Enriching WordNet with Derivational Subnets

Karel Pala and Radek Sedláček

Faculty of Informatics, Masaryk University,
Botanicka 68a, 60200 Brno, Czech Republic
{pala, rsedlac}@fi.muni.cz

Abstract. In this paper, we deal with the derivational (word formation) relations as they are handled by the Czech morphological module Ajka. First, we show that they represent empirically well-based semantic relations forming small semantic networks, and then we solve the problem how to integrate them into lexical database such as (Czech) WordNet. In this respect we examine the relation between the derivational relations and semantic roles (deep cases) defined as Internal Language Relations in EuroWordNet. An attempt is made to match up the inventory of the semantic roles in EWN with the derivational (semantic) relations. We also use a tool called SAFT that can process a raw (corpus) text in such a way that it uses module Ajka to find links relating the WordNet senses to the noun and verbal lemmata obtained from the raw (corpus) text. This technique allows us to enrich Czech WordNet with the derivational subnets and represent them in a XML format. The result is a new kind of the semantic network, which consists of two layers, upper and lower. The result is a more **powerful** and efficient resource for applications like tools for WSD, web searching or information extraction.

1 Derivational Relations as Semantic Networks

For computer processing highly inflected language like Czech it is necessary to have a high quality morphological module that can perform lemmatization of a given word form and yield all the grammatical categories that are carried by the word form. Such a tool for Czech is a morphological analyzer and generator called Ajka developed in NLP Lab at FI MU (Sedláček, 2001, 2004). Other tools exist for Czech as well (Hajič, 2004) but we prefer Ajka for its properties—it is able to deal with derivational relations automatically.

Ajka is based on the system of the (approx.) 2000 inflectional paradigms, contains about 350 000 Czech stems and is able to generate about 5,7 million Czech word forms. Its coverage/recall for Czech is about 96 % (tested on the corpus All containing 640 mil. Czech word forms and implemented in the NLP Lab at FI MU). It is based on the 'paradigmatic' model of morphology and though it has been primarily devised for Czech its engine can work also with other synthetic languages (such as Slavonic, e.g. Slovak, Serbian, Russian) as well as with analytic ones – there are versions for English, German, French, Dutch, Spanish, Italian.

As we said the morphological module Ajka captures not only the inflectional relations but also the derivation ones (word formation relations). For Czech we know

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 305 – 311, 2005.
© Springer-Verlag Berlin Heidelberg 2005

that approximately 67% of the word stock is obtained by means of the word formation and that the derivation of the new words is highly regular and can be described by the formal rules. The word formation rules have been recently integrated into Ajka (Sedláček, 2004) so that it is now able to generate and recognize word **derivational networks** (subnets) automatically. An example of such derivational nest is given in Fig. 1 (both for English and Czech, actual output from Ajka looks slightly different):

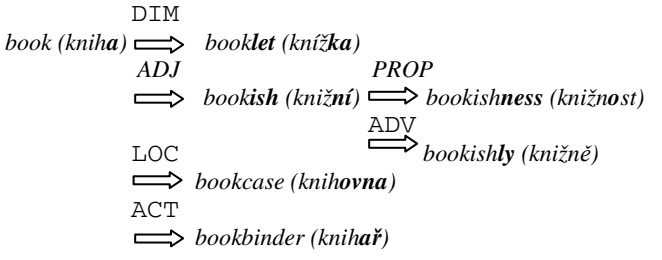

Fig. 1. Word derivation graph (subnet) for the root *book*

One can observe that the semantic relations between the above items are empirically well founded since they can be deduced from the given root or stem and the respective suffixes indicated in bold face (or prefixes as well). They represent a part of the language knowledge that speakers have. The derivational semantic network can be formally represented as a graph with one or more roots (see Figure 1). Its nodes represent the individual lemmata or word forms and edges can be labelled by the corresponding semantic relations, which follow from the relation between roots and derivational suffixes. The following example with *home* shows that derivational relations in English are reasonably rich, and in fact, follow almost the same principles as in Czech which, however, is more regular and productive (the data comes from BNC):

As a NOUN	In Real Estate domain:	As an ADJ	As an ADV	As a VERB
Home *homework* *homeland* *homecoming* *homeboy* *homestead* *homecare* *homebase* *homebrew* *(the) homeless* *homelessness* *homeliness*	*homeowner* *homebuyer* *homeloan* *homecover*	*homesick* *home-made* *home-based* *homegrown* *homeless* *homely*	*home* (e.g. *go home*, *I am at home*) *homeward* *homewards*	*home* (e.g. *If you want to home to a beacon…*) phrasal verb: *to home in on*

2 Are Derivational Relations Semantic?

In Czech linguistic works related to the word formation (Dokulil, 1962) the derivational relations are treated as a special group of the relations that express "semantic relations sui generis", i.e. they are understood as different from other „standard" semantic relations based on the sentence constituents. In our opinion, this differentiation can be empirically justified since the derivational relations are in fact morphological relations whereas "standard" semantic roles are viewed as the relations resting on sentence constituents.

However, if we have a look at the collection of the Czech derivational suffixes (67 in Ajka), we can distinguish various types of the derivational relations expressing the particular semantic relations as e.g. agentive (*to teach -.teacher*) or expressing property (*home – homelessness*). According to our intuition they can be seen as similar to other semantic relations usually characterized as "semantic roles" or "semantic cases" but they are obtained in a different way, i.e. derivationally (morphologically). While the semantic roles are typically associated with verbs and their arguments, the derivational relations hold between the four open parts of speech (in many languages), i.e. we have derivational pairs like *noun – adjective* or *adjective – noun*, *noun – verb* or *verb – noun*, *noun – noun*, *adjective – adverb.* Thus, formally they go across the individual parts of speech being XPOS relations. Though it requires more complete examination of the derivational data, the intuition is that basically there should not be an essential difference between "derivational" semantic relations and "sentence" semantic relations resting on predicate-argument structure of verbs. From the cognitive point of view the Occams Razor principle supports this intuition as well.

Below we are proposing a labelling (tagging) that can be used for the individual derivational relations and capture their semantic nature. It is tentative and it should be further refined when the empirical data becomes more complete. The labels in the list below are experimentally sub-classified (by numbers showing more detailed semantic differences) but it is not the only possible solution.

The tentative list of the derivational relations below is based on the rich Czech data but if the corresponding semantic relations can be considered rather universal (we think so), then we are convinced that they can be applied also in other languages such as English or German (not speaking about Slavonic ones).

- AG0, agent performing an action: *teach**er*** (*učit**el***) from *to teach* (*učit*),
- AG1, agent producing an object: *glass**maker*** (*skl**ář***) from *glass* (*sklo*),
- PROPB0, ownership of an object: *farmer* (*statk**ář***) from *farm* (*statek*),
- PROPB1, pertaining to an object: *villager* (*vesnič**an*** from *village* (*vesnice*),
- INS, means or instrument by which an action is performed: *excavat**or*** (*rypa**dlo***) from *to excavate* (*rýpat*),
- PAT, patient of an action: *prison**er*** (*vězeň*) from *imprison* (uvěznit),
- RES, result of an action: *printed **copy*** (***výt**isk*) from *to print* (*tisknout*),
- PROP (XPOS), property expressed by noun: *quick**ness*** (*rychl**ost***) from adjective *quick* (*rychlý*),
- PROP1, property, *dilig**ent*** (*pil**ný***) from *diligence* (*píle*),

- ACT (XPOS), action verb – noun: *the fall, falling* (*pád, padá**ní***) from *to fall* (*padat*),
- ACTPROP, property changing to an action: ***become*** *green* (*zelen**at***) from *green* (*zelený*),
- PROPMANN (XPOS), property of the action, i..e. manner: *quick**ly*** (*rychl**e***) from *quick* (*rychlý*),
- PROPDIM, diminutive relation: *book**let*** (kníž**ečka**) from *book* (*kniha*),
- PROPAUG, augmentative: ***big*** *oak* (*dub**isko***) from *oak* (*dub*),
- PROPGEN, shift of gender: ***female*** *teacher* (*učitel**ka***) from *teacher* (*učitel*), PROPYOUNG, young animal: *lion* ***cub*** (*lví**če***) from *lion* (*lev*),
- POSS, possessive, *father**'s*** (*otc**ův***) from *father* (*otec),*
- LOC, location: *battle**field*** (*boj**iště*** from *battle* (*boj*).

3 Adding Semantic (Derivational) Subnets into WordNet

As it follows from the above, we list above 17 derivational (semantic) relations that are morphologically well justified by the respective suffixes or morphemes (in English). The complete list will be a bit larger and the labelling is still tentative but the main point is that the indicated relations have a firm empirical (and formal) base. To prove the basis of the abovementioned intuition concerning the unity of the semantic relations we find it is useful to compare them with an inventory of semantic roles, particularly with the semantic roles that have been defined within the set of the Internal Language Relations introduced in EuroWordNet (Vossen, 1999). We find the following 15 roles there:

- ROLE_AGENT – INVOLVED_AGENT
- ROLE_PATIENT – INVOLVED_PATIENT
- ROLE_INSTRUMENT – INVOLVED_INSTRUMENT
- ROLE_LOCATION – INVOLVED_LOCATION
- ROLE_SOURCE_DIRECTION
- ROLE_TARGET_DIRECTION
- STATE_OF – BE_IN_STATE
- CAUSES – IS_CAUSED_BY
- HAS_SUBEVENT – IS_SUBEVENT_OF
- XPOS_NEAR_SYNONYM
- XPOS_NEAR_ANTONYM
- ROLE_RESULT – INVOLVED_RESULT
- IS_MANNER_FOR – IN_MANNER
- DERIVES – DERIVED FROM
- DERIVATIVE (defined in PWN v.2, not in EWN)

Obviously, roles like AGENT, PAT, INSTR, RES, LOC, MANN can be found in both lists but still, there is a question **how similar** they are. We have to be aware of the fact that ILRs are not always associated with the synsets while the derivational relations are always associated with the literals representing the individual items

within the synsets (being XPOS relations). In this way with derivational relations we obtain denser network containing not more relations but between more lexical items.

Some ILRs, e.g. DIRECTION, CAUSES, SUBEVENT, do not occur according to our knowledge (at least in Czech) as derivational so they can be kept and used in the same way as in EWN. The role SUBEVENT can be exploited to capture the aspect relations like Perfective – Imperfective – Iterative, which in Czech and other Slavonic languages are not treated as derivational but morphological, aspect is an obligatory grammatical category that has to be expressed by each Czech (Slavonic) verb. E.g. *přečíst* (*to read to the end, read through*) can be considered as a subevent of *číst* (*read*), but the category of the aspect is not so broad in Czech, so this is rather a tentative solution. In Czech WordNet we record aspect pairs (Perfective – Imperfective) associated with the individual verbs. The iterative verbs are obtained directly from Ajka through the respective derivational relation.

The special case is the relation DERIVED which was introduced into EWN to capture derivational relations existing in some EWN languages, however, according to our knowledge it was not elaborated in the way we do it here.

In fact, the role DERIVED was designed to cover any derivational relation that can occur between two synsets or literals, and in this respect, it is too general. However, it should be noted that in Princeton WordNet v.2 (PWN2) there is a relation DERIVATIVE which covers derivational relations between nouns and verbs (*teach – teacher*), adjectives and adverbs (*quick – quickly*) but not relations between nouns and adjectives like *stupidity – stupid*). Thanks to multilingual WordNets as in EuroWordNet or Balkanet the relation DERIVATIVE as it exists in PWN2 can be translated into other languages that are linked to English via Interlingual Index (ILI). But obviously, it can also work the other way around, i.e. if e.g. rich Czech derivational relations (together with their semantic labels) are integrated into Czech WordNet it is possible to exploit ILIs in another direction, i.e. from Czech to English and "derivational" semantic relation can be reflected in English as well.

Originally, the ILRs have been employed in the process of connecting hyperonyms with their respective hyponyms and holonyms with their meronyms, however the XML representation of the ILRs fundamentally allows us to capture any type of general relations.

The important result is: as we have indicated above – thanks to module Ajka we are able to work with the derivational relations automatically. Therefore, we can introduce them into Czech WordNet and exploit them in various ways there automatically as well. The derivational relations also can help considerably in a more reliable discrimination of the individual senses, which are sometimes too fine-grained (especially in PWN2).

4 Morphological Interface for Czech WordNet – Saft

WordNet synset literals are naturally stored as lemmata. That is why we cannot use plain text as an input stream for any kind of analysis. It is obvious that if we have large amount of data to be semantically tagged, we cannot use WordNet as it stands (at least in Czech).

The Ajka module mentioned above can be fully exploited as a bottom module for other applications. We want to exploit its ability to find a lemma for each word form in a text and associate it with its derivational subnet as we demonstrated above. The only problem is that Ajka as such is limited in one relevant respect: it can only process the words one by one as separate units.

For this purpose we have implemented a tool that employs Ajka and handles the multi-word expressions (collocations). It is named Mwe (Svoboda 2003) and uses Ajka as its bottom module. It recognizes multi-word expressions (MWE) that occur in Czech WordNet and many others. They are: collocations (*cumulative shot, diamond dust, dipterous insect*), proper, geographical and other names (*Albert Einstein, Lisabon, Kuril Islands, Matrix Reloaded*) and abbreviations (*NATO, colloq., A.D.*). Each recognized collocation is associated with its unique lemma. We can easily see from corpus texts or magazine articles that there is approximately one MWE in every other sentence.

In WordNet (both English and Czech) we find about 40 % collocations so it is obvious that if we want to semantically tag a sentence where the collocation *'cumulative shot*' occurs, we must recognize it as a whole. If simple analysis uncovers that '*cumulative*' is a lemma and '*shot*' is a lemma, and then we would manually look up for these two words, we will get plenty of false hits. It is likely that the desired synset will be among them but still the other synsets are unwanted when we process the data automatically. Collocations (or their lemmata) tend to display only one semantic unit, so if we recognize them as a whole, we practically recognize them unambiguously.

The implementation of the idea discussed above, i.e. interconnection of the functionalities of the Ajka and MWE tool with Czech WordNet can be found in the module called **Saft** (Čapek, 2004). It takes plain text as its input, parses it and recognizes the collocations, then looks them up in WordNet. Single-word expressions are processed by Ajka directly. Saft is now able to analyze and lemmatize any text in Czech and to associate relevant lemmata with the appropriate literals in Czech WordNet.

It should be noted that no attempt is made to disambiguate senses that may be associated with the individual literals. It is also possible to generate identification numbers of the synsets containing these literals and import them into VisDic (Smrž, Horák, 2004), which is the tool for storing, managing and editing WordNet lexical databases.

5 Conclusions

In the presented paper we offer the description of the selected derivational relations in Czech and their implementation in morphological analyzer Ajka, which is able to generate the derivational semantic networks). Then we show what semantic relations they capture and compare them briefly with the ILRs as they are defined in EuroWordNet. The comparison leads us to the conclusion that they are in many respects similar if not the same: ILRs are implicitly associated with the sentence constituents whereas the derivational relations (DR) rest on the morphological relations.

The derivational relations are labelled semantically and they are presented in the list containing 17 semantic relations. The list of ILRs from EuroWordNet contains 15 relations.

Then we show how the DR can be integrated into Czech WordNet. A tool called Saft is mentioned that makes it possible to process a free (corpus) text and search both for the individual synsets and literals linked with DRs. This does not mean that we do semantic disambiguation; the described processing is only a necessary first step that has to be done in any case.

As a result we obtain a Czech WordNet in which we have two levels of the semantic relations—the first one are ILRs and the second one are DRs. They are more subtle and detailed than ILRs, thus they yield more powerful resource for Information Extraction and Web searching. In this sense DRs represent a subnet that relates the individual literals above the standard synonymy/antonymy and hypero/hyponymy relations.

References

1. Baker, C. F., Fillmore, Ch. J., Lowe, J. B.: *FrameNet Project*, in: Proceedings of the Coling-ACL, Montreal, Canada (1998).
2. Dokulil, M. (1962) Tvoření slov v češtině 1 (Word-Formation in Czech 1). Nakladatelství ČSAV. Prague
3. Fellbaum, Ch.: (Ed.) WordNet: An Electronic Lexical Database, MIT Press (1998).
4. Horák, A., Smrž, P. New Features of WordNet editor VisDic, in: RomanianJournal of Information Science and Technology, Vol. 7, No 1-2, 2004, pp.201-214.
5. Klímová, J., Pala, K. (2000) Application of WordNet ILR in Czech Word-Formation. In Proceedings of LREC 2000. p. 987-992. ELRA.
6. Levin, B.: English Verb Classes and Alternations: a preliminary investigation, The University of Chicago Press, (1993).
7. Lopatková, M., Žabokrtský, Z.: Valency Dictionary of Czech Verbs, ELRA (2002).
8. Mráková-Žáčková, E.: Partial Parser DIS/VADIS (for Czech), Ph. D. Dissertation, Faculty of Informatics, Masaryk University, Brno (2002).
9. Pala, K., Rychlý, P., Smrž, P. (1997) DESAM—Annotated Corpus or Czech. In Proceedings of SOFSEM 97. Heidelberg: Springer Verlag. pp. 523--530.
10. Sedláček, R., Smrž, P.: A New Czech Morphological Analyser ajka, *Proceedings of TSD 2001*, Springer-Verlag, LNAI 2166, p.100-107.
11. Sedláček, R., Čapek, T., Svoboda, L.: Morphological Analysis and Czech WordNet, abstract of the non-published paper.
12. Smrž, P., Rychlý, P. (2001) Finding Semantically Related Words in Large Corpora. In Proceedings of TSD 2001. Berlin: Springer-Verlag, p. 108-115. LNAI 2166.
13. Vossen, P.: (Ed.), EuroWordNet: a multilingual database with lexical semantic networks for European languages, Kluwer Academic Publishers, (1999), Dordrecht