# Application of WordNet ILR in Czech Word-formation

# Jana Klímová\*, Karel Pala†

\*Institute of the Czech Language Academy of Sciences of the Czech republic Letenská 4, 118 51 Praha 1, Czech republic jana.klimova@ff.cuni.cz

<sup>T</sup>Dept. of Information Technologies, Faculty of Informatics, Masaryk University Brno Botanická 68a, 600 00 Brno, Czech Republic pala@fi.muni.cz

#### Abstract

The aim of this paper is to describe some typical word formation procedures in Czech and to show how the internal language relations (ILR) as they are introduced in Czech WordNet can be related to the chosen derivational processes. In our exploration we have paid attention to the roles of agent, location, instrument and subevent which yield the most regular and rich ways of suffix derivation in Czech. We also deal with the issues of the translation equivalents and corresponding lexical gaps that had to be solved in the framework of EuroWordNet 2 (confronting Czech with English) since they are basically brought about by verb prefixation (single, double, verb aspect pairs) or noun suffixation (diminutives, move in gender). Finally, we try to demonstrate that the mentioned derivational processes can be employed to extend Czech lexical resources in a semiautomatic way.

# 1. Background - EuroWordNet

In the present paper it is our aim to show how the techniques developed in the framework of EuroWordNet 1,2 (Vossen et al, 1999) can be employed to enlarge Czech lexical database of the WordNet type and that a part of this task can be done in a semiautomatic way.

Particularly, we are concerned with the internal language relations (ILR) as they are introduced and used in EuroWordNet 1,2. They represent a tool which makes it possible to have the roles like agent, patient, instrument, location, direction, subevent and others, marked in a lexical database and to associate them with the individual lexical units (or synsets in Czech WordNet).

Here we would like to demonstrate that an interesting link between the particular ILR's (roles) and some regular word formation processes in Czech can be found and explored in a more detailed way.

### 2. Word formation processes in Czech

In our research we have dealt with the ILR's, mentioned above and tried to relate them to the regular suffix derivation processes in Czech. In this way Czech lexical database can be considerably enriched and enlarged and its items can be associated with the respective ILR's. We also suggest a semiautomatic procedure that can serve as a means allowing us to perform this task, i.e. to modify a Czech lexical database of the WordNet type. This procedure takes advantage of Czech lemmatizer (LEMMA, Ševecek 1995, Osolsobě 1994) and the tool developed for the generation and analysis of derived words (Klímová 1999). These instruments can be used to determinate the paradigmatic and semantic aspects of the affixes, they enable to assign the grammatical categories to the words derived by certain suffixes by the means of a set of derivational patterns and rules.

# 3. Derivation rules

An example of the above mentioned procedure may have the following form:

Program LEMMA gives as one of the possible outputs for any Czech word form the information about its POS and the respective grammatical categories, which in the case of the nouns involve the gender. Take, for example, the quite frequent Czech suffix -tel (it may correspond to English -er, as in teach-er). If we have a noun derived by the suffix -tel typically expressing agentivity, say podnikatel (businessman) having the gender attribute with the value masculine animate, we can say that this noun can be labelled as ROLE AGENT. Then our rule tells us to remove the suffix -tel and as a result we get the respective verb stem having the form podnika-. In the next step the infinitive ending -t can be added and we obtain the complete infinitive form podnikat (do business) which can be correspondingly labelled as INVOLVED AGENT. In this way we get the required pair podnikatel - podnikat (businessman - do business) labelled as:

# 3.1 ROLE\_AGENT - INVOLVED\_AGENT

The suffixes representing the ROLE\_AGENT are as follows:

```
-tel: učit – uči-tel (teach – teacher), řídit – ředitel (manage – director)
```

 $-\check{c}$ :  $\check{r}idit - \check{r}idi-\check{c}$  (drive – driver),

-ec: lovit – lov-ec (hunt – hunter), chodit – chod-ec (walk – walker), střílet – střel-ec (shoot – shooter)

-ář: tunelovat – tunel-ář (tunnel – tunneler)

-ník: bojovat – bojov-ník (fight – fighter), útočit – útoč-ník (attack – attacker)

The corresponding derivation rule for all indicated suffixes may take the following shape:

- (1) if tag(word form) = k1gMnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)
- (2) then separate suffix { tel, ec, ář, ník }-> verb stem
- (3) possible change in stem, e.g. řed -> říd (for řídit ředi-tel (director))
- (4) add ending (-t) to the stem -> infinitive
- (5) ROLE\_AGENT = word form
- (6) INVOLVED AGENT = infinitive

Frequency of the above mentioned nouns in Czech National Corpus (CNC):

*-tel*: 366 933 (0,37%), from which 36 370 (10%) have the property gI (gender = inanimate)

-ec: 440 103 (0,44%), from which 168 032 (38%) have the property gI (gender = inanimate)

-ár: 162 191 (0,16%) from which 30 646 (19%) have the property gI (gender = inanimate)

-nik: 361 643 (0,36%) from which the property gI (gender = inanimate) 83 607 (23%) have

The set of such rules can be formulated and applied to other suffixes as well and they can generate pairs for the other mentioned ILR's in the similar fashion:

# 3.2 ROLE\_PATIENT - INVOLVED\_PATIENT

The rule for PATIENT would be practically the same as the one for AGENT, thus it is enough to give some examples showing what pairs of this sort can occur in Czech:

trestat – trestanec (punish – convict), odsoudit – odsouzenec (sentence – sentenced person), oběsit – oběšenec (hang – hanged man), zmocnit – zmocněnec (empower – representative), vyhnat – vyhnanec (expel – exile), pojistit – pojištěnec (insure – insured person), vystěhovat se – vystěhovalec (emigrate – emigrant)

# 3.3 ROLE\_INSTRUMENT – INVOLVED\_INSTRUMENT

The suffixes representing the ROLE\_INSTRUMENT are as follows:

-dlo: sed-at – seda-dlo (sit – seat), plavit – plavi-dlo (float – vessel), čerpat – čerpa-dlo (draw – pump)

-ák: bodat – bod-ák (stab – bayonet), šroubovat – šroubov-ák (screw – screwer)

-tko: ukazovat – ukazová-tko (point at – pointer), koukat – kuká-tko (peep – peep hole, opera glasses), klekat – kleká-tko (kneel – kneeler)

The corresponding derivation rules may take the following shape:

- (a) for neuter nouns
- (1) if tag(word form) = k1gNnSc1 (POS = noun, gender = neuter, number = singular, case = 1)
- (2) then separate suffix  $\{dlo, tko\}$  -> verb stem
- (3) possible change in stem, e.g. kuká -> kouka (for koukat kuká-tko (peep hole, opera glasses)
- (4) add ending (-t) to the stem -> infinitive
- (5) ROLE\_INSTRUMENT = word form
- (6) INVOLVED\_INSTRUMENT = infinitive
- (b) for masculine nouns
- (1) if tag(word form) = k1gNnSc1 (POS = noun, gender = masculine inanimate, number = singular, case = 1)
- (2) then separate suffix (ák) -> verb stem
- (3) possible change in stem,
- (4) add ending (-t) to the stem -> infinitive
- (5) ROLE\_INSTRUMENT = word form
- (6) INVOLVED\_ INSTRUMENT = infinitive

frequency of *-dlo* nouns in CNC: 126 009 (0,13%) frequency of *-tko* nouns in CNC: 15 807 (0,015%) frequency of *-ák* nouns in CNC: 124 079 (0,12%)

# 3.4 ROLE\_LOCATION – INVOLVED\_LOCATION

The suffixes representing the ROLE\_LOCATION are as follows:

rna: kovat – ková-rna (stud – smithy), sušit – sušá-rna (dry – drying room)

-iště: klouzat – kluz-iště (glide – skating ring), brouzdat – brouzdal-iště (paddle – paddling pool), hnízdit – hnízd-iště (nest – nesting place).

The corresponding derivation rules may take the following form:

- (a) for neuter nouns
- (1) if tag(word form) = k1gNnSc1 (POS = noun, gender = neutre, number = singular, case = 1)
- (2) then separate suffix (iště) -> verb stem
- (3) possible change in stem, e.g. kluz klouz for klouzat kluz iště
- (4) add -t to the stem -> verb
- (5) ROLE LOCATION = word form
- (6) INVOLVED\_LOCATION = infinitive
- (b) for feminine nouns
- (1) if tag(word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (2) then separate suffix (rna) -> verb stem
- (3) possible change in stem
- (4) add -t to the stem -> verb
- (5) ROLE\_LOCATION = word form
- (6) INVOLVED\_LOCATION = infinitive

frequency of -iště nouns in CNC: 61 277 (0,061%) frequency of -rna nouns in CNC: 54 123 (0,054%) Examples: hrát – hřiště (play – field), bojovat – bojiště (battle – battlefield),

# 4. New Internal Language Relations for Czech

In the paper we also touch the problem of the translation equivalents and the corresponding gaps that appear when one attempts to apply the present collection of ILR's in Czech, with comparison to English. It can be seen that the confrontation of the highly inflected language as Czech with analytic English yields an interesting view on ILR table and also shows that interesting "holes" can be found in this respect. If we take a more general view at this we can discover that regular morphological phenomena in Czech, as e.g. formation of verb aspect pairs (triples), verb prefixation (single, double), derivation of diminutives (noun suffixation), derivation of gender pairs (noun suffixation) produce lexical units in Czech that have no lexical counterparts not only in English but also in other European languages.

This exercise leads us to the conclusion that it might be useful to introduce some new ILR's into the existing ILR table, such as e.g. attributes

# 4.1 X\_HAS\_IMPF, X\_HAS\_PERF

This pair captures the aspect opposition that is typical for Czech verbs: each Czech verb belongs either to the imperfective aspect (denoting an action not limited in time) or to the perfective one (expressing an action that was finished, completed). The information about aspect is as a rule given in Czech dictionaries for each verb. As we indicated, the aspect opposition is not present in the set of ILR's used within EuroWordNet 2 since it occurs only in Czech and it would lead to some problems for the other 7 languages included presently in EuroWordNet 2. It is our opinion that the aspect opposition can be also captured by the role SUBEVENT existing in ILR's, however this role is more general and therefore it would yield overspecifications in many cases.

Examples:

Typically aspect pairs are derived by prefixes vy-, pri-, do-, za-, pro-:

učit – vyučit (teach – train), doučit (tutor), zaučit (introduce a p. to sth)

nést – vynést (carry : carry off), přinést (bring), donést (fetch), zanést (carry sth to a place), pronést (carry through)

jít – vyjít (go – go out), přijít (come), dojít (get to), zajít (go for), projít (pass through)

# **4.2** X\_HAS\_ITER for verb aspect pairs (triples)

We would not like to go into details here but, in fact, the picture concernig the aspect opposition in Czech is more complicated than we just indicated above. Particularly, some authors (Panevová, Benešová, Sgall 1971) treat the aspect not as a binary opposition but as a ternary one with values: perfective: imperfective: iterative. Thus we think it useful to offer few examples giving, we hope, at least the basic idea about the iteratives in Czech. Their formation is very regular and applies very frequently, e.g.

číst – přečíst – čítávat (read : read through : read regularly, use to read)

dělat – udělat – dělávat (do – perform – do regularly, use to do)

kupovat – koupit – kupovávat (buy – buy completely – buy regularly).

# 4.3 X MAN, X WOMAN for gender pairs

Another pair of the attributes that is not present within EuroWordNet 2 set of ILR's are gender pairs. In Czech there is a regular derivation which allows to form a feminine counterpart to each animate masculine noun denoting e.g. a profession or similar function. Nouns belonging to this class are mainly created by suffixes -ka, -yně.

#### 4.3.1 Suffix -ka

Examples: učitel – učitelka (teacher – she-teacher), kuchař – kuchařka (cook – she-cook), doktor – doktorka (doctor – she-doctor), řidič – řidička (driver – she driver), manžel – manželka (husband – wife):

The corresponding derivation rule may take the following form:

- (1) if tag(word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (2) then separate suffix  $\{ka\}$  -> noun stem
- (3) noun stem = noun
- (4)  $X_WOMAN = word form$
- (5)  $X_MAN = noun$

# 4.3.2 Suffix -yně

(a) Examples:

žák – žákyně (male pupil – female pupil), bůh – bohyně (god - goddess), ministr – ministryně (minister – female minister),

The corresponding derivation rule may take the following form:

- (1) if tag(word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (2) then separate suffix (-yně) -> noun stem
- (3) possible change in stem, e.g. for boh bůh for bůh bohyně
- (4) noun stem = noun
- (5)  $X_WOMAN = word form$
- (6) X\_MAN = noun & tag(noun) = k1gFnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)
- (b) The masculine nouns created by the suffix -ce derive the feminine nouns by replacing -ce with -yně, e.g.: vládce vládkyně (male ruler female ruler), soudce soudkyně (judge female judge)

The corresponding derivation rules may take the following shape:

- (1) if tag(word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (2) then separate suffix (-yně) -> noun stem
- (3) add suffix (-ce) to the stem -> noun
- (4)  $X_WOMAN = word form$
- (5) X\_MAN = noun & tag(noun) = k1gInSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)

The complete explanation of the algorithm would require to go into technical details.

# 4.4 Derivation of diminutives

Diminutives represent an interesting word formation category in Czech and they are created by following suffixes (see table 1). Diminutives express either smaller quantity of what is denoted by the basic word or better quality of the entity denoted by the basic word (in typical cases this involves emotional attitudes). Thus in Czech there are many regular (and regularly formed) lexicalizations that do not exist in English and other European languages (German is the exception here) and have to be translated by means of various ad hoc formed syntactic constructions. Thus we are convinced that it would be useful to have diminutives in the set of ILR's. The present attempt to capture the formation rules for diminutives that are highly regular also shows the expressive power of Czech in this respect. Including them into ILR's certainly will pay off later when other highly inflectional languages (like Slavonic ones) will be attracted to building the WordNet like lexical databases.

It is useful (see in Table 1) to distinguish primary, secondary and tertiary diminutives. The difference between these three classes is in the degree of intensifying the diminution or the emotional feature. The table and examples of diminutives below illustrate the rich productivity of Czech diminutive suffixes.

1			
	masculine	feminine nouns	neuter
	nouns		nouns
primary	-ík, -ek	-ka	-ko
suffixes			
secondary	-ček	-čka	-čko
suffixes			
	-ínek	-inka, -e/ěnka,	-inko
		-unka	
	-oušek	-u/ouš ka	-átečko
tertiary	-číček	-čička	-čičko
suffixes			
	-čínek	-čenka,-činka	-činko
	-eneček,	-enečka, -inečka	-enečko
	-íneček		
	- ulínek	-ulinka, -ulenka	-inečko,
			-ulinko

Table 1: Some Czech suffixes for the derivation of diminutives

#### 4.4.1 X IS NORM, X IS SMALL

(a) masculine diminutives

Examples:

vlak – vláček (train – little train),

hřib – hříbek – hříbeček (mushroom – little mushroom – more little mushroom),

dům - domek - domeček (house - small house - tiny house),

pytel - pytlík - pytlíček (bag - small bag - tiny bag)

#### (a1) primary diminutives:

The corresponding derivation rule for primary diminutives may take the following form:

- (1) if tag(word form) = k1gMnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)
- (2) then separate suffix (ek) -> noun stem
- (3) possible change in stem, e.g. dom dům for dům domek
- (4) noun stem = noun
- (5)  $X_{IS}_{SMALL} = word form$
- (6) X\_IS\_NORM = noun & tag = (noun) = k1gMnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)

#### (a2) secondary diminutives:

The corresponding derivation rule for primary diminutives may take the following form:

- (1) if tag(word form) = k1gMnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)
- (2) then separate suffix {eček, íček, áček, ouček} -> noun stem
- (3) possible change in stem, e.g. dom dům for dům domeček
- (4) noun stem = noun
- (5) X IS SMALL = word form
- (6) X\_IS\_NORM = noun & tag = (noun) = k1gMnSc1 (POS = noun, gender = masculine animate, number = singular, case = 1)

# (b) feminine diminutives:

#### Examples:

kniha – knížka – knížečka (book – little book – tiny book) hůl – hůlka – hůlčička (stick – cane – little stick) židle – židlička (chair – small chair)

váza – vázička (vase – small vase)

ruka – ručka – ručička (hand – small hand – tiny hand)

liška - lištička (fox - small fox)

# (b1) primary diminutives

The corresponding derivation rule may take the following shape:

- (1) if tag (word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (2) then separate suffix (-ka) -> verb stem
- (3) possible change in stem, e.g. kul koul for koule kulka
- (4) noun stem = noun
- (5) X IS SMALL = word form
- (6) X\_IS\_NORM = noun & tag = (noun) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)

#### (b2) secondary diminutives

The corresponding derivation rule may take the following shape:

- (1) if tag (word form) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)
- (1) then separate suffix {ečka, ička, ačka} -> verb stem
- (3) possible change in stem, e.g. lišt lišk for liška lištička
- (4) add ending (-e,-a) to the stem -> noun
- (5) X IS SMALL = word form
- (6) X\_IS\_NORM = noun & tag(noun) = k1gFnSc1 (POS = noun, gender = feminine, number = singular, case = 1)

# (c) neuter diminutives:

#### Examples:

jablko – jablíčko (apple – small apple)

oko − očko − očičko (eye − small eye − tiny eye)

pole – políčko (field – small field)

náměstí – náměstíčko (square – small square)

#### (c1) primary diminutives

The corresponding derivation rule may take the following form:

- (1) if tag (word form) = k1gNnSc1 (POS = noun, gender = neuter, number = singular, case = 1)
- (2) then separate suffix  $(ko) \rightarrow \text{verb stem}$
- (3) possible change in stem, e.g.  $o\check{c} ok$  for  $oko o\check{c}ko$
- (4) add ending  $\{-e,-o\}$  to the stem -> noun
- (5) X IS SMALL = word form
- (6) X\_IS\_NORM = noun & tag (noun) = k1gNnSc1 (POS = noun, gender = neuter, number = singular, case = 1)

# (c2) secondary diminutives

The corresponding derivation rule may take the following form:

- (1) if tag (word form) = k1gNnSc1 (POS = noun, gender = neuter, number = singular, case = 1)
- (2) then separate suffix {-ačko, -áčko, -ečko, -éčko, -íčko, -ičko} -> verb stem
- (3) possible change in stem, e.g.  $o\check{c} ok$  for  $oko o\check{c}ko$
- (4) add ending  $\{-e,-o,-i\}$  to the stem -> noun
- (5)  $X_IS_SMALL = word form$
- (6) X\_IS\_NORM = noun & tag (noun) = k1gNnSc1 (POS = noun, gender = neuter, number = singular, case = 1)

#### 4.4.2 X EXPRESSES EMOTIONAL ATTITUDE

The derivation rules are the same as in 4.4.1. The following examples show the diminutives expressing the positive (eventually also slightly negative) emotional attitude taken by the speaker to the entity denoted by the basic word. These diminutives represent an interesting part of Czech vocabulary (similarly as in other languages) since they relate to not quite well explored part of Czech word stock which deals with emotions and emotional attitudes. This can be partly seen in our attempts to give possible (approximative) English translations but we are well aware that this area calls for the systematic exploration:

# (a) masculine diminutives

táta – tatík – tatíček (daddy – dad – my dear dad) zloděj – zlodějík – zlodějíček (thief – small thief – pilferer, unsucessful thief)

zadek – zadeček (bottom – nice small bottom)

vtip – vtipek – vtipeček (joke – quip – short, witty joke)

voják – vojáček (soldier – toy soldier or poor soldier) děda – dědek – dědeček, dědoušek (old man, grandfather

old man, grandfather (a bit derogatory) – grandpa)
(b) feminine diminutives

máma – mamka – maminka (mother – mom – ma) bába – babka – babička (beldame – little old woman –

very small old woman or granny) slečna – slečinka (Miss – Miss (a bit derogatory) or small

sestra – sestřička (sister – little, younger sister or nurse) teta – tetka – tetička (aunt – aunt (slightly derogatory) :

#### (c) neuter diminutives

auntie)

pohodlí – pohodlíčko (comfort – my private (special) comfort)

pivo – pivko – pivečko – pivíčko (beer – wallop – my beer that I like)

peklo – peklíčko (hell – small nice pleasant hell)

nebe – nebíčko (heaven or sky – small nice pleasant heaven or sky)

psaní – psaníčko (letter – nice pleasant letter).

# 5. Conclusion

In the presented contribution we have tried to apply the approach developed within EuroWordNet 1, 2 that describes selected semantic relations between lexical units by means of ILR to Czech word formation. In our view the attempt was successful and it confirms our intuition concerning the high regularity of the word formation processes in Czech. This is demonstrated by the number of the rules presented in our text and also by the fact that they are relatively easy to formulate in a formal way.

We also offered few new ILR's that are not included in the EuroWordNet 1,2 original list of ILR's: they reflect semantic relations that have lexicalized expression, and, as it can be expected, in other (Slavonic) languages as well.

Thus, we assume that sooner or later the extended list of ILR's will be appear that may include the suggested semantic relations and possibly the new ones.

# 6. Acknowledgement

This research was supported by the GAČR, grant no. 405/96/K214.

#### 7. References

CNC (Czech National Corpus): http://ucnk.ff.cuni.cz/cnc Dokulil, M., 1962. *Tvoření slov v češtině* (Word-formation in Czech). Praha: Academia.

Klímová J., 1999. System of Computerised Czech Wordformation. In Kiefer F., Kiss G., Pajzs J.(eds.), *Papers in Computational Lexicography, Complex'99*. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Osolsobě, K., 1994. An Algorithmic Description of Czech Morphology, PhD dissertation, Brno.

Pala, K., Ševecek, P., Haldur Oim, Kadri Vider, Leho Paldre, Heili Orav, 1999. Tools and Resources for the Czech and Estonian Wordnets, Bimonthly Report for EuroWordNet-2, Version 1. Brno, Tartu, January 15, 1999.

Panevová, Benešová, Sgall, 1971. *Čas a modalita v češtině*, AUC philologica Pragensia XXXIV.

Slavíčková, E., 1975. Retrográdní morfematický slovník češtiny (Inverted Dictionary of Czech Morphemes). Praha: Academia.

Slovník spisovného jazyka českého 1-4 (Dictionary of Literary Czech), 1957-1971, Praha: Academia.

Ševeček, P., 1995. *LEMMA* (morphological analyser and lemmatizer for Czech), program in C. Brno.

Vossen, P. et al., 1999. EuroWordNet Final Report, Version 2. University of Amsterdam.