cycles, used Princeton WordNet and the respective Czech data as input, implemented in C under Linux;

- program able to compute MI (mutual information) score and other word-association measures for word forms as well as lemmata from Czech text corpora (particularly from the corpus ESO built and mantained at Faculty of Informatics and also from the Czech National Corpus), written by Pavel Rychly [1], (in C and Python, developed under the Linux platform). It was used to obtain the list of present-day Czech collocations, approx. 2 000 were included;

- morphological module AJKA by Radek Sedlacek and Pavel Smrž [2], containing 350 000 Czech stems and able to perform full morphological analysis of an arbitrary Czech text (also used for tagging corpora texts in the mentioned corpora). The program has been written in C and runs under Unix as well as MS Windows platforms;

- VisDic – Wordnet editor and browser implemented at the Faculty of Informatics, Masaryk University, described in this issue [3]. The tool replaced Polaris that has been exploited during the EuroWordNet 1, 2 Project.

The following resources have been used in the course of building Czech wordnet:

- Dictionary of Written Czech (DWC), by J. Filipec et al, Academia, Prague, 1986, size: approx. 50 000 entries, in electronic form, with automatic lemmatization, word form recognition and basic word derivation;

- Bilingual Lingea Lexicon: Czech-English-Czech, by Lingea Ltd., Brno, size: approx. 125 000 entries, in electronic form (on CD ROM), with automatic lemmatization, word form recognition and basic word derivation. It was mainly used as a source for finding English equivalents matching Princeton WordNet;

- Dictionary of Czech Synonyms, by Karel Pala and Jan Vsiansky, NLN, Prague, 1994–5, size: approx. 22 000 entries, electronic version with automatic lemmatization, word form recognition and basic word derivation;

- Czech Synonymical Dictionary and Thesaurus I, II, III (Český slovník věcný a synonymický I, II, III), ed. by J. Haller, SPN Prague 1969–1977. This is the only Czech thesaurus-like dictionary, it has not been finished because almost all its authors have died one after another. It does not exist in an electronic form. The dictionary contains useful collection of the data that had been consulted when the hyperonym trees were created.

- Fully tagged and disambiguated corpus DESAM (both structurally and grammatically) that has been compiled at the Faculty of Informatics [4]. It contains mostly newspaper and magazine texts from 1992–1996, with size about 1 mil. Czech words;

- Corpus ESO built at the Faculty of Informatics in the course 1998 from newspaper and magazine texts (1996–1998), size about 61 mil. Czech word forms, partially tagged (lemmatized);
- Czech National Corpus, compiled at Institute of Czech National Corpus, Prague, with the size currently more than 250 mil. tokens;
- A list of Czech collocations containing approximately 100 000 lines was obtained from the above-mentioned text corpora, sorted according several criteria (MI score, relative frequencies, syntactic criteria) and prepared for the processing of the present-day Czech collocations. About 2 000 collocations obtained in this way were integrated into Czech wordnet. The same procedure was then applied also to Czech National Corpus;

4. ILRs in Czech Wordnet

The hyper-/hyponym and most other relations have been built mostly automatically taking advantage of Princeton WordNet. However, we had to pay a special attention to the problem of the translation equivalents. The obtained experience can be generalized in the following way: the highly inflectional nature of Czech with its rich formal and derivational morphology causes that in some standard cases the straightforward English translation equivalents either cannot be easily found or have to be substituted by the various syntactic constructions often depending on context.

For example, we do not have any natural Czech counterpart for instrumentality though an artificial equivalent "nástrojovost" can be formed but our main doubt is whether this concept is not a bit artificial even in English. Generally, in our view, many hyperonym trees in Princeton WordNet typically suffer from a kind of redundancy and we are convinced that for some future applications it will be necessary to flatten them to a certain degree. It is our intuition that the steep trees frequent in wordnets may not be in a good agreement with human semantic memory but this intuition has to be explored and verified experimentally.

We have paid attention to ILRs in Czech wordnet and especially we have explored the possibilities to exploit selected word formation relations which are both rich and regular in Czech. It appears that for ILR shown below some typical groups of word formation suffixes can be found and applied to establish at least the following ILRs:

1. ROLE_AGENT – INVOLVED_AGENT
   Many noun synsets can be labeled as ROLE_AGENT on the ground of the regular derivations between verbs (verb stems or roots) and corresponding nouns. The following suffixes can serve for this purpose: -tel as in *učit – učitel (teach – teacher)*, *mučit – mučitel (torture – torturer)*, *kázat – kazatel (preach – preacher)*.
   Other suffixes of this kind are:
   -ač, kopat – kopáč (dig – digger),
   -eč, svářet – svářeč (weld – welder),
   -ič, řídit – řidič (drive – driver),

-ec, střílet – střelec (shoot – shooter),
-ák, zpívat – zpěvák (sing – singer),
-ač, hlídat – hlídač (watch – watcher),
-ík, jíst – jedlík (eat – eater),
-ér, trénovat – trenér (train – trainer),
-átor, programovat – programátor (program – programmer),
-ář, lhát – lhář (lie – liar),
-ant, okupovat – okupant (occupy – occupier),
-ící, pracovat – pracující (work – worker),
-čí, nakupovat – nákupčí (buy – buyer),
-ej, čarovat – čaroděj (conjure – sorcerer),
-al, mazat – mazal (paint – dauber),
-il, žvanit – žvanil (talk – chatterbox)
and others.

2. ROLE_PATIENT – INVOLVED_PATIENT
   -anec, trestat – trestanec (punish – convict)
   -ek, utrhnout – útržek (tear off – scrap)

3. ROLE_LOCATION – INVOLVED_LOCATION
   -iště, lovit – loviště (hunt – hunting ground)
   kotvit – kotviště (anchor – berth)
   bojovat – bojiště (fight – battle ground)

   The suffix –iště is used generally to derive nouns denoting places or locations, it is a typical suffix for this purpose.

4. ROLE_INSTRUMENT – INVOLVED_INSTRUMENT
   -ák, bodat – bodák (stab – bayonet)
   -átko, ukazovat – ukazovátko (point – pointer)
   -ítko, tlačítko – stlačit (push – button)

   Here also suffixes -átko or -ítko normally indicate that nouns with them denote instruments.

5. HAS_SUBEVENT – IS_SUBEVENT_OF
   přibližovat se – jít (approach – advance)
   přijít – jít (arrive – go)
   vejít – jít (enter – go)
   odejít – jít (leave – go)
   vyslovovat – mluvit (articulate – speak)
   vymyslet – myslet (invent – think)
   odnést – nést (carry away – carry)

Closer examination of the verb pairs shows that aspect pairs in Czech can be captured in this way since it is obvious that the relation between ongoing activity (imperfective aspect) and finished one (perfective) can be quite naturally interpreted as the relation HAS_SUBEVENT – IS_SUBEVENT_OF. This seems to offer an interesting possibility how to solve aspect problems within Czech wordnet but more aspect pairs will have to be examined and checked to see if the final solution can be obtained in this way.

Nevertheless, some problems with aspect pairs in Czech remain, particularly the fact that the members forming an aspect pair as a rule do not display always the same senses. The solution should be arrived at when a larger number of the aspect pairs (about 5 000 pairs) will be collected with all their senses – this work is going on now. We have to say, however, that no Czech dictionary, including the representative DLC, does not list the aspect pairs in Czech, not speaking about the description of the relations between the members of the aspect pairs and their respective senses.

5. Translation Equivalents and Nonlexicalized Concepts

The BalkaNet partners decided to systematically record concepts from other languages (mainly from English based on Princeton WordNet) that are not lexicalized in their particular languages. Several examples of such problems are presented in this section.

When working both with the noun and verb synsets in Czech we have faced the problem of the translation equivalents and corresponding gaps with regard to English. There are two kinds of cases where it is not possible to find the equal synonyms (or even near synonyms):

1. The Czech synsets do not have corresponding counterparts in Princeton WordNet due to differences in lexicalizations and conceptualizations between Czech and English:
   (a) Czech synsets do not have equivalents in English at all;
   (b) Czech synsets do not have equivalents in Princeton WordNet but we have been able to find their English equivalents in general.
2. The second collection of Czech items without equivalents in English comprises Czech items that are typologically different due to the highly inflectional nature of Czech with its rich formal and derivational morphology. Because of that in some typical cases the straightforward English translation equivalents either cannot be easily found or have to be substituted by the various syntactic constructions or context dependent equivalents have to be searched for. At the present moment, at least four types of basically morphological phenomena causing the gaps have to be mentioned:
   (a) verb aspect;
   (b) reflexive verbs;
   (c) verb prefixation (single, double);

(d) diminutives (noun derivation by suffixation);

(e) move in gender (noun derivation by suffixation).

The described phenomena, in our view, are relevant in the wordnet context and they can be generalized in the following way:

1. aspect opposition in Czech is a ternary relation: imperfectives – unbounded in time – perfectives – bounded in time – iteratives – iterative, bounded in time,

   The question to be answered is: shall we have one synset for each of the mentioned verb types or should we keep the information about them in one synset for all of them? In our opinion, the appropriate solution, at least within the Czech wordnet, would be to introduce appropriate internal language relations that could link together the respective synsets. In the case of the aspect we would suggest to introduce internally in Czech wordnet as a new kind of ILR something like X_HAS_IMPF, X_HAS_PERF, X_HAS_ITER attributes.

2. reflexives

   At least the three relevant types of the reflexive verbs have to be taken into consideration:

   - reflexiva tantum;
   - verbs expressing reflexivity;
   - verbs expressing reciprocity;

   The suggestion is to have them systematically as separate synsets (as it is standard in Czech dictionaries). Other cases of reflexive verbs and their meanings belong to the syntactic level.

3. prefixed verbs

   They combine the aspect distinctions with iterativity and various types of the distributed actions, thus they are sources of many gaps. Some of them can be translated by (English) phrasal verbs but there is not very much regularity in this respect. The use of the relation HAS_NEAR_SYNONYM appears as a possible solution, though within Czech wordnet we have to mark them completely.

4. gender pairs

   They display binary semantic opposition – male : female, and the question is again similar as above: shall we have this distinction in one synset as it is in English WordNet or it is reasonable to keep them apart as a separate synsets and have a special attribute with two values for them, say: X_HAS_MALE – X_HAS_FEMALE?

5. diminutives

   They display a sort of ternary semantic opposition as in the case of aspect, however, there is a relevant difference: one of the attributes expresses an emotional attitude of the speaker in an lexicalized way. Thus, as we have indicated above, the following cases can be found with Czech diminutives:

- standard as in "dům – domek" (in English house, cottage);
- small thing as in "domek – domeček" (small house, Wendy house);
- emotional attitude (something like my dear little house).

To preserve this information in Czech wordnet we suggest to introduce (temptatively) the following attributes: X_IS_SMALL and X_EXPRESSES _POSITIVE_ EMOTION.

6. Verb Valency Frames and Verb Senses in Czech Wordnet

When building Czech verb synsets we paid a systematic attention to verb valency frames. This follows from inflectional nature of Czech which displays a rich declension structure – each Czech noun (as well as adjective, pronoun, numeral) can appear in one of the seven cases: Nominative, Genitive, Dative, Accusative, Vocative, Locative and Instrumental. For verbs it means that their arguments (participants) represented by nouns or noun groups also come in the mentioned cases. Thus, we have decided to include valency frames into Czech verb synsets. The frame displays the information about the corresponding morphological cases that are obligatorily (or optionally) associated with the given literal and also about semantic roles of the participants represented by the surface cases.

It is our strong opinion (and experience too) that due to the inflectional character of Czech the valency frames should be given for each verb synset. The usefulness of verb frames in wordnet lies especially in their ability to capture the links between the valency frames and verb senses. Moreover, in our view the valency frames should be enriched in such a way that apart from the semantic roles of the participants given symbolically they also should include the typical lexical items reflecting the most frequent lexical collocations. Thus, the valency frames have to be devised as richer structures than the case frames in their standard shape. The enriched valency frames may take the following form (for verb *otevřít/open*:

```
=1 kdo1/Ag/člověk -otevřít- co4/Pat/(láhev, pivo, krabici)
   (who1/Ag/Person -open- what4/Pat/bottle,box)

=2 kdo1/Ag/člověk -otevřít- co4/Body Part/(oči, ústa)
   (who1/Ag/Person -open- what4/Body Part/(eyes, mouth)

=3 kdo1/Ag/člověk -otevřít- komu3/Adr/člověk co4/Pat/(dveře, bránu)
   (who1/Ag/Person -open- to whom3/Adr/Person what4/Pat/(door, gate)

=4 kdo1|co1/(člověk, instituce) -otevřít- co4/Pat/(školu, obchod)
   (who1|what1/(Person,Institution) -open- what4/Pat/(school,business)
```

(Remark: =1, =2, denote the particular senses, the individual numbers indicate the surface cases 1: Nominative, 3: Dative, 4: Accusative, Czech pronoun expressions kdo1, co1, co4, komu3, indicate the surface cases directly by their forms.)

According to our experience almost no Czech lexical resource offers the valency frames together with capturing systematically the relations between the various combinations of verb arguments (verb valencies) and their senses. However, quite recently Lopatkova and Zabokrtsky [5] have started to develop Valency Lexicon of Czech (ValLex) based on PDT [6] but it is not related to Czech wordnet. Therefore we decided to have Czech valency frames in verb synsets in Czech wordnet.

At the present moment, the situation with valency frames is quite favourable as we have at our disposal reasonably large list of Czech verb surface valency frames containing approximately 15 000 items, however the links between valencies and senses have been systematically prepared for some 5 000 items so far, particularly for those being included to Czech wordnet (the estimated number of verbs in Czech is about 36 000 items). Recently a list of approx. 1 000 Czech verbs with their both surface and deep valency frames has been compiled [7].

At the last Progress Meeting of the BalkaNet Project in Sofia (May 2004) new results related to the valency frames mentioned above have been presented also for Bulgarian and Romanian. They are based on the list of approx. 1 000 Czech valency frames. The presented comparison has shown that the individual valency frames originally developed for Czech display relevant semantically universal properties and can work for Czech, English and two mentioned languages. This is highly relevant for validation of the wordnets in general, particularly, with respect to the discrimination of the verb senses. A strong hypothesis can be formulated that a respective valency frame can apply for translational equivalents (captured by the respective IDs) covering all BalkaNet languages.

7. Conclusions

To conclude we can say that the presented core of the Czech wordnet has been prepared in a good quality covering the base of the Czech word stock and can be regarded as a reliable starting point for further testing and validation. We tried to include as many polysemous lexical units as we could – they are most difficult to prepare and at the same time they represent the most frequent lexical units in language. The synsets with one sense can be processed semi-automatically and added to the present Czech wordnet rather fast and hopefully in a relatively short time.

ILRs have been integrated into Czech wordnet so far mostly using EuroWordNet relations, some links have been obtained manually but when the derivation processes are captured more thoroughly by morphological module AJKA this part of work will be partly automated. In this respect we expect that new ILRs may appear that will hold between the literals, not between synsets. Apart from this the derivational relations seem to make it possible to add derivational nests into Czech wordnet, which can be considered as special subnets within the wordnet as such.

References

[1] RYCHLÝ, P., *Corpus Managers and Their Effective Implementation*, PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2000.

[2] SEDLACEK, R., SMRŽ, P., *A new Czech morphological analyser ajka*, in *Proceedings of the TSD*, Czech Republic, 2001, 100–107.

[3] HORAK, A., SMRŽ, P., *New features of wordnet editor VisDic*, Romanian Journal of Information Science and Technology, vol. **7**, 1–2, 2004 (in this volume).

[4] PALA, K., RYCHLÝ, P., SMRŽ, P., *Corpus annotation in inflectional languages: Czech*, in *Ninth International Workshop on Database and Expert Systems Applications* (A Min Tjoa, Roland R. Wagner, ed.), IEEE Computer Society, 1998, 149–153.

[5] LOPATKOVA, M., ZABOKRTSKY, Z., *Valency dictionary of Czech verbs*, in *Proceedings of LREC*, 2002.

[6] BÉMOVÁ, A., HAJIC, J., HLADKÁ, B., PANEVOVÁ, J., *Morpho-logical and syntactic tagging of the prague dependency treebank*, *Journées ATALA – Corpus annotées pour la syntaxe, ATALA Workshop – Treebanks*, 1999.

[7] PALA, K., SMRŽ, P., *Top ontology as a tool for semantic role tagging*, in *Proceedings of LREC*, 2004 (to be published).