Derivational Relations in Czech WordNet

Karel Pala
Faculty of Informatics
Masaryk University Brno
Czech Republic
pala@fi.muni.cz

Dana Hlaváčková
Faculty of Informatics
Masaryk University Brno
Czech Republic
ydana@aurora.fi.muni.cz

Abstract

In the paper we describe enriching Czech WordNet with the derivational relations that in highly inflectional languages like Czech form typical derivational nests (or subnets). Derivational relations are mostly of semantic nature and their regularity in Czech allows us to add them to the WordNet almost automatically. For this purpose we have used the derivational version of morphological analyzer Ajka that is able to handle the basic and most productive derivational relations in Czech. Using a special derivational interface developed in our NLP Lab we have explored the semantic nature of the selected noun derivational suffixes and established a set of the semantically labeled derivational relations – presently 14. We have added them to the Czech WordNet and in this way enriched it with approx. 30 000 new Czech synsets. A similar enrichment for Princeton WordNet has been reported in its recently released version 3.0, we will comment on the partial similarities and differences.

1 Introduction

WordNets as such represent huge semantic networks in which the basic units – synsets – are linked with the ‚main‘ semantic relations like synonymy, near_synony my, antonymy, hypero/hyponymy, meronymy and others. In the EuroWordNet project (cf. Vossen, 2003) Internal Language Relations (ILR) have been introduced such as Role_Agent, Agent_Involved or Role_Patient, Patient_Involved etc., as well as the relation Derivative capturing derivational relations between synsets. The semantic nature of the derivational relations, however, was not systematically analyzed and labeled in EuroWordNet project.

If we try to label the derivational relations semantically and include them in WordNet as a result we get two level network where on the higher level we have the ‚main‘ semantic relations between synsets such as synonymy, near_synonymy, antonymy, hypero/hyponymy, meronymy and others and on the lower level there are relations like the derivational ones that hold rather between literals than between synsets.

In the highly inflectional languages the derivational relations represent a system of semantic relations that definitely reflects cognitive structures that may be related to a language ontology. Such ontology undoubtedly exists but according to our knowledge it has not been written down yet. However, for language users derivational affixes (morphemes) function as formal means by which they express semantic relations necessary for using language as a vehicle of communication. In our view, the derivational relations should be considered as having semantic nature though a question may be asked what kind of semantics we are dealing with (see Sect. 3). It has to be remarked that grammatical categories such as gender or number display a clear semantic nature.

2 Derivational Morphology in Czech

In Czech words are regularly inflected (declined, conjugated) as they express different grammatical categories (gender, number, case, person, tense, aspect etc.) using affixes. This is what is called

Balto-Slavonic Natural Language Processing 2007, June 29, 2007, pages 75–81,
Prague, June 2007. ©2007 Association for Computational Linguistics

formal morphology in Czech grammars and its description mostly deals with the system of the inflectional paradigms. Then there is a *derivational morphology* which deals with deriving words from other words, e.g. nouns from verbs, adjectives from nouns or verbs etc. using affixes again. The derivations are closely related to the inflectional paradigms in a specific way: we can speak about derivational paradigms as well (cf. Pala, Sedláček, Veber, 2003).

For Czech inflectional morphology there are automatic tools – morphological analyzers exploiting the formal description of the inflection paradigms – we work with the analyzer called Ajka (cf. Sedláček, Smrž, 2003) and developed in our NLP Lab. Its list of stems contains approx. 400 000 items, up to 1600 inflectional paradigms and it is able to generate approx. 6 mil. Czech word forms.

We are using it for lemmatization and tagging, as a module for syntactic analyzer, etc. We have also developed a derivational version of Ajka (D-Ajka) that is able to work with the main regular derivational relations in Czech – it can generate new word forms derived from the stems. Together with D-Ajka an MWE preprocessing module with the database containing approx. 100 000 collocations is exploited as well.

2.1 Derivational relations in Czech

The derivational relations (D-relations) in Czech cover a large part of the word stock (up to 70 %). Thus we are interested in describing derivational processes (see examples) by which new words are formed from the corresponding word bases (roots, stems). In Czech grammars (Mluvnice češtiny, 1986) we can find at least the following main types (presently 14) of the derivational processes:

1. mutation: noun -> noun derivation, e.g. *ryba -ryb-ník* (*fish -> pond*), semantic relation expresses location – between an object and its typical location,
2. transposition (existing between different POS): noun -> adjective derivation, e.g. *den -> den-ní* (*day ->daily*), semantically the relation expresses property,

3. agentive relation (existing between different POS): verb -> noun e.g. *učit -> uči-tel (teach -> teacher*), semantically the relation exists between action and its agent,

4. patient relation: verb -> noun, e.g. *trestat -> trestanec* (*punish ->convict*), semantically it expresses a relation between an action and the object (person) impacted by it,

5. instrument (means) relation: verb -> noun, e.g. *držet -> držák* (*hold ->holder*), semantically it expresses a tool (means) used when performing an action,

6. action relation (existing between different POS): verb -> noun, e.g. *učit -> uče-n-í* (*teach -> teaching*), usually the derived nouns are charaterized as deverbatives, semantically both members of the relation denote action (process),

7. property-va relation (existing between different POS): verb -> adjective, e.g. *vypracovat -> vypracova-ný* (*work out -> worked out*), usually the derived adjectives are labelled as de-adjectives, semantically it is a relation between action and its property,

8. property-aad relation (existing between different POS): adjective -> adverb, e.g. *rychlý -> rychl-e* (*quick -> quickly*), semantically we can speak about property,

9. property-an (existing between different POS): adjective -> noun, e.g. *rychlý -> rychl-ost* (*fast -> speed*), semantically the relation expresses property in both cases,

10. gender change relation: noun -> noun, e.g. *inženýr -> inženýr-ka* (*engineer -> she engineer*), semantically the only difference is in sex of the persons denoted by these nouns,

11. diminutive relation: noun -> noun -> noun, e.g. *dům -> dom-ek -> dom-eček* (*house -> small house -> very little house* or *a house to which a speaker has an emotional attitude*), in Czech the diminutive relation can be binary or ternary,

12. augmentative relation: noun -> noun, e.g. *bába -> bab-izna* (*beldame -> hag*), semantically it expresses different emotional attitudes to a person,

13. prefixation: verb -> verb, e.g *myslet -> vy-myslet* (*think -> invent*), semantically prefixes in Czech denote a number of

different relations such as distributive, location, time, measure and some others. We will not be dealing with this topic here, it calls for a separate examination (project),

14. possessive relation (existing between different POS): noun -> adjective *otec* -> *otcův* (*father* -> *father's*), semantically it is a relation between an object (person) and its possession.

We should mention two more relations that are sometimes regarded inflectional but in our view they belong here as well: gerund relation - verb -> adjective: (*bojovat* ->*bojující, fight* -> *fighting*) and passive relation – verb -> adjective (passive participle): (*učit* -> *učen*, *teach* -> *taught*).

These 14 (+2) relations have been taken as a starting point for including derivational relations in Czech Wordnet. The main condition for their including is whether they can be generated by the derivational version of the analyzer Ajka. In this way we have been able to obtain automatically a precise specification what literals are linked together. It was also necessary to introduce the labels for the individual relations in a more systematic way. As a result we have obtained the following list of 10 derivational relations with their semantic labels that are given in the brackets and hold between the indicated POS:

1. deriv-na: noun -> adjective (property)
2. deriv-ger: verb -> adjective (property)
3. deriv-dvrb: verb -> noun (activity as a noun)
4. deriv-pos: noun -> adjective (possessive relation)
5. deriv-pas: verb -> adjective (passive relation)
6. deriv-aad: adjective -> adverb (property of property)
7. deriv-an: adjective -> noun (property)
8. deriv-g: noun -> noun (gender relation)
9. deriv-ag: verb -> noun (agentive relation)
10. deriv-dem: noun -> noun (diminutive relation)

The location and patient relation will be included in CzWn when the D-Ajka will be able to handle them (in the near future).

2.2 Derivational nests – subnets

If we have a look at the data, i.e. at the list of Czech stems and affixes and try to see how the just described relations work we obtain the typical derivational clusters – we will prefer to call them derivational nests (subnets). To illustrate their regularity we adduce an example of such nest for the Czech roots – *prác*/*prac*- (*work*). The main relations holding between these roots and the corresponding suffixes are:

roots: -prác-/-prac-e-

deriv-act - *prac-ova-t* (*to work*)

deriv-loc1- *prac-ov-iště* (*workplace*)

deriv-loc2 - *prac-ov-na* (*study*)

deriv-ag1- *prac-ov-ník* (*worker*),

deriv-g - *prac-ovn-ice* (*she-worker*),

deriv-ag2 - *prac-ant* (*plodder*)

deriv-ger - *prac-uj-ící* (*working - person*)

deriv-pro - *prac-ov-ní* (*professional, working*)

deriv-pro - *prac-ov-i-t-ý* (*diligent, hardworking*)

deriv-pro - *prac-ov-i-t-ost* (*diligence*)

The proposed labels are not final yet – the number of the productive derivational relations that have to be examined in Czech is larger, certainly up to 15. Number of the derivational suffixes in Czech is higher – more than 80.

At the moment the derivational Ajka is not able to generate the full nests automatically but we continue processing the remaining Czech derivational suffixes for this purpose.

2.3 Processing derivational suffixes

So far we have not said much about the affixes, i.e. prefixes, stem-forming infixes and suffixes used in derivations. In this analysis we pay attention mainly to the suffixes, prefixes are related mostly to verbs and in this sense they represent a separate and rather complicated derivational system. Infixes or intersegments are basically covered by the list of stems – instead writing rules for changes in stems we just use more variants of one stem. But the root analysis is possible and if we want to describe the derivational processes in Czech as completely as possible we have to return to them.

As starting data we have used a list of noun stems taken from the stem dictionary of the D-Ajka analyzer – their number is approx. 126 000. The derivations have been analyzed by means of the web interface developed just for this purpose. Noun derivations are performed in the three basic steps:

1. a set of words is defined by means of the (prefix), suffix and morphological tag;
2. defining a derivational rule – typically a substitution of morphemes (suffixes) at the end of the word;
3. manual modification of the results – usually correcting or deleting cases that cannot be regarded as properly derived forms though they may follow the given rule.

An example of the derivational analysis for Czech sufix *–ík*: it occurs with the nouns denoting agent or instrument (means), e.g. *zed-n-ík (bricklayer)* or *kapes-ník (hankerchief)*.

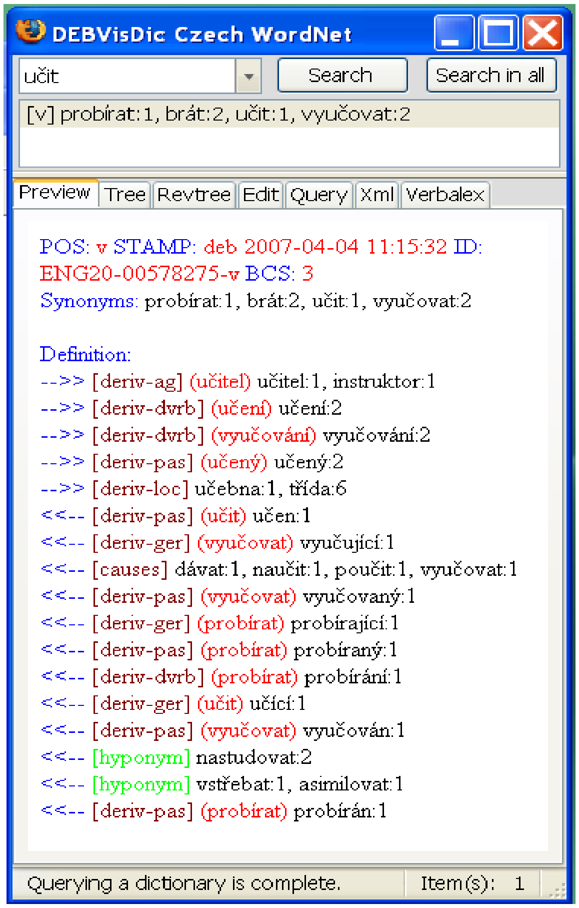
First we want to derive agentive nouns: so we enter the suffix *–ík* and tag k1gM (noun, masculine animate) and generate the list of all words ending with *-ík*. The output is a list of 1210 nouns including proper names (from the original list of 126 000 Czech nouns). To obtain instrument nouns we input the tag k1gI (noun, masculine inanimate). As an output result we get a list of 715 nouns including proper names. The number of all words ending with suffix *-ík* (disregarding the grammatical tag) in stem dictionary of Ajka is 1830. The difference in the given numbers follows from the homonymy, for instance, some nouns can be both masculine animate and masculine inanimate (e.g. *náčelník* can denote – *chief* as well as *čelenka – headband*. Such cases have been checked manually.

In a similar way we have processed 22 Czech derivational suffixes and as a result we have obtained a detailed classification of the indicated derivations capturing agentive, instrumental, location and also resultative relations, for instance *spálit* -> *spálenina* (*to burn* -> *a burn*) which has not been mentioned before. At the same time the complete lists of all stems with the indicated suffixes together with labeling their semantic relations between the stems and respective suffixes was obtained as well. For the processed suffixes the coverage is complete (with regard to the list of 126 000 of the Czech noun stems).

Thus using the described procedure we are trying to find pairs of the word forms in which the first one is considered basic and the second one derived. The direction of the derivations is not always unambiguous but the most important goal is to establish the relation itself not its direction. The cases when changes in stem take place have to be checked and added manually.

2.4 D-relations in Czech and English WordNet

In Figure 1 we show how the D. relations are implemented in Czech Wordnet. As an example we show verbal synset {učit:1, vyučovat: probírat:1, brát:2}and the similar English one {teach:1, instruct:1}). It can be seen that there is a derivational subnet with five D-relations associated to



```
[v] probírat:1, brát:2, učit:1, vyučovat:2

POS: v STAMP: deb 2007-04-04 11:15:32 ID:
ENG20-00578275-v BCS: 3
Synonyms: probírat:1, brát:2, učit:1, vyučovat:2

Definition:
-->> [deriv-ag] (učitel) učitel:1, instruktor:1
-->> [deriv-dvrb] (učení) učení:2
-->> [deriv-dvrb] (vyučování) vyučování:2
-->> [deriv-pas] (učený) učený:2
-->> [deriv-loc] učebna:1, třída:6
<<-- [deriv-pas] (učit) učen:1
<<-- [deriv-ger] (vyučovat) vyučující:1
<<-- [causes] dávat:1, naučit:1, poučit:1, vyučovat:1
<<-- [deriv-pas] (vyučovat) vyučovaný:1
<<-- [deriv-ger] (probírat) probírající:1
<<-- [deriv-pas] (probírat) probíraný:1
<<-- [deriv-dvrb] (probírat) probírání:1
<<-- [deriv-ger] (učit) učící:1
<<-- [deriv-pas] (vyučovat) vyučován:1
<<-- [hyponym] nastudovat:2
<<-- [hyponym] vstřebat:1, asimilovat:1
<<-- [deriv-pas] (probírat) probírán:1

Querying a dictionary is complete.    Item(s):  1
```

Figure1: D-relations in Czech WordNet

{učit:1, ...} (in fact 14 but they are repeating with other literals in the synset as well). Each D-relation is labeled semantically so we have here the following D-relations: agentive, location, deverbative, gerund, passive – the last two may be characterized as more morphological (surface, see Sect. 2.1) than the first three.

In Princeton WordNet 3.0 we can observe the following three D-relations associated with the synset {teach:1, learn:5, instruct:1}

S: (v) **teach**, learn, instruct (impart skills or knowledge to) *"I taught them French"; "He instructed me in building a boat"*
derivationally related form

- W: (adj) teachable [Related to: teach] (ready and willing to be taught) *"docile pupils eager for instruction"; "teachable youngsters"*
- W: (n) teacher [Related to: teach] (a person whose occupation is teaching)
- W: (n) teacher [Related to: teach] (a personified abstraction that teaches) *"books were his teachers"; "experience is a demanding teacher"*
- W: (n) teaching [Related to: teach] (the activities of educating or instructing; activities that impart knowledge or skill) *"he received no formal education"; "our instruction was carefully programmed"; "good classroom teaching is seldom rewarded"*
- W: (adj) instructive [Related to: instruct] (serving to instruct or enlighten or inform)
- W: (n) instruction [Related to: instruct] (the activities of educating or instructing; activities that impart knowledge or skill) *"he received no formal education"; "our instruction was carefully programmed"; "good classroom teaching is seldom rewarded"*
- W: (n) instructor [Related to: instruct] (a person whose occupation is teaching)

It is not surprising that the full agreement between Czech and English D-relations includes only the agentive relation (*teach* -> *teacher*) and gerund relation (*teach* -> *teaching*). The relation *teach* -> *teachable* is not included among Czech relations at the moment but it will be easy to add it. The location relation is missing in English and also some others characterized usually as morphological. We included them in Czech WordNet – they belong to the set of the Czech derivational relations.

If we compare semantic labeling of the D-relations in both Wordnets we observe that they are more explicitly formulated in Czech Wordnet. The question that remains to be answered is how the different senses may be or are reflected in the individual derivations. In PWN 3.0 the derivation *teach* – *teacher* is given twice because there are two different senses of *teach* in PWN 3.0. In our view, it is enough to give this derivational relation just once because it is agentive in both cases. Of course, in Czech there are frequent cases like *držet* -> *držák* (*hold* -> *holder*) and *držet* -> *držitel* (*hold* -> *holder*) where the first one is instrument relation and the second agentive but in Czech the different suffixes have to be used (*-ák* vs. *-tel*) indicating a difference in gender as well (masculine inanimate vs. masculine animate).

3 What is the nature of the D-relations?

In the previous sections we have introduced the labeling of the Czech D-relations. The question may be asked what is the real nature of D-relations, whether it is semantic or rather morphological (formal). The D-relations exist between morphemes, typically between stems and corresponding suffixes. This formal feature makes them different from the relations between sentence constituents, as e.g. between verbs and their arguments. However, the main criterion for us is whether the particular relation affects meaning irrespective of its formal realization.

If we apply this criterion to the D-relations discussed above, such as deriv-ag, deriv-loc, deriv-instr, deriv-g, deriv-dem, deriv-pos, deriv-pro, we definitely come to the conclusion that their nature is semantic.

Then there are relations like deriv-an, deriv-na, deriv-dvrb, deriv-ger, deriv-aad, deriv-pas that are sometimes characterized as morphological only and their semantics is left aside. The first two relations hold between nouns and adjectives and both denote properties (e.g. deriv-an: *nový* -> *novost* (*new* -> *newness*)), but we have to take into account that there is something that may be called semantics of the parts of speech, i.e. in one case property is expressed by the adjective and then by the noun which is derived from the adjective. Deriv-na denotes property as well but here the adjective is derived from noun as in *boj -> bojovný* (*fight -> combative*). The relation deriv-dvrb exists between a verb and noun, e.g. *učit -> učení* (*teach -> teaching*), and it denotes action which is first expressed by the verb and then by the deverbative noun. We can say that in these cases the only difference lies in the optics of the individual parts of speech but this difference should be understood as semantic as well. However, it should be remarked once more that quite often the differences in the semantics of the parts of speech are not treated as truly semantic.

If we have look what standard Czech grammars (see e.g. Karlík et al, 1995) say about the semantics of the parts of speech we find the formulations such as: nouns denote independent entities, i.e. persons, animals and things and also properties and actions. Verbs then denote states and their changes and processes (actions) and their mutations. These descriptions certainly refer to the semantics of the nouns and verbs. They are usually followed by the explanations about morphological processes, i. e. usually derivations by which some parts of speech are formed from the others, as we have described them above. What is relevant and what is missing in the standard grammars are more detailed and extensive semantic classifications of nouns, verbs, as well as adjectives and numerals. They are beginning to appear only recently and have the form of ontologies – the standard grammars do not use this term at all.

Until we have such semantic classifications describing semantic relations between the individual parts of speech we can hardly have a full picture that is necessary for automatic processing of the derivational relations.

This issue certainly calls for a more detailed examination, which would be a topic for another paper.

4 The implementation of D-relations in Czech WordNet

The existing software tools (e.g.Visdic, cf. Horák, Smrž, 2004) used for building Wordnet databases standardly work with semantic relations between synsets and they treat them as atomic units. In fact, the synsets are not atomic as such and they consist of the smaller units called literals, i.e. for instance the synset {teach:1, instruct:1} contains two literals (lemmas).

If we want to deal with the D-relations automatically we immediately face a problem: because of their nature they typically hold not between synsets but between literals that as a rule belong to the different synsets, e.g. teach:1 and teacher:1. Therefore we need a tool that is able to define and create derivational links between the literals. According to our knowledge the only tool that can do this is DEBVisdic editor and browser developed at our NLP Lab at FI MU (cf. Horák, Pala, 2006, it can be downloaded from: http://nlp.fi.muni.cz/projekty/deb2/clients/).

We have used it for the implementation of the D-relations in Czech WordNet (the result is shown in Sect. 2.4). The DEBVisdic tool is now used for representing and storing all the semantic relations including the D-relations. It is also exploited for building Wordnets in other languages such as Polish, Slovenian, Hungarian and others.

In our view, the way in which the D-relations (and other relations as well) are represented relevantly depends on the software tools used. This can be demonstrated if we compare the representation of the Czech D-relations in DEBVisdic with the one in PWN 3.0 (see Sect. 2.4) which appears to be less explicit and rather verbose. This also means that the representation used in PWN 3.0 will be probably less suitable for possible applications.

5 The results

As we said above after processing all D-relations by the derivational Ajka we have added the derived literals (lemmas) to the Czech WordNet. The final result – the number of the literals generated from the individual D-relations is given below together with their semantic labels:

deriv-na	641	(property, noun -> adj)
deriv-ger	1951	(property, verb -> adj)
deriv-dvrb	5041	(action, verb -> noun)
deriv-pos	4073	(possessive, noun -> adj)
deriv-pas	9801	(passive, verb -> adj)
deriv-aad	1416	(property, adj -> adverb)

deriv-an ………...1930 (property, adj -> noun)

deriv-g …………2695 (gender, noun -> noun)

deriv-ag …………186 (agentive, verb -> noun)

deriv-dem ………3695 (diminutive, noun -> noun)

Total ………… 31429 literals

These numbers also tell us how productive the particular relations are. Note that the most frequent is passive relation which is followed by the deverbative (action) relation. The third most frequent relation is a possessive one. It would be interesting to examine what these facts can tell us about semantic structure of texts.

6 Conclusions

In the paper we present the first results of computational analysis of the basic and most regular D-relations in Czech using derivational version of the morphological analyzer Ajka.

Though the analysis is far from complete at the moment the number of the generated items has led us to the decision to include them in Czech WordNet and enrich it considerably with the derivational nests (subnets). In our view, this kind of enrichment makes Czech WordNet more suitable for some applications, namely for searching.

The second and even more important reason for doing all this is a belief that the derivational relations and derivational subnets created by them reflect basic cognitive structures existing in natural language. More effort is needed for exploring them from the point of view of now so popular ontologies – they certainly offer a formal ground (they are expressed by the individual morphemes) for natural language based ontologies.

We have also included a brief comparison with the recently released Princeton WordNet 3.0 which now contains derivational links for English as well. As we expected the comparison confirms the known fact that English as an analytic language is much poorer with regard to the derivational relations than the inflectional ones.

From the technical point of view PWN 3.0 is still not using the representation in XML format (as DebVisdic does) and this, we think, in certain degree limits the possibilities to express some of the links in a standard way. The present web interface where Princeton WordNet 3.0 can be browsed: http://wordnet.princeton.edu/perl/webwn) does not seem to be able to work directly with the links between literals.

On the other hand, we are well aware that adding D-relations to PWN 3.0 is very stimulating and useful though it will be quite demanding to establish the derivational links between English and other languages (through Interlingual Index). This makes it a new challenge for the whole WordNet community.

Acknowledgements

The research was supported by the grant projects GA 201/05/2871, 1ET100300419 and NPVII 2C06009.

References

Horák A., Pala K., Rambousek A., and Povolný M. 2006. First version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International WordNet Conference – GWC 2006*, p. 325-328, Jeju, South Korea, Masaryk University, Brno.

Horák A., Smrž P. 2004. Visdic – WordNet Editing and Browsing Tool, Proceedings of the 2nd GWC, Brno, Masaryk University.

Karlík P. et al. 1995, Příruční mluvnice češtiny (Every day Czech Grammar), Nakladatelství Lidové Noviny, Prague, pp. 229, 310.

Pala K., Sedláček R., Veber M. 2003. Relations between Inflectional and Derivation Patterns, Proceedings of EACL, Budapest.

Petr J. et al. 1986. *Mluvnice češtiny 1*, Praha: Academia.

Sedláček R., Smrž P. 2001. A New Czech Morphological Analyser Ajka. Proceedings of the 4th International Conference on Text, Speech and Dialogue, Springer Verlag, Berlin, s.100-107.

Vossen P. 2003. EuroWordNet General Document, Version 3, University of Amsterdam.

Web address of the Princeton WordNet 3.0: http://wordnet.princeton.edu/perl/webwn.