Derivational relations in English, Czech and Zulu wordnets

S. Bosch
Department of African Languages
University of South Africa
PRETORIA
E-mail: boschse@unisa.ac.za

C. Fellbaum
Department of Computer Science
Princeton University
PRINCETON, USA
E-mail: fellbaum@princeton.edu

K. Pala
Faculty of Informatics
Masaryk University
BRNO, CZECH REPUBLIC
E-mail: pala@fi.muni.cz

Abstract

Derivational relations in English, Czech and Zulu wordnets

This article investigates one kind of cross-part-of-speech relation for English, Czech and Zulu lexical resources in the form of semantic networks (wordnets). Many languages have rules whereby new words are derived regularly and productively from existing words via morphological processes. The morphologically unmarked base words and the derived words, which share a semantic core with the base words, can be interlinked and integrated into wordnets, where they typically form "derivational nests", or subnets. Efforts are described to capture the morphological and semantic regularities of derivational processes in English, Czech and Zulu to compare the linguistic mechanisms and to exploit them for suitable computational processing and wordnet construction. While some work has been done for English and Czech already, wordnets for Zulu and other Bantu languages are still in their infancy. This article illustrates how Zulu can benefit from existing work.

Opsomming

Afleidingsverhoudings in Engelse, Tsjeggiese en Zoeloe woordnette

In hierdie artikel word een tipe kruis-woordsoortverhouding vir Engelse, Tsjeggiese en Zoeloe leksikale bronne in die vorm van semantiese netwerke (woordnette) ondersoek. Baie tale beskik oor reëls waarvolgens nuwe woorde reëlmatig en produktief van bestaande woorde via morfologiese prosesse van bestaande woorde afgelei word. Die morfologies ongemarkeerde basiswoorde en die afgeleide woorde, wat 'n semantiese kern met die basiswoorde in gemeen het, kan met woordnette verbind en geïntegreer word. Hulle vorm dan tipies "afgeleide neste" of subnette. Ons beskryf pogings om morfologiese en semantiese reëlmatighede van afleidingsprosesse in Engels, Tsjeggies en Zoeloe vas te lê sodat linguistiese meganismes vergelyk en ontgin kan word vir geskikte rekenaarmatige prosessering en woordnetkonstruksie. Terwyl werk reeds in 'n mate vir Engels en Tsjeggies gedoen is, is woordnette vir Zoeloe en ander Bantoetale nog in die beginstadium van ontwikkeling. In hierdie artikel word aangedui hoe Zoeloe deur bestaande werk kan baat vind.

1. Introduction: background and motivation

Arguably the greatest challenge for natural language processing (NLP) is the discrimination of distinct senses associated with one word form. The most frequently used words are also the most polysemous, a great problem for texts that are not restricted to a specific domain like finance or medicine. Many systems rely on lexical resources: traditional dictionaries that have been converted to be machine-readable, or electronic lexicons whose formats may be designed specifically for NLP applications. *Word sense disambiguation* can be defined as the task of matching a word token in a text with the appropriate sense entry in the lexical resource, which serves as a de-facto standard of the sense inventory of a language.

Like a human user who does not know a word, automatic systems need as much information as possible about the word they are trying to disambiguate in order to distinguish it from similar but inappropriate senses. A good lexicon for NLP therefore connects as many semantically related words to one another as possible, in the form of definitions, example sentences, or semantic pointers. Wordnets are electronic lexical resources that contain all of these, and their appeal for NLP lies in the way they interconnect word forms and senses by means of semantic relations into a giant network. Each word form

with a specific meaning occupies a unique position in that network and can be identified by virtue of its particular constellation in relation to other words. Wordnets are most useful when their network is dense, i.e. when a given word is connected to many other words, as more links mean more semantic information and thus better discrimination of individual word senses. In this article we describe relations among words that are both formal and semantic, and that are useful enrichments to wordnets.

2. Morphology

The lexicon of a language is large, irregular, and open-ended, yet acquiring, storing, and retrieving lexical items is an amasing feat that human speakers perform with great ease. By contrast, morphology is a rule-based system whereby a finite number of affixes modify lexical items in regular and productive ways. Inflectional morphology, also called grammatical morphology, is concerned with affixes that have purely grammatical function. Thus, most Indo-European languages have (or once had) verbal inflection to mark person, number, tense and aspect as well as nominal inflection to indicate categories like gender, number and case.

Czech exploits what can be called a “cumulation” of functions, that is, one inflectional suffix conveys as a rule several grammatical categories; for nouns, adjectives and pronouns (as well as numerals) the categories expressed by the affixes are gender, number and case. While Czech is a richly inflected language, English has developed characteristics of an analytic language where some grammatical functions are assumed by free morphemes, for example future tense, unlike past and present, is marked by *will*. As in Czech, a single morpheme can have several grammatical functions, for instance *-s* in (*he*) *runs* is a marker of more than one grammatical function, such as present tense, third person and singular. Zulu is an agglutinative language and uses affixes to express a variety of grammatical relations and meanings. These morphemes “glue” onto stems or roots. The morphemes are not polysemous, as one of the principles that characterises agglutinating languages is the one-to-one mapping of form and meaning (Kosch, 2006:135), and each morpheme therefore conveys one grammatical category or distinct lexical meaning. In all languages, the inflected word belongs to the same form class (i.e. represents the same part of speech (POS)) as the base. Inflectional morphology encompasses regular and productive rules that are an important part of speakers’ grammar. Given a

new (or nonce) word like *wug*, even young children effortlessly produce the (inflected) plural form *wugs* (Berko Gleason, 1958).

In contrast to inflectional morphology, derivational morphology often yields words from a different form class via affixation. For example the English verb *soften* is derived from the adjective *soft* by means of the suffix *-en*; adding *-ness* or *-er* to the adjective yields the nouns *softness* and *softener*. Speakers know that *-ness* can attach to adjectives but not to nouns and that the derived nouns refer to the quality expressed by the base adjective. Derivational morphology is thus a mechanism that generates new lexical items whose meanings are systematically related to those of the base forms.

In addition to affixation, new words can be derived from existing ones by compounding. Examples are English *flowerpot, bittersweet,* and *dry-clean.* In Czech, compounding is a regular word derivation procedure but it is considered rather marginal and not so productive. An example: *česko+slovenský* (*Czecho-Slovak*) or *bratro+vrah* (*murderer of the brother*).

In Zulu, compounding is a productive and regular way of creating new words and it has its own rules, e.g.

(1) a.	*abantu*	+	*inyoni*	>	*abantunyoni*
	(people)		*(bird)*		*(astronauts)*
b.	*umkhumbi*	+	*ingwenya*	>	*umkhumbingwenya*
	(boat)		*(crocodile)*		*(submarine)*

This article focuses on derivational morphology and addresses the question as to how we can exploit its regularity to populate wordnets and to characterise both formal and semantic relations. We explore and formulate derivational rules (D-rules) allowing one to generate automatically as many word forms as possible in the three languages in focus (English, Czech and Zulu) and to assign meaning to the output of these rules. Formulating D-rules would bypass the task of compiling and maintaining large lists of base forms (stems) and would allow one to generate automatically the core of the word stock of the individual languages.

When trying to write the formal D-rules that allow us to generate new words automatically we are confronted with the problem of over- and undergeneration of derived forms. That is, the D-rules could either produce forms that are possible but not actually occurring forms (in corpora or dictionaries), or they could fail to generate all attested

forms. To avoid errors as well as undergeneration, the manual checking of the output is currently primarily relied on, but procedures that can semi-automatise this process by comparing the output of the D-rules to corpora or dictionaries are being developed. Addressing the overgeneration problem requires revisiting the D-rules and correcting those that generate ill-formed strings.

Derivational affixes in Czech and English are associated with meanings; as shown later in this article, these may be polysemous, and their attachment to polysemous base words can generate multiple new senses. In contrast to Czech and English, D-affixes in Zulu only acquire meaning by virtue of their connection with other morphemes (for example *agent*, result of an action, instrument of an action, et cetera) and cannot always be assigned an independent semantic value.

3. Derivational relations in Czech

We discuss the two main mechanisms of Czech derivational morphology, namely suffixation and prefixation. Morphemes are classified semantically.

3.1 Suffixes

First, the basic and most productive derivational relations expressed by suffixes or, more precisely, the rules describing them were formulated and integrated into the Czech morphological analyser Ajka yielding its D-version. It is an automatic tool which is based on the formal description of the Czech inflection paradigms that include declension, conjugation and comparison (cf. Sedláček & Smrž, 2001) and was developed by the NLP Centre at the Faculty of Informatics, Masaryk University, Brno. Its list of stems contains approximately 400 000 items, up to 1 600 inflectional paradigms[1] and it is able to generate approximately 6 million Czech word forms. The tool is used for lemmatisation and tagging, as a module for the syntactic analyser.

Second, we have also developed a derivational version of Ajka (D-Ajka) that is also able to work with other derivational relations in Czech: it can generate new word forms derived from the stems

1 Although 1 600 inflectional paradigms may seem excessive, this number allows us to handle the morphological changes (alternations) that are characterised as irregular in grammars. In this way all exceptions can be solved.

using rules capturing suffix and prefix derivations. A special derivational version of Ajka has been developed in the above-mentioned NLP Centre in Brno which enabled us to further explore the semantic nature of the selected noun derivational suffixes as well as verb prefixes and establish a set of semantic labels associated with the individual D-relations. With regard to the verbs we have focused on exploring the derivational relations between four selected prefixes and corresponding Czech verb stems or basic non-derived verbs just for one verbal semantic class, namely verbs of motion.

Inflectional paradigms are mentioned here because prior to the derivation of new words such as *novost* (*newness*) from *nový* (*new*) we first have to know to which inflectional paradigms they belong. In this sense there is a mapping between inflectional and derivational paradigms (at least in Czech). By using results obtained from the analyser Ajka and the mentioned D-interface we added the selected noun and verb D-relations to the Czech WordNet and in this way enriched it with approximately 31 000 new Czech synsets using the DebVisdic editor and browser (DebVisDic Manual, 2008) as illustrated in Fig. 1 towards the end of this article.

3.2 Characterisation of Czech D-morphology

Our starting data include 126 000 noun stems and 22 noun suffixes, 42 745 verb stems (or basic verbs) and fourteen verb prefixes. There are also alternations (infixes) in stems but we set them aside here. The complete inventory of the main noun suffixes is of course larger (approximately 120) and the same holds for the set of verb prefixes (approximately 240) – here we are paying attention only to the primary prefixes (fourteen in total, and of those we select only four due to space constraints). The higher number of prefixes in Czech follows from the fact that for each primary prefix there are about fifteen secondary (double) prefixes which consist of two primary ones, e.g. po-vy- in the verb *po-vy-skočit* (*to jump up a little*).

In Czech grammars (Karlík *et al*., 1995) we can find the following main types (presently fourteen) of the derivational processes exploiting suffixes and prefixes:

- Mutation: noun → noun derivation, e.g. *ryba -ryb-ník* (*fish → pond*). This semantic relation expresses the typical location of an entity.

- Transposition (the relation existing between different POS): noun → adjective derivation, e.g. *den → den-ní* (*day → daily*), semantically the relation expresses property.

- Agentive relation (cross-POS): verb → noun e.g. *myslit → myslitel* (*think → thinker*). This relation links an event and its agent. In the screenshot (Fig. 1 in section 7) a similar example for English is given where for the synset *dance*:1 there is a pair *dance → dancer* derived according the agentive D-rule. In this example, which is taken from the Princeton WordNet 2.0, the D-rule is not stated explicitly but the agentive relation is of course exploited here.

- Patient relation: verb → noun, e.g. *trestat → trestanec* (*punish → convict*). This relation expresses a relation between an action and the object (usually a person) impacted by it.

- Instrument (means) relation: verb → noun, e.g. *drž-et → drž-ák* (*hold → holder*). This relation links a tool (means) used to perform an action.

- Action relation (cross-POS): verb → noun, e.g. *uč-it → uč-e-n-í* (*teach → teaching*). Usually the derived nouns are characterised as deverbatives. Both members of the relation denote an event. An example of this relation is given in the screenshot (Fig. 1 in section 7) where for the synset *tancovat:2*, *tančit:2* the pairs *tancovat → tancování, tančit → tančení* (*dance → dancing*) can be found. They follow D-rule [deriv-dvrb].

- Property1 – verb-adj relation (cross-POS): verb → adjective, e.g. *vypracovat → vypracova-ný* (*work out → worked out*). Usually the derived adjectives are labelled as de-verbal, semantically it is a relation between an event and its resultant state.

- Property2 – adj-adv relation (cross-POS): adjective → adverb, e.g. *rychlý → rychl-e* (*quick → quickly*). Both adjective and adverb refer to the same property.

- Property3 – adj-noun relation (cross-POS): adjective → noun, e.g. *rychlý → rychl-ost* (*fast → speed*). The noun is the attribute and the adjective.

- Gender change relation: noun → noun, e.g. *inženýr → inženýr-ka* (*engineer → female engineer*). This relation links male and female referents.

- 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 endearing of belittling emotional attitude).
- Augmentative relation: noun → noun, e.g. *dub → dub-isko* (*oak tree → huge*, *strong oak tree*). The use of this suffix indicates means that the object denoted by augmented expression is really large or that the speaker use it pejoratively or expresses a positive emotional attitude.
- Possessive relation (existing between different POS): noun → adjective *otec → otcův* (*father → father's*). This is a relation between an object (person) and his/her possession.
- This D-relation exploits prefixes, in fact, it represents a whole complex of D-relations holding between verbs only, i. e.: verb → verb, e.g *nést → od-nést* (*carry → carry away*), *tancovat → dotancovat* (*dance → finish dancing*). We will say more about these relations below.

Table 1: Selected D-relations with suffixes implemented in Czech Wordnet

Label	Parts of speech	Meaning	No. of literals	Suffix
deriv-na	noun → adj	Property	641	*-í*,
deriv-pos	noun → adj	Possessive	4 037	*-ův*, *-in*
deriv-an	adj → noun	Property	1 930	*-ost*
deriv-aad	adj → adverb	Property	1 416	*-e*
deriv-dvrb	verb → noun	Action	5 041	*-í*, *-ní*
deriv-ag	verb → noun	Agentive	186	*-tel*, *-ík*, *-ák*, *-ec*
deriv-instr	verb → noun	Instrument	150	*-tko*, *-ík*
deriv-loc	verb → noun	Location	340	*-iště, -isko*
deriv-ger	verb → adj	Property	1 951	*-ící*, *-ající*, *-ející*
deriv-pas	verb → adj	Passive	9 801	*-en*, *-it*
deriv-g	noun → noun	Gender	2 695	*-ka*
deriv-dim	noun → noun	Diminutive	3 695	*-ek*, *-eček*, *-ička, -uška*
Total			31 429	

The 25 selected suffixes in Table 1 express a number of semantic relations, particularly action (deverbative nouns), property, possessive, agentive, instrument, location, gender variation and diminutive. Result and augmentative relation have not been included in Table 1 because the analyser Ajka does not currently handle them.

The abbreviated labels used in the Czech WordNet can be seen in Tables 1 and 2.

3.3 Prefixes

The core of the primary fourteen prefixes contains the following: *do-* (*to*), *na-* (*on*, *at*), *nad-* (*above*, *up*), *od-* (*from*, *away*), *pro-* (*for*, *because*), *při-* (*by*, *at*), *pře-* (*over*), *roz-* (*over*), *s-/se-* (*with*, *by*), *u-* (*at*, *near*), *v-/ve-* (*in*, *up*), *vy-* (*out*, *off*), *z-/ze-* (*of*, *off*), *za-* (*over*, *behind*).

Prefix D-relations in Czech hold only among verbs, typically between a stem or basic form and the respective prefix. It can be seen that the semantics of the prefix D-relations is different from the suffix D-relations because they hold between verbs that usually denote actions, processes, events and states.

Four of the fourteen Czech prefixes mentioned above have been selected as example in Table 2 to give an indication of the semantic nature of the D-relations and show the number of literals generated by the individual D-relations. They denote a number of semantic relations such as location, time, intensity of action, various kinds of motion (see Table 2), iterativity (repeated motion) and some others. It is obvious that they differ significantly from suffix based D-relations since they hold only between verbs. In the following table we will show how they combine with the selected verbs of motion.

Table 2: D-relations with 4 prefixes implemented in Czech WordNet

Label	Parts of speech	Meaning	No. of literals	Prefixes
prefix *do-* (*to*, *at*)				
deriv-act-t	verb → verb	finishing motion	173	*do-*
deriv-act-t-iter	verb → verb	finishing motion iterative	24	*do-*
prefix *od-* (*from*, *off*)				
deriv-mot-from	verb → verb	motion from	187	*od-*
deriv-mot-from-iter	verb → verb	motion from iterative	25	*od-*
deriv-oblig	verb → verb	obligation	2	*od-*
prefix *pře-* (*over*)				
deriv-mot-over	verb → verb	motion over a place	207	*pře-*
deriv-mot-over-iter	verb → verb	motion over a place iteratively	21	*pře-*
prefix *při-* (*to*, *at*)				
deriv-mot-to	verb → verb	motion to a place	171	*při-*
deriv-mot-to-iter	verb → verb	motion to a place iteratively	18	*při-*
deriv-add	verb → verb	additivity	3	*při-*
Total			743	

It should be remarked that the D-relation iterative is a subset of the verbs of motion, thus we do not count iterative verbs here as a new group. We also deal only with verbs of motion that have one argument, i.e. moving agent *jít* (*walk*/*go*). Verbs of motion with two arguments like *nést* (*carry*) are not included here though they represent quite a large number of the motion verbs. They are also

not pure motion verbs but cross over into contact and transfer (e.g. I bring you flowers).

3.4 Semantic classes of verbs and prefixes

The relation between semantic classes of verbs and verb prefixes should be mentioned here because in the Czech WordNet we adduce for each verb the semantic class it belongs to. The approaches to the semantic classes of verbs, particularly Levin's classification of English verbs (Levin, 1993) and its extension by Dang *et al*. (1998), are based on argument alternations whose nature is mostly syntactic, e.g. verbs that show a transitive-inchoative alternation (like *break*) not only share this particular syntactic behaviour but are semantically similar in that they denote changes of state or location.

Levin' s list of the most frequent verbs in English falls into over 50 classes (most of them with several subclasses); Palmer' s VerbNet project has extended this work to 395 classes. These verb classes have been translated and adapted for the Czech language. Presently, we work with approximately 82 semantic verb classes in the VerbaLex database of Czech valency frames containing approximately 12 000 verbs (Hlaváčková & Horák, 2006).

In this approach to the verb classification in Czech we exploit the verb valency frames that contain semantic roles. It appears that the verb classes established using semantic roles can be well compared with the classes obtained by the alternations. However, according to our results the classes obtained by means of the semantic roles appear to be semantically more consistent.

The third approach is based on the meanings of prefixes. If we have a look at how prefixes function in Czech we can see that they classify verbs yielding rather small and even more consistent semantic classes of verbs. Using prefixes as sorting criteria we obtain classes that are visibly closer to the real lexical data due to the fact that the prefixes are well established formal means. For example if we take prefix *do*- (it corresponds to the English preposition *to* or *at*) and apply it to the larger group of verbs of motion (approximately 1 200), the result is a group containing 173 Czech verbs denoting finishing motion and nothing more. The verb classes based on prefix criteria will be examined more thoroughly in future research.

Finally we would like to stress that not all the D-relations exploiting prefixes have been integrated into the Czech WordNet (see Fig.1

towards the end of this article) as yet. The relations based on prefixes will be integrated in the near future.

4. Derivational relations in English

Many traditional paper dictionaries include derivational word forms but list them as run-ons without any information on their meaning, relying on the user' s knowledge of morphological rules. Habash and Dorr (2003), recognising the importance of morphology-based lexical nests for NLP, created *CatVar*, a large-scale database of categorial variations of English lexemes. CatVar relates lexemes belonging to different syntactic categories (parts of speech) and sharing a stem, such as *hunger* (n.), *hunger* (v.) and *hungry* (adj.). CatVar is a valuable resource containing some 100 000 unique English word forms; however, no information is given on the words' meanings.

Miller and Fellbaum (2003) describe the addition of "morphosemantic links" to WordNet (Miller, 1995; Fellbaum, 1998), which connect words that are similar in meaning and where one word is derived from the other by means of a morphological affix. For example, the verb *direct* (defined in WordNet as "guide the actors in plays and films") is linked to the noun *director* (glossed as "someone who supervises the actors and directs the action in the production of a show"). Another link was created for the verb-noun pair *direct/-director*, meaning "be in charge of" and "someone who controls resources and expenditures", respectively. Most of these links connect words from different classes (noun-verb, noun-adjective, verb-adjective), though there are also noun-noun pairs like *gang-gangster*. English has many such affixes and associated meaning-change rules (Marchand, 1969).

When the morphosemantic links were added to WordNet, their semantic nature was not made explicit, as it was assumed – following conventional wisdom – that the meanings of the affixes are highly regular and that there is a one-to-one mapping between the affix forms and their meanings. But ambitious NLP tasks and automatic reasoning require explicit knowledge of the semantics of the links. Fellbaum *et al*. (2007) describe on-going efforts to label noun-verb pairs with semantic "roles" such as agent (*direct-director*) and result (*produce-product*). The assumption was that there was a one-to-one mapping between affixes and meanings. Fellbaum *et al*. (2007) extracted all noun-verb pairs with derivational links from WordNet and grouped them into classes based on the affix. They manually inspected each affix class expecting to find only a limited number of exceptions in each class. Instead, they found that the affixes in each

class were polysemous, that is, a given affix yields nouns that bear different semantic relations to their base verbs.

Table 3 shows Fellbaum *et al*.' s (2007) semantic classification of *-er* noun and verb pairs, with the number of pairs given in the right-hand column.

Table 3: Distribution of *-er* verb-noun pair relations in English

Agent	**2 584**
instrument	482
inanimate agent/Cause	302
event	224
result	97
undergoer	62
body part	49
purpose	57
vehicle	36
location	36

Examination of other morphological patterns showed that polysemy of affixes is widespread. Thus, nouns derived from verbs by *-ion* suffixation exhibit regular polysemy between event and result readings (*the construction took three years/the construction is shoddy and unsafe*; cf. Pustejovsky, 1995).

Fellbaum *et al*. (2007) also found one-to-many mappings for semantic patterns and affixes: a semantic category can be expressed by means of several distinct affixes, though there seems to be a default meaning associated with a given affix. Thus, many *-er* nouns denote Agents (*write-writer*, *drive-driver*, et cetera), and event nouns are regularly derived from verbs via *-alt* suffixation (r*everse-reversal, deny-denial,* et cetera). Patterns are partly predictable from the thematic structure of the verb. Thus, nouns derived from unergative verbs (intransitives whose subject is an agent) are agents, and the pattern is productive: *runner, dancer, singer, speaker, sleeper*, et cetera. Nouns derived from unaccusative verbs (intransitives whose subject is a patient/undergoer) are patients: *breaker* (wave), *streamer* (banner), et cetera. This pattern is far from productive:

*faller, ?arriver, ?leaver, et cetera. Many verbs have both transitive (causative) and intransitive readings (cf. Levin, 1993):

(2) a. The cook roasted the chicken

b. The chicken was roasting

For many such verbs, there are two corresponding readings of the derived nouns: both the *cook* in (2)a and the *chicken* in (2)b can be referred to as a *roaster.* Other examples of agent and patient nouns derived from the transitive and intransitive readings of verbs are *(best)seller, (fast) developer, broiler.* But the pattern is not productive, as nouns like *cracker, stopper*, and *freezer* show.

For virtually all *-er* pairs that we examined, the default agentive reading of the noun is always possible, though it is not always lexicalised. Thus a person who plants trees could well be referred to as a *planter*, but under this reading the noun seems infrequent enough not to deserve an entry in most lexicons. Speakers easily generate and process ad hoc nouns like *planter* (*gardener*), but only in its (non-default) location reading (*pot*) is the noun part of the lexicon, as its meaning cannot be guessed from its structure.

The semantic relations that were identified by Fellbaum *et al*. (2007) are doubtless somewhat subjective. Other classifiers might well come up with more coarse-grained or finer distinctions. Nevertheless, it is encouraging to see that this classification overlaps largely with that for Czech suffixes, which was arrived at independently. In addition, the English relations are a subset of those identified by Clark and Clark (1979), who examined the large number of English noun-verb pairs related by zero-affix morphology, i.e., homographic pairs of semantically related verbs and nouns (*roof, lunch, Xerox*, et cetera). This is the largest productive verb-noun class in English, and Clark and Clark's relations include not only agent, location, instrument and body part, but also meals, elements, and proper names.

In the context of the EuroWordNet project Vossen (1998) and Peters (1998) manually established noun-verb and adjective-verb pairs that were both morphologically and semantically related. Of the relations that Peters considered, the following match the ones we identified: agent, instrument, location, patient, cause. Peters' methodology differed from that of Fellbaum *et al*. (2007), who proceeded from the previously classified morphosemantic links and assumed a default semantic relation for pairs with a given affix. Peters selected pairs of

word forms that were both morphologically related and where at least one member had only a single sense in WordNet. These were then manually disambiguated and semantically classified, regardless of regular morphosemantic patterns.

5. Derivational relations in Zulu

Derivational morphology in Zulu constitutes a combination of morphemes, which may either produce a new word in a different word category or may leave the word category (class membership) unchanged. The first type of derivation results in a change in word class, and produces words which include nouns, verbs, adverbs and ideophones derived from other word categories. The derivation process of nouns from verbs (deverbatives) is the most productive, and is therefore singled out in this discussion.

When nouns are derived from verb roots, a noun prefix as well as a deverbative suffix is required, as illustrated in the following examples of nouns formed from the verb root *-fund-* learn:

(3)	a.	*u-m(u)-**fund**-i student*	in Czech the corresponding root is ***uč-***
	b.	*i-m-**fund**-o education*	in Czech the corresponding root is ***uč**-e-n-í*
	c.	*i-si-**fund**-o lesson*	there is no appropriate equivalent in Czech

The deverbative suffixes in the above example are *-i* and *-o.* Such nouns may have more than one suffix if the deverbative noun is derived from a verb root that has been extended, e.g.

(4)	*u-m(u)-**fund**-is-i teacher*	in Czech we have ***uč**-i-t-el (teach-er)*

The suffix *-is-* is a causative extension which changes the meaning of *-fund- learn* to *cause to learn* i.e. t*each*. Compare with English, where causatives are usually not morphologically derived, with very few exceptions like *rise-raise* and *fall-fell;* in most cases, causative and non-causatives are different morphemes: *kill-die, show-see*, et cetera. The last suffix *-i* is the deverbative suffix.

Table 4 shows the general rules for the formation of nouns from verb stems, however, not every verb can be treated in this way (cf. Doke, 1973:66):

Table 4: D-relations in Zulu

Personal deverbatives			
Prefix of personal class (i.e. noun class 1/2 or 7/8 or 9/10)	**Verb Root**	**Suffix -i**	
umu/aba (class 1/2) (personal class only) (most common)	fund (*learn*)	-i	umfundi *student*
	hamb (*go*, *walk*)	-i	umhambi *traveller*
	theng (*buy*)	-i	umthengi *customer*
	shumayel (*preach*)	-i	umshumayeli *preacher*
isi/izi (class 7/8) (personal as well as impersonal class)	eb (*steal*)	-i	isebi *thief*
	thul (*be silent*)	-i	isithuli *a mute*
	gijim (*run*)	-i	isigijimi *runner*, *messenger*
in/izin (class 9/10) (personal as well as impersonal class)	bong (*praise*)	-i	imbongi *royal praiser*
Impersonal deverbatives			
Prefix of impersonal class (i.e. noun class 3/4 or 5/6 or 7/8 or 9/10 or class 11)	**Verb root**	**Suffix -o**	
umu/imi (class 3/4) (impersonal class only)	buz (*ask*)	-o	umbuzo *question* (result)
i(li)/ama (class 5/6) (personal as well as impersonal class)	ceb (*devise, contrive*)	-o	icebo *plan, scheme* (result)
isi/izi (class 7/8) (personal as well as impersonal class)	aphul (*break*)	-o	isaphulo *rupture* (result)
in/izin (class 9/10) (personal as well as impersonal class)	phuc (*shave*)	-o	impuco *razor* (instrument)
u(lu) (class 11) (impersonal class only)	thand (*love*)	-o	uthando *love* (abstract)

Personal deverbative nouns signify the agent of the action expressed by the relevant verb while impersonal deverbatives indicate the following semantic relations:

- instrument of the action signified by the verb;

- result of an action is conveyed; and/or
- abstract idea conveyed by the verb.

As indicated in column 1 of Table 4 under impersonal deverbatives, there is an overlap in the semantic content of classes (i.e. personal and impersonal), which makes the choice of the correct class prefix rather unpredictable. However, certain regularities can be identified in the sense that personal deverbatives usually suffix *-i* to the verb root while impersonal deverbatives suffix *-o*.

Exceptions to the general rule also occur, e.g. the impersonal noun *umsebenzi* (um-sebenz-i) *work* is derived from the verb root-*sebenz*- (*work*), but uses the "personal" suffix *-i.*

The second type of derivation creates derived forms within the same word class, and produces words like diminutives, feminine gender, augmentatives and locatives, as illustrated in the following table:

Table 5: Same word class derivations in Zulu

Noun	Prefix	Suffix	Derived form
isitsha (*dish*)		-ana (diminutive)	isitshana (*small dish*)
intaba (*mountain*)		-kazi (augmentative)	intabakazi (*big mountain*)
imvu (*sheep*)		-kazi (feminine gender)	imvukazi (*ewe*)
ikhaya (*home*)	e- (locative)		ekhaya (*at home*)
indlu (*house*)	e- (locative)	-ini (locative)	endlini (*in the house*)

As can be gleaned from Table 5, semantic categories such as diminutives, feminine gender and augmentatives are regularly and productively derived in Zulu by means of suffixation, while in the case of locatives, prefixation takes place often in combination with the suffix *-ini.*

It is noteworthy that although locativised nouns such as *ekhaya* (*at home*) may also be used to function as adverbs, they continue to exhibit certain characteristics of regular nouns, for instance functioning as subjects and objects and in the process triggering agreement.

The Zulu wordnet, being part of the African Languages Wordnet, is still in a conceptualisation phase although experimental work on noun and verb synsets has begun (cf. Bosch *et al*., 2007; Moropa *et al*., 2007). The African Languages Wordnet effort, aiming to create

an infrastructure for wordnet development for African languages, began with a week-long workshop funded by the Meraka Institute (CSIR) in Pretoria in March 2007. Christiane Fellbaum (Princeton), Piek Vossen (Amsterdam) and Karel Pala (Brno) facilitated. Linguists and computer scientists representing nine official South African languages were introduced to WordNet lexicography and familiarised with the lexicographic editing tools DebVisDic (DebVisDic Manual, 2008).

6. Semantics of the D-relations in English, Czech and Zulu

The semantic classifications done for Czech and English suggest that the meanings of affixes can be classified into a finite, relatively small number of semantic categories. It is important to note that the inventory of relations is somewhat arbitrary; one could certainly propose a more fine-grained or a more coarse-grained one. We expect to encode additional relations as we consider other types of morphosemantic pairs for the languages discussed here and languages not yet thus analysed. Nevertheless, we anticipate a fairly small number of relations, most likely a subset of those discussed by Clark and Clark (1979) and Fillmore' s (1968) Cases.[2]

6.1 Similarities and differences for English, Czech and Zulu

Now we can attempt to compare the D-relations in three languages. It may be surprising to find that Czech and a Bantu language such as Zulu are in a certain respect formally closer than Czech and English. This is due to a rich system of affixes in both languages though they are not exploited in the same way in Czech and Zulu. Similarity consists in highly developed prefixation and suffixation but in Zulu both are used in a way typical for agglutinative languages, this certainly holds for noun prefixes. In Czech, prefixation is typical mostly for verbs and deverbatives which are, in fact, verbs as well. English also has verbal prefixes (e.g. *out-* prefixes to intransitive

2 The labels we assigned refer to well known semantic categories and have been studied or applied in different contexts. The Cases proposed by Fillmore (1968) and the FrameElements of FrameNet (Ruppenhofer *et al*., 2002) also refer to agents, undergoers/patients, instruments, and so forth. The work of Tony Veale and his group (Veale & Hoa, 2007; Butnariu & Veale, 2007) shows how WordNet can be enriched with new semantic information by searching corpora for lexical patterns. The semi-automated methods of Veale *et al*. are clearly to be preferred over painstaking and intuitive manual procedures, and the present article similarly hopes to highlight their advantages.

verbs and makes them transitive: *I outran the bear*) but makes regular use of separate particles to form phrasal verbs *(look up/down/ away,* et cetera).

What all three languages have in common is the small number of semantic relations expressed by morphemes that create new words. The analyses of Czech, English and Zulu presented here allow us to predict that these D-relations are likely to be universal. All three languages use morphological processes to regularly and productively derive semantic categories such as agent, instrument, location, gender, diminution, augmentation, result as well as others.

The following is an example to illustrate the similarities and differences of D-relations for Czech and English. Although the Czech D-relations are much more regular, the English counterparts are given in brackets: In Czech (English) there are thousands of adjectives like *nový* (*new*), *chytrý* (*clever*) to which a suffix *-ost* (*-ness*) can be added. In this way we obtain *nový – novost* (*new – newness*) or *rychlý – rychlost* (*fast – speed*, here English is different). Thus a D-rule can be formulated that captures this relation and allows us to form the noun *novost* from the adjective *nový* (frequency: 143 303 occurences in the Czech National Corpus (CNK)). This D-rule can also be used for recognising all the pairs *nový – novost* in arbitrary Czech texts. Semantically, this rule can be labelled as adj-noun property since both the adjective and noun derived from it denote property.

A similar D-rule can be formulated for approximately a thousand Czech verbs if the suffix *-tel* is used, for example if the suffix *-tel* (*-er*) is applied to the verb *učit* (frequency 8 639 in CNK) we get the noun *učitel* (*teacher*) (frequency 9 924 in CNK).

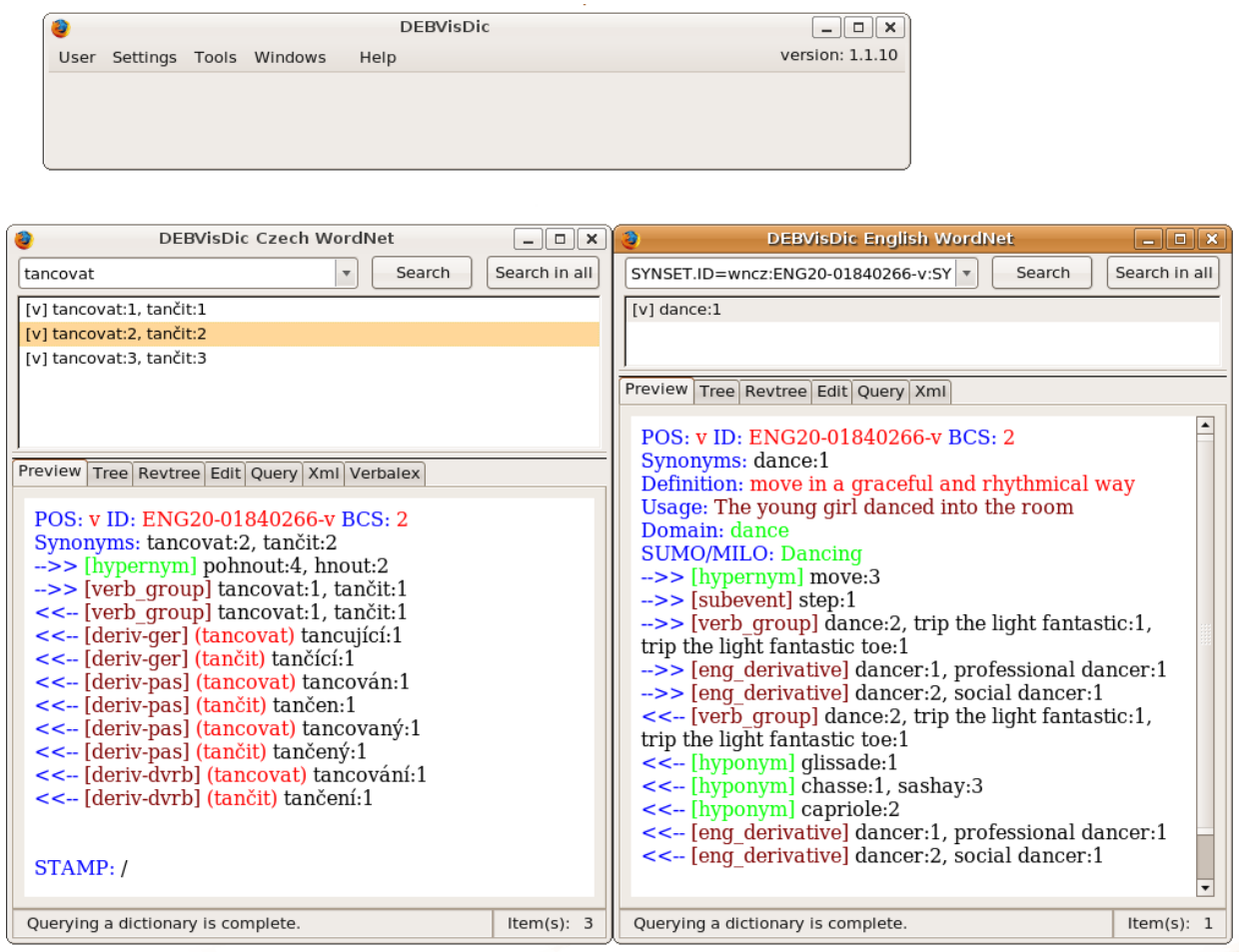
It was mentioned earlier that D-rules could cause either over- or undergeneration of derived forms. As for the D-rules of Zulu, a semi-automated process of dealing with these challenges within a morphological analyser is described in Bosch and Pretorius (2006). This process involves the extraction of derived forms from corpora by means of morphological analysis which exploits the D-rules of the language. However, when it comes to assigning of meaning to the output of the D-rules, ways now need to be explored in which Zulu can benefit from existing work such as the derivational version of Ajka for Czech. This will assist to further explore the semantic nature of noun derivational affixes. This aspect will form part of future work.

7. Representing D-relations in WordNet and relations between literals (screenshots of Czech and Princeton WordNets)

The screenshot below (Figure 1) indicates how D-relations are visualised (represented) in the Czech and English WordNet using the browser and editor DebVisdic (DebVisDic Manual, 2008). The DebVisDic tool uses the XML database in which synsets are stored in highly configurable XML format. D-relations are represented in the XML format as well and they link together the individual literals of which the synsets consist. In the text form this can be seen in the screenshot. It should be added that D-relations are stored in the same way as other semantic relations used in wordnets, i.e. relations like synonymy, antonymy, hypero/hyponymy, meronymy, and so forth. The exploited XML format, in fact, makes it possible to introduce any kind of semantic relation into the WordNet database.

The example shows the verb *tancovat*:1*/tančit* :1 -*dance*:1 in Czech WordNet and PWN 2.0. It would be better to show the verb *dance* in PWN 3.0 (Wordnet a lexical database for the English language, 2006) where the respective D-relations are more complete but it has not been converted yet for browsing in DebVisdic. Nevertheless, it should be clear from the screenshot which relations are captured and how they are visualised.

Fig. 1: D-relations in Czech and English WordNet



To give an idea how the equivalent or near equivalent look like for all three languages we offer the following example (in text form only but in DebVisDic it will look very similar):

```
Eng: POS: v ID: ENG20-01840266-v BCS: 2
     Synonyms: dance:1
     Definition: move in a graceful and rhythmical way
     Deverbative: 1. dancer

Cz: POS: v ID: ENG20-01840266-v BCS: 2
     Synonyms: tancovat:2, tančit:2
     Deverbative: 1. tanečník

Zulu: POS: v ID: ENG20-01840266-v BCS: 2
     Synonyms: dansa:1
     Definition: dance in a European fashion
     Deverbative: 1. umdanso (pl. imidanso) – noun class 3/4
     Definition: European dance
      Deverbative: 2. umdansi (pl. abadansi) – noun class 1/2

     Definition: dancer (in European fashion)
```

8. Conclusions

In the article we present the first results of the analysis of basic and most regular D-relations in English, Czech and Zulu. We offer basic lists of the D-relations for all three languages. Though the analysis is currently still far from complete we have decided to include the items related with D-relations in the Czech and English Wordnet, respectively. This makes it possible to enrich both WordNets considerably with the derivational nests (subnets). In our view, this kind of enrichment makes both WordNets more suitable for applications such as searching. For Zulu we try to show how the Czech and English experience could be applied in building wordnets for Bantu languages in general.

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

We hope that the work reported here will stimulate similar work in other languages and allow insights into their morphological processes as well as facilitate the computational representation and treatment of crosslinguistic morphological processes and relations.

List of references

BERKO GLEASON, J. 1958. The child's learning of English morphology. *Word*, 14:150-77.

BOSCH, S., FELLBAUM, C., PALA, K. & VOSSEN, P. 2007. African Languages WordNet: laying the foundations. Presented at the 12th International Conference of the African Association for Lexicography (AFRILEX), Soshanguve.

BOSCH, S. & PRETORIUS, L. 2006. A finite-state approach to linguistic constraints in Zulu morphological analysis. *Studia orientalia*,103:205-227.

BUTNARIU, C. & VEALE, T. 2007. A hybrid model for detecting semantic relations between noun pairs in text. (*In* Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), ACL, Prague. p. 378-381.)

CLARK, E. & CLARK, H. 1979. When nouns surface as verbs. *Language*, 55:767-811.

DANG, H.T., KIPPER, K., PALMER, M. & ROSENZWEIG, J. 1998. Investigating regular sense extensions based on intersective Levin classes. (*In* Coling/ACL-98, 36th Association of Computational Linguistics Conference. Montreal 1, p. 293-300.)

DEBVISDIC MANUAL. 2008. http://nlp.fi.muni.cz/trac/deb2/wiki/DebVisDic Manual Date of access: 22 Feb. 2008.

DOKE, C.M. 1973. Textbook of Zulu grammar. Johannesburg: Longman.

FELLBAUM, C. 1998. WordNet: an electronic lexical database. Cambridge: MIT.

FELLBAUM, C., OSHERSON, A. & CLARK, P.E. 2007. Adding semantics to WordNet's "morphosemantic" links. (*In* Proceedings of the Third Language and Technology Conference, Poznan. p. 226-230.)

FILLMORE, C. 1968. The case for case. (*In* Bach, E. & Harms, R., *eds*. Universals in linguistic theory. New York: Holt. p. 1-88.)

HABASH, N. & DORR, B. 2003. A categorial variation database for English. (*In* Proceedings of the North American Association for Computational Linguistics, Edmonton. p. 96-102.)

HLAVÁČKOVÁ, D. & HORÁK, A. 2006. VerbaLex – new comprehensive lexicon of verb valencies for Czech. (*In* Computer treatment of Slavic and East European languages. Bratislava: Slovenský národný korpus. p. 107-115.) (Slovak National Corpus.)

KARLÍK, P., NEKULA, M. & RUSÍNOVÁ, Z. 1995. Příruční mluvnice češtiny. (Reference Czech Grammar.) Prague: Nakladatelství Lidové Noviny.

KOSCH, I.M. 2006. Topics in morphology in the African language context. Pretoria: Unisa.

LEVIN, B. 1993. English verb classes and alternations. Chicago: University of Chicago.

MARCHAND, H. 1969. The categories and types of present-day English word formation. Munich: Beck.

MILLER, G.A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39-41.

MILLER, G.A. & FELLBAUM, C. 2003. Morphosemantic links in WordNet. *Traitement automatique de langue*, 44(2):69-80.

MOROPA, K., BOSCH, S. & FELLBAUM, C. 2007. Introducing the African languages WordNet. Presented at the 14th International Conference of the African Language Association of Southern Africa, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

PETERS, W. 1998. The English wordnet, EWN Deliverable D032D033. Sheffield: University of Sheffield.

PUSTEJOVSKY, J. 1995. The generative lexicon. Cambridge: MIT.

RUPPENHOFER, B.C. & FILLMORE, C.J. 2002. The FrameNet database and software tools. (*In* Braasch, A. & Povlsen, C., *eds*. Proceedings of the 10th Euralex International Congress. Vol. 1. Copenhagen. p. 371-375.)

SEDLÁČEK, R. & SMRŽ, P. 2001. A new Czech morphological analyser Ajka. (*In* Proceedings of the 4th International Conference on Text, Speech and Dialogue. Berlin: Springer Verlag. p. 100-107.)

VEALE, T. & HAO, Y. 2007. Making lexical ontologies functional and context-sensitive. (*In* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague. p. 57-64.)

VOSSEN, P., *ed.* 1998. EuroWordNet. Dordrecht: Kluwer.

WORDNET. 2006. A lexical database for the English language. http://wordnet.princeton.edu/ Date of access: 10 Sept. 2007.

Key concepts:

derivational relations
lexical resources
semantic relations
wordnets

Kernbegrippe:

afgeleide verhoudings
leksikale bronne
semantiese verhoudings
woordnette