Lexical Resources for Noun Compounds in Czech, English and Zulu

Karel Pala[1], Christiane Fellbaum [2], Sonja Bosch[3]
[1]Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2] Department of Computer Science, Princeton University, USA
[3] Department of African Languages, University of South Africa, Pretoria, South Africa
pala@fi.muni.cz, fellbaum@princeton.edu, boschse@unisa.ac.za

Abstract

In this paper we discuss noun compounding, a highly generative, productive process, in three distinct languages: Czech, English and Zulu. Derivational morphology presents a large grey area between regular, compositional and idiosyncratic, non-compositional word forms. The structural properties of compounds in each of the languages are reviewed and contrasted. Whereas English compounds are head-final and thus left-branching, Czech and Zulu compounds usually consist of a leftmost governing head and a rightmost dependent element. Semantic properties of compounds are discussed with special reference to semantic relations between compound members which cross-linguistically show universal patterns, but idiosyncratic, language specific compounds are also identified. The integration of compounds into lexical resources, and WordNets in particular, remains a challenge that needs to be considered in terms of the compounds' syntactic idiosyncrasy and semantic compositionality. Experiments with processing compounds in Czech, English and Zulu are reported and partly evaluated. The obtained partial lists of the Czech, English and Zulu compounds are also described.

1. Introduction

Word formation is a productive process that generates new forms and meanings, making the lexicon huge and open-ended and confronting builders of lexical resources with the decision as to what to include and what to omit.

We discuss noun compounding, a particular kind of word formation, in three distinct languages: Czech, English and Zulu, building on earlier work on the analysis and representation of morphologically related word forms (Bosch, Fellbaum & Pala, 2008) and broadening the English-Zulu-based perspective (Bosch & Fellbaum, 2009) with Czech data.

From a semantic perspective, compounds, like all lexemes arising from derivational morphology, represent a large grey area between regular, compositional word forms on the one hand and idiosyncratic, non-compositional ones on the other hand. In Czech, however, noun compound formation is exceptionally regular; most compounds in Czech are compositional but some noncompositional compounds can be found too. Some of the compounds may not be understood as compounds by an average user and certain etymological knowledge may be necessary to recognize their members, e. g. *morfo-logie* (morphology) is a case in point, though its Czech equivalent *tvaro-sloví* is a bit more transparent.

The members of some compounds may be etymologically obscure to most contemporary speakers (e.g., *-diction* in *jurisdiction*) but in many cases, their meanings are sufficiently clear to make them productive nevertheless; this is the case for *–logy* and *-itis*, which give rise to new and immediately understandable words like *biotechnology* and *telephonitis*.

The semantics of compounds can be added to the word derivation relations found already in Czech and English wordnets (Hlaváčková & Pala, 2007; Fellbaum 1998). For instance, if we have a literal 'computer:3' in the Princeton WordNet (PWN) v. 2., we can connect it with the literal 'microcomputer' via compound relation.

From a formal perspective, English compounds are for the most part noun phrases consisting of two or more nouns or adjectives and nouns. We consider such compounds but also others, which we call morphological compositions. Composites or compositions denote single word compounds in comparison with multiword compounds. Czech examples of composites are: *moře-plavba* (sea-sailing) is a composite, *vysoká škola* (university, literally high school).

In Czech almost all nouns are inflected and there are also compositions where the first member is a verb which is inflected as well. Moreover, lexicalized compounds (compositions) in Czech are single words only, in which only the second member is inflected in accordance with the synthetic nature of the language. Expressions consisting of two (or more) words (as opposed to one-word compositions) are considered collocations (or MWEs) and not compounds in Czech.

In Zulu most compounds are also single words that are formed on a morphological level, and are not merely two words that have been juxtaposed.

Formally, we can say that one word is a string of letters without space. We also consider words connected by a hyphen as single words.

2. Formal properties of Czech, English and Zulu compounds

We briefly review and contrast the structural properties of Czech, English and Zulu nominal compounds.

2.1 Czech

In Czech derivational morphology (Petr et al., 1986), compounds or more precisely composites, are defined

as single words consisting of bases, i.e. morphemes that are stems or roots, and connecting infixes (connectors). Thus they are single lexicalized items that can be found in Czech dictionaries as headwords. In this respect Czech is different from English where compounds are typically formed by multi-word expressions (MWEs), words that are arguably lexical units but that are separated by spaces.
The main patterns of Czech composites are pairs which mostly consist of the governing (head) and dependent member:

N-N:
conjunctive (coordination), reciprocal, going in both directions, typically N-N, e. g. *jazyk-o-věda* (literally language-science, linguistics) *les-o-step* (forest stepp, prairie), les-o-park (forest park), *lid-o-op* (anthropoid ape), *čas-o-prostor* (space-time), *moř-e-plavba* (sea voyage, sailing). The *–o-* and *–e-* here are semantically empty connectors that are sometimes required to bind the two compound members together, a phenomenon found also in German.

A-N compositions (modifications):
in which A modifies N, e. g. *star-o-usedlík* (old resident), *čern-o-býl* (mugwort). Frequent are compounds with the adjectives *malý* (*mal-o-myslnost,* faint-heartedness) / *velký* (*velk-o-myslnost,* high-mindedness)*, nový/starý, rychlý* (new/old, fast). Quite frequent are also hybrid compounds with Greek or Latin bases like *makro/mikro, mono-, elektro-, auto-*. Similar compounds appear in English as well, e.g. *micro-computer*, *macroeconomics*, *auto-suggestion*, but also compounds like *teo-logie* (theo-logy), *ideo-logie* (ideo-logy).

Num-N:
here the first member of the compound is a numeral, e. g. *prv-o-číslo* (prime number), *dvou-hra* (double – in tennis), *troj-hlas* (composition for three voices), *čtyř-takt* (four-stroke engine).

Pron-V:
the first member of the compound is a pronoun, the second is a verb, e. g. *vše-věd* (literally all know, wise man), *samo-pal* (literally: self shoot, sub-machine gun), observe that there is more than 150 Czech compounds have the first member with *samo* (auto, himself, itself) as the first member.

V-N:
the first member of the compound is a verbal base, the second is Noun, e. g. *svíti-plyn* (coal gas), *prši-plášť* (raincoat, macintosh).

N-V:
here a verbal base appears as a second member of the compound, e. g. *pivo-var* (literally: beer brew, brewery), *vodo-vod* (literally: water lead, water supply).

Adv-V:
the first member of the compound is an adverb, the second is a verb, e. g. *těsno-pis* (literally: tight write*,* stenography), *darmo-šlap* (literally: uselessly walk, do-nothing).

Adjective compounds:
A – A, e. g. *trestně-právní* (of criminal law)
Adv – A, e. g. *tmavo-modrý* (dark-blue)
Pron – A, e. g. *samo-volný* (unprompted, spontaneous), *vše-objímající* (all-embracing)
Noun – A, e. g. *sněh-o-bílý* (snow-white), *vod-o-těsný* (water-resistant, leakproof)
Adv – Adv, e. g. *dennodenně* (literally: day by day, every single day)
These compositions, e.g. of the A-N kind - (*velk-o-myslnost,* high-mindedness) reflect the structure of a similar (corresponding) NP consisting of the constituents A and N as they occur in a sentence, such as in NP *velká mysl* (high mind).

2.2 English

Compounds in English can be formally described as follows. Noun compounds are head-final and thus left-branching. English has very few compounds like *attorney general*, where the phrasal head is not the rightmost member. In most compounds, the phrasal head is also the semantic head, i.e., the constituent that expresses the basic meaning of the compound (in WordNet terms, the superordinate, more general, concept). For example, the semantic heads of *windowsill* and *attorney general* are *sill* and *attorney*, respectively. Some noun compounds have a head that is not a noun; nevertheless, the category of the phrase is nominal. Examples include the large class of English nouns derived from phrasal verbs: *cutoff, pick-up, set-up, look-up, push-up, knock-out, drop-in, stowaway, shut-down,* etc. (Note that not all phrasal verbs can be nouns: *cut up, set off, look on, push away, knock back, drop down* and *shut away* have uses as verbs only). Noun compounds composed of verbs include *wannabe, must-have* and *knock-me-down*. In each case, the semantic head is unexpressed and cannot be easily inferred, as there is no obvious or regular semantic relation between the noun and the implied head. Because of their semantic idiosyncracy, such compounds must be listed in the lexicon.
Under the classical view of the Lexicon, only irregular words should be included. However, there is no clear-cut division between regular and irregular lexemes. Moreover, NLP clearly benefits when multi-word phrases can be looked up and interpreted wholesale rather than be submitted to the parser first. We loosely agree with Agirre, Aldezabal and Pocielli (2006) who propose that any MWE that is included in a standard lexical resource should be considered as "lexicalized" and thus worthy of inclusion in a lexical database like

WordNet (though the authors concede that dictionaries do not agree with respect to their entries).

2.3 Zulu

In Zulu, a lexeme containing at least two stems or roots is considered to be a compound. But unlike in English, the process of compounding in Zulu is not merely a concatenation of independent words. Words contributing to a compound may undergo phonological and morphological changes. Parts of these words "may be elided, replaced or adapted in some way or another, so that the lexical components do not appear as full words, but as bound stems and roots." (Kosch, 2006:122).
Zulu has a rich system of prefixation, and compounding affects prefixation. A new prefix is added for every compound headed by a verb as in *-lala* (sleep) + *indle* (wilderness) > *umlalandle* (wild animal) to which the class 3 (*umu-*) prefix has been added. The class prefix of a noun as head may even get changed in a compound, as in *indaba* (information) + *emlonyeni* (in the mouth) > *undabamlonyeni* (news of the day) where the prefix for class 9 (*in-*) becomes class 1a (*u-*) in the compound.
The main patterns of Zulu compounding are similar to Czech with pairs which usually consist of a leftmost governing (head) and a rightmost dependent element:

N-N
intabamlilo (vulcano)
intaba (mountain) + *umlilo* (fire)
The initial vowel of the noun *umlilo* (fire) is discarded in the compound.

N-Pron
udadewethu (my/our sister)
udade (sister) + *wethu* (my/our)
There are no morphological changes in this compound.

N-Qual
usibalukhulu (person in authority)
usiba (feather*)* + *olukhulu* (big)
The initial vowel of the qualificative *olukhulu* (big) is discarded in the compound.

N-Adv
umushonje (simple sentence)
umusho (sentence) + *nje* (in this way/manner)
There are no morphological changes in this compound.

N-Ideo[1]
uthambophoqo (fracture)
ithambo (bone) + *phoqo* (of snapping/breaking)
The ideophone consists only of a root which simultaneously functions as a stem and a fully-fledged word. This is in contrast to the linguistic word in Zulu, which is characterised by a number of morphemes such as prefixes and suffixes, as well as a root or stem. The ideophone *phoqo* (of snapping/breaking) is simply attached to the noun in this compound.

N-V
uyihlozala (your father-in-law)
uyihlo (your father) + *-zala* (give birth)
The verb stem *-zala* (give birth) is used without any concordial prefixes in the compound.

uxamukavinjelwa (strong-headed person)
uxamu (monitor lizard) + *akavinjelwa* (it is not stopped)
The initial vowel of the verb *akavinjelwa* (it is not stopped) is discarded in the compound. Instead of merely a verb stem, the rightmost element of the compound consists of a complete verb with negative concordial prefix as well as passive suffix.

V-N
umlindasimu (scarecrow)
-linda (guard) + *insimu* (field/cultivated land)
The initial vowel of the noun *insimu* (field/cultivated land) is discarded in the compound. The compound is given a class 3 noun prefix *um(u)-*.

V-Pron
umhlalawodwa (hermit)
-hlala (sit/stay/live) + *wodwa* (it alone/only it)
The compound is given a class 3 noun prefix *um(u)-*.

V-Qual
umbonanhle (good omen, literally: pleasant sight)
-bona (see) + *enhle* (which is pleasant)
The compound is given a class 3 noun prefix *um(u)-*.

V-Adv
isibonakude (binoculars, telescope)
-bona (see) + *kude* (far)
The compound is given a class 7 noun prefix *isi-*.

umhlalaphansi (pension)
-hlala (sit/stay/live) + *phansi* (down/beneath)
The compound is given a class 3 noun prefix *um(u)-*.

V-V
umaganedlula (woman without morals)
-gana (marry) + *-edlula* (pass)
The compound is given a class 1a noun prefix *u-,* as well as a feminine prefix *–ma-*. The terminative vowel of the verb *–a* is discarded.

[1] A feature particular to the Bantu languages (to which Zulu belongs) is the POS known as “ideophone”, a word category which describes a predicate, qualificative or adverb in respect to manner, colour, sound, smell, action, state or intensity.

V-Ideo
ibuyambo (recurrence of an action/rebounding, boomerang action)
-buya (return) + *mbo* (of striking)
The ideophone *mbo* (of striking) is simply attached to the verb stem in this compound, which is given a class 5 noun prefix *i(li)-*.

According to Ungerer (1983:249) the majority of noun compounds in Zulu consist of either a verb stem (63,9%) or a nominal element (30,2%) as head.

3. Semantic properties of compounds

Compounding is a highly generative, productive process; speakers make up compounds on the fly and hearers decode them effortlessly. This means that the majority of compounds are compositional, though their meaning may depend on the context in which they are embedded (e.g., *banana war* is best interpreted in the context of trade relations among nations that are exporting and importing the fruit). For example, the following compounds (with A-N structure) headed by *chair* specify the LOCATION of the *chair* (where it is used): *office chair, kitchen chair.* Other compounds specify salient or characteristic PARTS of the head: *armchair, wing chair, wheelchair.* In English the head can be modified by a verbal participle: *stacking chairs* and *folding chairs* CAN BE stacked and folded, respectively.
Compounds are found particularly in terminology-rich domains, such as specific professions, nationalities, animals, plants, chemicals, names of physical as well as frozen expressions. It would be very useful to have compounds marked with this essentially ontological information. The book dictionaries usually give labels like (bio.) or (eng.). English WordNet (WordNet…, 2010) provides many synsets with domain labels, such as *mathematics* and *cooking*; these domain labels are linked to the appropriate synsets. For cases like professions, animals, plants, etc. a superordinate provides the 'domain'. Moreover, for Balkanet the domains developed by Bentivogli et al. (2004) have been introduced. Linking to SUMO/MILO (Niles & Pease, 2003) also yields domain-like information.

3.1 Semantic relations between compound members

In Czech, English and Zulu N-V and V-N compounds, the noun can encode semantic roles associated with the event denoted by the verbs, such as Theme/Patient, Location, Instrument and Manner. A cross-linguistic perspective shows the universality of the patterns and, conversely, identifies idiosyncratic, language-specific compounds. The PWN explicitly encodes the semantic relation between a verb and derived noun (Fellbaum, Osherson & Clark, 2007). For example, a specific sense of the verb *direct* is linked to the appropriate noun *director*, and the link is labelled AGENT.

3.2 Metaphoric compounds

Some semantically idiosyncratic (or idiomatic) compounds have metaphorical character. *Ladyslipper* and *buttercup* are flowers whose folk names derive from their similarity to the objects referred to under the literal readings. Similarly, *skyscrapers* (Czech equivalent is figurative in a similar way, *mrak-o-drap,* literally: cloud scraper) are not *scrapers* under any readings of this noun. Corresponding examples in Zulu are *ihelanjadu* (scandalmonger), composed of *-hela* (sniff up snuff) plus *injadu* (snuff-box) and *umkhabimalanga* (fellow or mate), which combines *inkabi* (ox) and *amalanga* (days). Such idiomatic compounds often denote common concepts that are therefore lexicalised and feature in lexical databases.

4. Compounds in NLP

How can compounds be recognized automatically if they are not found in the lexical database (or if a system determines that the lexical entry is not the context-appropriate one)? In such cases, the semantic relation between the compound members should be determined. For example, the relation may be expressible with phrases like "X is performed on Y" (*baking potato*) or "X is the location for Y" (*baking dish*). Work by Kim and Baldwin (2008) and Kim, Mistica and Baldwin (2007) has recast the interpretation of noun compounds as a Semantic Role labelling task; the Semantic Roles express the relations among the members of different compounds as examined by, e.g., Levi (1978). Kim, Mistica and Baldwin (op cit) took advantage of the WordNet hypernym structure to identify concepts like PRODUCTS and PRODUCER, which form a semantic subclass of noun compounds such as *honey bee* and *music clock*. The semantics among verbs and derived nouns in WordNet are similarly expressed in a way that captures Semantic Relations (Fellbaum, Osherson & Clark, 2007).

4.1 Resources

We propose the construction of the following resources.
a) Czech: the first resources of the compounds are two Czech representative dictionaries: the smaller one is the Dictionary of Written Czech (SSC, 1986), which contains 4, 75% compounds, in the second one – the Academic Dictionary of Literary Czech (SSJC, 1960) we find 8, 95% the marked compounds. The second resource is a list of stems containing approx. 400 000 items (Sedláček, 2004). It is integrated in the tool Deriv, which allows us to recognize frequent types of compounds using rules with regular expressions (Hlaváčková et al., 2009). We performed a simple experiment in which we used the Deriv tool to search for compounds with some selected bases, for example the base *-pis-* (write) gives noun *leto-pis* (literally *year-write,* i.e. chronicle) and further 49 ones plus 50

adjective compounds with the same base. The base *-logie* (-logy) yields 282 nouns and 280 adjective compounds such as *morfo-logický* (morpho-logical). Similarly, the base *auto-* gives 186, e. g. *auto-baterie* (car battery). More examples of this sort can be offered.

We have tried to estimate the percentage of the compounds produced with the tool Deriv If we combine it with the numbers obtained from the Czech dictionaries we come to approx. 20%, i.e. at the moment we can say that the compounds cover not less than 20% of the Czech word stock. What remains to be done is to automatize the searching of the compounds as completely as possible.

b) English: there is no easy way to determine the number of compounds in English, especially not given the broad definition we adopt in this paper. We are not aware of any dictionary that includes all these types of compounds and composites. An experiment involving extracting the most frequent English compounds from BNC using Word Sketch Engine (Kilgarriff et al., 2004) will therefore need to be performed.

c) Zulu: the most comprehensive available resource is a list of 1227 lexicalised noun compounds compiled by Ungerer (1983:26-347). The list is compiled mainly from dictionaries and orthography lists. Zulu being a lesser resourced language, an experiment with a small electronic list of noun and verb stems as well as compound formation rules (some of which are discussed in 2.3) with regular expressions, was implemented in a Perl program. Possible prefixes were specified according to the typical noun and verb prefix patterns of the language, viz. optional V is followed by any number of CV or CCV and optionally followed by one or more C. The results produced by the Perl script are given in Table 1.

```
>> $ perl zulu.pl nstems vstems < corpus
>> ihlalankomo can be a case of 1.1:
i-hlala-0-nkomo
>> ihlalankomo can be a case of 1.2:
i-hlal-a-nkomo
>> ugoningane can be a case of 1.2:
u-gon-i-ngane
>> umhlalabantu can be a case of 1.3:
um-hlala-b-a-ntu
>> idlalesula can be a case of 2.4:
i-dlale-0-sula
>> ihambahlale can be a case of 2.4:
i-hamba-0-hlale
>> umkhathizwe can be a case of 3.5:
um-khathi-0-zwe
>> umphathisihlalo can be a case of 3.6:
um-phathi-s-i-hlalo
>> nehlalankomo can be a case of 1.1:
ne-hlala-0-nkomo
>> nehlalankomo can be a case of 1.2:
ne-hlal-a-nkomo
>> nogoningane can be a case of 1.2:
no-gon-i-ngane
>> njengomhlalabantu can be a case of
1.3:
njengom-hlala-b-a-ntu
>> onehambahlale can be a case of 2.4:
one-hamba-0-hlale
>> ungumkhathizwe can be a case of 3.5:
ungum-khathi-0-zwe
>> omphathisihlalo can be a case of 3.6:
om-phathi-s-i-hlalo
>>
```

Table 1: Zulu examples produced by the Perl script

The evaluation of the results of this small experiment is as follows: in each example the verb root/stem or noun stem is correctly identified, e.g. *–hlala, -hamba, -khathi, -ntu* etc. Provision is made for a possible connecting vowel between the two elements of the compound, which however, results in overgeneration in an example such as *–hlal-a,*. In a more extended experiment involving a list of several thousands of noun and verb stems, it became clear that the compound formation rules need to be refined somewhat for an effective (semi-) automated compound extracting tool. Since the compounding process is extremely productive and compounds are developed on the fly (cf. Bosch & Fellbaum, 2009:39), such a tool will be most useful for extracting 'new' compounds from real life corpora such as newspaper resources.

4.2 Compounds in wordnets

A compound that is judged syntactically or semantically idiosyncratic needs to be listed in the WordNet of that language and represented in the interlingual ontology. It need not be included in the WordNets of other languages where it is compositional as translation can proceed via the ontology. We expect this to be the case for many, if not most, noun compounds. In the EuroWordNet (Vossen, 1998) the relation Deriv existing e.g. between *teach* and *teacher* was introduced and similar links can be found in the PWN version 3.0. (Fellbaum, Osherson & Clark, 2007). Previous versions of WordNet included manually added links between words that are both morphologically and semantically related. For example, (only) the appropriate sense of the noun *direction* is linked to the appropriate sense of the verb *direct*. Fellbaum, Osherson and Clark (op cit) identified a subset of these relations: noun-verb pairs whose semantic relation could be typed. They classified the noun-verb pairs according to the semantic role that the noun plays with respect to the event denoted by the verb. For example, one sense of *ruler* refers to an ARTIFACT with which one *rules* paper. Another pair *ruler-rule* is labelled ACTOR, as the noun refers to a head of state who governs over a

people.

5. Conclusion

We discussed basic types of noun compounds in Czech, English and Zulu. There are many commonalities as well as some compound-formation processes that are specific to Czech and Zulu that provide a still broader perspective.

The integration of compounds into lexical resources, and WordNets in particular, remains a challenge that needs to be considered in terms of the compounds' syntactic idiosyncrasy and semantic compositionality. The experiments with processing compounds in Czech, English and Zulu are reported and partly evaluated. The obtained partial lists of the Czech, English and Zulu compounds are also described.

6. Acknowledgements

This work has been partly supported by the Ministry of Education of Czech Republic under the projects LC526, NPV II 2C06009 and by the Czech Grant Agency under the project GA 201/05/2781 and project GA 407/07/0679. Christiane Fellbaum's work is supported by the U.S. National Science Foundation grant CNS 0855157. We would like to express our thanks to Pavel Šmerk, MA, who programmed the Perl script for processing Zulu compounds as well as the tool Deriv for working with Czech compounds.

7. References

Agirre, E., Aldezabal, I., Pocielli, E. (2006). Lexicalization and multiword expression in the Basque WordNet. *Proceedings of Third International WordNet Conference*, pp.131--138. ISBN 80-210-3915-9. Jeju Island (Korea).

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing, in COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, Switzerland, August 28, 2004, pp. 101--108.

Bosch, S., Fellbaum, C. (2009). A Comparative View of Noun Compounds in English and Zulu. In: Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.): *After Half a Century of Slavonic Natural Language Processing*, pp.35--44. Brno: Masaryk University.

Bosch, S., Fellbaum, C., Pala, K. (2008). Derivational Relations in English, Czech and Zulu Wordnets. *Literator* 29(1):139--162.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fellbaum, C., Osherson, A., Clark, P.E. (2007). Putting Semantics into WordNet's ``Morphosemantic'' Links. In: *Proceedings of the Third Language and Technology Conference, Poznan, Poland*, October 5-7, 2007. Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies, eds. Z. Vetulani and H. Uszkoreit. Springer Lecture Notes in Informatics vol. 5603:350--358 (2009).

Hlaváčková, D., Osolsobě, K., Pala K., Šmerk, P. (2009). Exploring Derivational Relations in Czech with the Deriv Tool. In: *Proceedings of the Slovko Conference, Bratislava*, in print.

Hlaváčková, D., Pala, K. (2007). Derivational Relations in Czech Wordnet, *Workshop BSNLP, ACL*, Prague 2007, pp.75--81.

Kilgarriff A., Rychlý P., Smrž P., Tugwell D. (2004). The Sketch Engine. In: *Proceedings of the Eleventh EURALEX International Congress, Lorient*, 2004, pp. 105--116

Kim, S.N., Baldwin, T. (2008). An Unsupervised Approach to Interpreting Noun Compounds. In: *Proceedings of the International Conference on NLP and KE, Beijing, China*.

Kim, S.N., Mistica, M., Balwin, T. (2007). Extending Sense Collocations in Interpreting Noun Compounds. In: *Proceedings of the Australasian Language Technology Workshop*, pp.49--56.

Kosch, I.M. (2006). *Topics in Morphology in the African Language Context*. Pretoria: Unisa Press.

Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Niles, I., Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412--416. See also http://www.ontologyportal.org/

Petr, Jan et al. (1968). *Mluvnice češtiny II*, (Grammar of Czech II), Prague, Academia, p. 451--485.

Sedláček, R. (2004). *Morphematic analyser for Czech*. PhD thesis, Faculty of Informatics, Masaryk University, Brno.

Slovník spisovného jazyka českého (Academic Dictionary of Literary Czech). (1960), Academia, Prague.

Slovník spisovné češtiny (Dictionary of Written Czech). (1986). Academia, Prague.

Šmerk, P. (2006). Deriv. Web application interface (in Czech), accessible at: http://deb.fi.muni.cz/deriv.

Ungerer, H.J. (1983). Komposita in Zulu. Unpublished Doctoral Thesis. Johannesburg: Randse Afrikaans University.

Vossen, P. (1998, ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

WordNet a lexical database for the English language. (2010). Available: http://wordnet.princeton.edu/ Accessed on 8 March 2010.