More Than Only Noun-Noun Compounds: Towards an Annotation Scheme for the Semantic Modelling of Other Noun Compound Types

Ben Verhoeven
CLiPS - Computational Linguistics Group
University of Antwerp
Antwerp, Belgium
ben.verhoeven@ua.ac.be

Gerhard B. van Huyssteen
Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Abstract

The computational processing of compound semantics poses several interesting challenges. Up to now, the processing of nominal compounds with non-noun left-hand constituents (henceforth XN compounds) has not received any attention, despite the fact that these also seem to be rather productive in Germanic languages. In our research project, we aim to fill this hiatus by investigating various kinds of compounds in Afrikaans and Dutch, develop annotation protocols and data sets, and model the semantics of such compounds. In this publication we present the alpha version of an annotation protocol that was designed for both descriptive linguistic and computational linguistic purposes. We describe the protocol development and discuss the current version.

1 Introduction

Within the field of natural language understanding, the semantic processing of compounds poses several interesting challenges, including issues related to compositionality, ambiguity, and contextual interpretation (see Girju *et al.* (2005) for a more elaborate discussion). The majority of research up to now has focused on English, but surprisingly, virtually no research has been done for other (Germanic) languages (cf. Verhoeven, 2012; Verhoeven, Daelemans & van Huyssteen, 2012). Also, given that noun-noun (NN) compounds are by far the most productive form of compounding in English (Plag, 2003: 145), it is to be expected that research on the semantic analysis of English compounds (both in descriptive linguistics and computational linguistics) has thus far focused almost exclusively on NN compounds (see Ó Séaghdha (2008) for a comprehensive overview, as well as Adams (2001: 83ff) for a synopsis). A computational understanding of compound semantics is of importance for commercial applications such as machine translation systems, where one often has to paraphrase compounds (i.e. make the compound semantics explicit at surface level) to be able to translate them into languages that are not as productive in compounding, or that has different compound constructions (Nakov, 2008).

Second to NN compounds, nominal compounds with non-noun left-hand constituents (henceforth XN compounds; i.e. other nominal compound types) seems to be the most productive in English (see Lieber, 2009), and probably in other Germanic languages as well. However, as far as we could establish, no research has been done on the computational modelling of the semantics of XN compounds in any language; hence, no annotation guidelines, data sets or prior experiments are available. In our research project, we aim to fill this hiatus by investigating various kinds of compounds in Afrikaans and Dutch, develop annotation protocols and data sets, and model the semantics of such compounds.

In this contribution, we present a first version of an annotation protocol for XN nouns, specifically for Afrikaans and Dutch (but also referring to English in passing). The next section presents a brief linguistic description of XN compounding in Afrikaans and Dutch. In section 3 we discuss some general principles for compound annotation, before presenting the detailed protocol. In section 4 we

	Afrikaans (Afr.)	**Dutch (Du.)**	**English (Eng.)**
NN	*tafelblad* 'table top'	*pannenkoek* 'pancake'	*car key*
VN[1]	*faksmasjien* 'fax machine'	*leesbril* 'reading glasses'	*skateboard*
AN[2]	*geelwortel* 'carrot'	*geelzucht* 'yellow fever; jaundice'	*lightweight*
PN[3]	*onderrok* 'under skirt; petticoat'	*achterlicht* 'back light'	*undertone*
QN[4]	*agthoek* 'octagon'	*eenoog* 'cyclops	*twoface*

Table 1: Examples of NN and XN compounds in Afrikaans, Dutch and English.

conclude with a view on future research.

2 XN Compounding in Germanic Languages

Compounding is a highly productive word-formation process in most languages (Plag, 2003), and as such has received much attention in research literature (e.g. Lieber and Štekauer, 2009). With regard to typologies of compounding, Scalise and Bisetto (2009) provide a comprehensive overview, and also present the most recent morphological compound classification scheme that is based on the compound's internal syntactic function. With regard to the syntactic form of compounds, Plag (2003) indicates that nominal compounds occur widely in English (with NN compounds the most common type); Don (2009: 370-371) maintains the same for Dutch: "Nominal compounds are by far the most productive type, although other types (adjectival and verbal) exist and can also be formed productively". Since the same holds true for German (Neef, 2009: 388) and Danish (Bauer, 2009: 404), we may safely assume that it also applies to Afrikaans, a West Germanic language, closely related to Dutch. Compare Table 1 for examples of NN and XN compounds in Afrikaans, Dutch and English.

The most important challenge with regard to interpreting VN compounds is whether the V should be interpreted as a V or an N in languages where a distinction between these forms are not marked overtly, or where the (lack of) morphology could lead to ambiguous interpretations. For example, in *swimming pool*, the question is whether *swimming* should be interpreted as a V ('pool where one swims') or as a N ('pool for the act of swimming') (see Lieber, 2009: 361). When using the continuous participle form of English verbs (the *-ing* forms) as a noun, it "does not describe a single episode of the process, but instead rather refers to it in a generalised, even generic fashion" (Langacker, 1987: 208). It is therefore natural to assign an N interpretation to *swimming*, and consequently regard *swimming pool* (and the likes) as an NN compound.

In contrast, in the Dutch *zwembad* **swim+bath** 'swimming pool', a V interpretation is assigned to the first constituent (a verbal stem), i.e. 'bath where one swims'. Most of the time, a verbal interpretation is the only option, since the infinitive form of the verb (e.g. *zwemmen*) is usually used as the nominalised form (as in *Ik hou van zwemmen* **I like of swim-INF** 'I like swimming'). Hence, in Dutch we often find VN compounds.

Since Afrikaans does not have an overtly marked infinitive form of the verb, it might seem to be more ambiguous to distinguish whether *swem* in *swembad* **swim+bath** 'swimming pool' is a verb or a noun (i.e. the part-of-speech category of *swem* remains ambiguous). However, because of the close relationship between Dutch and Afrikaans, we will treat their compounds equally and thus consider these stems as verbs; see Section 3.1. below.

With regard to AN compounds, we should note that none of the three Germanic languages under discussion allow for productive AN compounding, since an A and N usually forms a noun phrase (NP), e.g. *white cloud* is considered an NP, and not an AN compound. However, all three languages do allow for compounding when there are signs of extension of meaning. For example, a *blackboard* is more than just 'a board with the colour black' - it is more specifically 'a dark-coloured surface where one could write on with chalk'. In all three lan-

[1]VN = Verb-Noun Compound
[2]AN = Adjective/Adverb-Noun Compound
[3]PN = Preposition-Noun Compound
[4]QN = Quantifier/Numeral-Noun Compound

guages, such cases are most often written as one word, and thus more easily distinguishable from NPs. (Of course, the orthography is a result of the compounding process, rather than a cause.) Note that it seems as if this phenomenon is found more frequently in Afrikaans orthography than in English or Dutch, e.g. Afr. *witwyn*, Du. *witte wijn*, Eng. *white wine*; or Afr. *swartmark*, Du. *zwarte markt*, Eng. *black market* 'underground economy for trading illegal goods'; further comparative linguistic research is needed to confirm this observation. All cases of AN compounds should therefore be considered lexicalised, although certain patterns in the semantics of such compounds might become apparent (see Section 3.2 below).

Similarly, all QN compounds in these three languages are lexicalised - see Afr. *agthoek*, Du. *eenoog*, Eng. *twoface* in Table 1 above. However, a special case of phrasal compound could be distinguished: $[[\mathrm{Q\ N}]_{\mathrm{N}P}\ \mathrm{N}]_{\mathrm{N}}$, as in Afr. *derdejaarstudent*, Du. *tweepersoonsbed*, Eng. *three-phase electricity*. Booij (2002: 150-151) presents an argument that one could consider such constructions also as NN compounds (i.e. $[[\mathrm{QN}]_{\mathrm{N}}\ \mathrm{N}]_{\mathrm{N}}$), but we are of contention that it makes more sense in the context of compound semantics to consider it phrasal compounds, i.e. a *derdejaarstudent* is 'a student in his/her third year', a *tweepersoonsbed* is 'a bed for two people', and *three-phase electricity* is 'electricity with three phases'. Currently, such compounds are excluded from our focus, since this protocol only deals with two-part compounds, as will be indicated in the next section.

3 Protocol Design

The design of our XN compound semantics protocol is based on the work by Ó Séaghdha (2008) on NN compounds. We adopted his approach of semantic categorisation and used his categories as basis for the construction of a protocol for XN compounds in Dutch and Afrikaans. The protocol deals mainly with two-part compounds, and hence phrasal compounds and recursive compounds are excluded from the scope of our current research.

Also note that the version of the protocol presented here is still an alpha version, and has not yet been verified (i.e. tested and extended) on a representative dataset of compounds. A complete version of this protocol, as well as subsequent updated versions of the protocol are available on the Sourceforge page of the AuCoPro project[5].

In concordance with the approach of Verhoeven (2012) and Verhoeven, Daelemans and Van Huyssteen (2012) on the computational understanding of compounds, all compounds that are listed in a standard explanatory dictionary are considered lexicalised when using the protocol for computational experiments. These lexicalised words do not need a computational interpretation, because their meanings are already present in the dictionary glosses. For purposes of descriptive linguistics, using dictionary compounds in non-lexicalised categories is allowed when their meanings are the product of a clear relation between the two constituents. This distinction between lexicalised and non-lexicalised leaves room for interpretation in descriptive linguistics, but it is a practical measure for computational purposes.

Exocentric compounds, such as Afr. *banggat* **afraid+bottom** 'person that is easily frightened'; Du. *kaalkop* **bald+head** 'person with a bald head', Eng. *uphill* (i.e. $[\mathrm{PN}]_{\mathrm{A}dv}$) are always lexicalised and thus also tagged as lexicalised, following the LEX category of Ó Séaghdha (2008). Endocentric compounds can be either lexicalised or non-lexicalised (and thus productive). Endocentric compounds with lexicalised meanings do not explicate the relation between the constituents in a predictable manner, i.e. they are fully non-compositional. There is thus one more differentiation within the lexicalised category: such compounds can be classified as either endocentric or exocentric.

The main distinction between compound types in our protocol is between the parts-of-speech of the first constituent. We consider the following main categories: verb, adjective (or adverb), quantifier (or numeral), or preposition.

3.1 Verb-Noun Compounds (VN)

This category contains two-part compounds that take a verb as a first constituent and a noun as a second constituent. The first constituent will only be considered a verb if it cannot be interpreted as a noun. That is, in *zwembad* **swim+pool** 'swimming

[5] https://sourceforge.net/projects/aucopro/

pool', the constituent *zwem* can only be interpreted as a verb, and is hence assigned a V interpretation (unlike the case in English; see discussion above).

3.1.1. Event

This category is based on the INST and ACTOR categories in Ó Séaghdha's protocol (2008). In our protocol, the verb describes an action in which the noun is some sort of participant. There are three subcategories to this rule: the nominal element can be the subject, object or instrument of the action described by the verb. Although it might be interesting to consider using semantic roles (e.g. from Frame Semantics) as subcategories, this might also lead to an abundance of fine-grained semantic classes, resulting in more problems than gains for automatic classification. We opine that such a classification task (i.e. using fine-grained semantic roles) would be a particularly hard task even for human annotators, while the semantic role information could be deduced broadly from the combination of the verb semantics with the syntactic role of the noun.

- Subject
  ('N that Vs; the goal of N is to V')
  Afr. *snydokter* **cut+doctor** 'doctor that cuts; surgeon'
  Du. *gloeilamp* **glow+lamp** 'lamp that glows; lightbulb'

- Object
  ('N that is (being) V-ed; VN is the result of V-INF; the goal of N is to be V-ed')
  Afr. *snyblomme* **cut+flowers** 'the goal of the flowers is to be cut'
  Du. *werpbal* **throw+ball** 'ball that is thrown'

- Instrument
  ('N is used to V-INF')
  Afr. *kapbyl* **chop+axe** 'axe used to chop down trees'
  Du. *leesbril* **read+glasses** 'glasses that are used to read; reading glasses'

3.1.2. Location

This category practically equals Ó Séaghdha's IN category (2008). It contains those VN compounds in which the noun is a spatial or temporal location (two subcategories) of the action described by the verb.

- Space
  ('V in (neighbourhood of) N; N where one Vs')
  Afr. *herstelsentrum* **recover+centre** 'centre where people recover from injuries or operations'
  Du. *slaapkamer* **sleep+room** 'room where one sleeps; bed room'

- Time
  ('N during which one Vs')
  Afr. *bakleifase* **quarrel+fase** 'fase during which one quarrels'
  Du. *regeerperiode* **rule+period** 'period during which someone rules'

3.1.3 Composed of

This category can best be compared with the part-whole and group interpretation of the HAVE category in Ó Séaghdha (2008). The noun is some sort of collection of the action described by the verb. The compound can best be paraphrased as 'N consists of V', e.g.:

Afr. *skokterapie* **shock+therapy** 'therapy that consists of shocking the patient'
Du. *niesbui* **sneeze+shower** 'rapid succession of sneezes'

3.1.4. Lexicalised

As indicated above, lexicalised compounds can be either endocentric or exocentric; both subcategories are excluded from computational experiments.

- Endocentric
  Afr. *snyhou* **cut+stroke** 'kind of tennis stroke'
  Du. *draaibal* **turn+ball** 'ball that is kicked with a turning effect'

- Exocentric
  Afr. *speeltuin* **play+garden** 'playground'
  Du. *verzamelwoede* **collect+anger** 'urge or mania to collect things'

3.2 Adjective-Noun Compounds (AN)

In our research thus far, we found all AN compounds to be lexicalised, since the normal pattern in Germanic languages is to consider A + N as a syntactic phrase (see Section 2 above). We will therefore not consider this category for computational experiments, but for descriptive

completeness, we do posit some subcategories for concatenated AN compounds.

3.2.1. Lexicalised

- Endocentric

Most examples under this category can be matched to certain aspects of Ó Séaghdhas (2008) ABOUT category, where the first constituent (A) describes a characteristic of the concept defined by the second constituent (N). Note that the A provides a more precise, fuller specification of the concept in the domain of instantiation (Langacker, 2008: 134-136), invoking a variety of cognitive domains (Langacker, 1987: 117). From our initial data analyses, we posit "Duration" and "Colour" as prototypical domains (specifically for Afrikaans), but we also posit an "Other" category, leaving the door open that more subcategories could be defined in further linguistic research and data analysis.

- Duration
('kind of N that is A')
Afr. *langverlof* **long+leave** 'kind of leave that is longer than what is normally taken'
Du. no examples found

- Colour
('kind of N that is A')
Afr. *geelrys* **yellow+rice** 'kind of rice that is yellow'
Du. *rodekool* **red+cabbage** 'kind of cabbage that is red'

- Other qualities
('kind of N that has the quality expressed by A')
Afr. *sterkstroom* **strong+current** 'high voltage; the power current is strong'
Du. *hogeschool* **high+school** 'school for higher education'

- Exocentric

This category of lexicalised AN compounds contains those compounds of which the semantic head is not present in the compound. Often, they are possessive compounds where the compound is an entity that has the characteristic described by the noun modified by the adjective.

- Attributive (Scalise and Bisetto, 2009: 36); also known as possessive or bahuvrihi compounds (Bauer, 2004: 21)
Afr. *luigat* **lazy+bottom** 'person that is lazy'
Du. *kaalkop* **bald+head** 'person that has a bald head'

- Other
Afr. *groenskrif* **green+script** 'first draft of legislation; green paper'
Du. *blijspel* **happy+game** 'theatre play that is supposed to amuse people'

3.3 Quantifier-Noun Compounds (QN)

In this category, we consider quantifiers and numerals as first constituent of a two-part compound that has a noun as a second constituent.

3.3.1. Quantity-Object
The quantifier that specifies the quantity of N within a larger phrasal compound (i.e. [$[\text{Q+N}]_{NP}$ $\text{N}]_{\text{N}}$) is the only productive form of QN compounding (e.g. Afr. *sewejaardroogte* **seven+year+drought** 'seven-year drought') (see Section 2 above). Since these are not two-part compounds, they fall outside the scope of our current research project.

3.3.2. Lexicalised
Many of the lexicalised QN compounds are exocentric compounds, with a notable number of them being plant and animal names.

- Endocentric
No examples in Afrikaans or Dutch have been found yet.

- Exocentric - Attributive
(compound is 'entity that has Q number of N')
Afr. *vierkleur* **four+colour** 'flag of the old Transvaal Republic'
Du. *duizendpoot* **thousand+leg** 'centipede'.

3.4 Preposition-Noun Compounds (PN)

All compounds that have a preposition as a first constituent and a noun as the second constituent belong in this class, even when the prepositions have adopted a more abstract or metaphorical meaning.

3.4.1. Location

This category also relates to Ó Séaghdha's IN category (2008). The concept described by N is at a position P of an undefined other concept. In three different subclasses, the preposition describes a spatial, temporal, or more abstract/metaphorical position. The paraphrases of these categories contain an undefined concept 'G' that is used as reference point (i.e. grounding point).

- Space
  ('N is spatially at position P relative to G')
  Afr. *onderrok* **under+skirt** 'skirt worn under other skirt'
  Du. *achterlicht* **behind+light** 'light at behind of car or bike; rear light'

- Time
  ('N is temporally at position P relative to G')
  Afr. *voormiddag* **before+noon** 'forenoon'
  Du. *nagesprek* **after+talk** 'conversation after previous event'

- Abstract/Metaphorical
  ('N is at abstract position P relative to G')
  Afr. *byverdienste* **by+income** 'additional income to normal income'
  Du. *overgewicht* **over+weight** 'the weight that is over the normal'

3.4.2. Process-based

We assume this kind of PN compound to be related to some kind of process. The noun goes in the direction described by the preposition ('N goes in direction P'), e.g.:

Afr. *opmars* **up+march** 'march'
Du. *overstap* **over+step** 'transfer on public transport'

3.4.3. Lexicalised

- Endocentric
  Afr. *optog* **up+trip** 'procession'
  Du. *uitgroeisel* **out+growth** 'excrescence'

- Exocentric
  Afr. *insig* **in+sight** 'insight'
  Du. *nageboorte* **after+birth** 'afterbirth'

4 Conclusion and Future Work

We have presented the alpha version of an annotation protocol for the semantics of Dutch and Afrikaans noun compounds which have a non-noun as a first constituent. Although this protocol is primarily designed for computational linguistic purposes, we have also indicated some categories relevant to (comparative) descriptive linguistics. Basic points of departure (based on work by Ó Séaghdha (2008)) have also been described.

During the development of the protocol, we came across some interesting findings that should be verified in further research. For example, it seems as if all two-part AN and QN compounds are lexicalised, probably because the more regular A + N and Q + N constructions in Germanic languages are syntactic phrases. In some categories we could not find examples yet, these should be investigated in further corpus-based/-driven research.

Also, the way we constructed the event-based category for VN compounds (see Section 3.1.1. above) is open for closer scrutiny. Having separate subcategories for subject, object, instrument and goal/result relations seems an interesting adaptation of the INST and ACTOR categories in Ó Séaghdha (2008). We believe it is worth considering the adjustment of Ó Séaghdha's INST and ACTOR categories to be more like our categories in combining the several participants of the event on which the compound is based. This would, in our opinion, make the annotation process easier because it does away with the 'direction' of the annotation rules that Ó Séaghdha uses.

As part of the continuous development of our current protocol, we are currently in the process of annotating Dutch and Afrikaans compounds, using this protocol. The annotation process will proceed as described by Verhoeven (2012). We are using the compound database CKarma (CTexT, 2005) for Afrikaans and a compound list extracted from the e-Lex corpus for Dutch[6]. Eventually, this annotated data will be used in computational experiments to predict the semantics of a variety of compounds in these two languages. The results of these experiments will be published later.

[6]This list was extracted from the e-Lex corpus and annotated by Lieve Macken from LT3 at Ghent University College.

Future work on the semantics of compounds includes, but is not limited to: the investigation of affixoid-noun compounds where an adverb-like affixoid combines with a noun, such as Afr. *laatherfs* **late+autumn** 'late autumn'; Du. *tegenaanval* **against+attack** 'counter-attack'; and Eng. *co-inhabitant*; investigation of the semantics of compounds with different parts-of-speech such as XA (e.g. Afr. *bloedrooi* **blood+red** 'very red') and XV (e.g. Afr. *stofsuig* **dust+suck** 'vacuum/hoover') compounds; research into regularities that could be found in the construction and meaning of phrasal compounds.

Acknowledgments

Automatic Compound Processing (AuCoPro[7]) is a collaborative project by research groups of the North-West University (Potchefstroom, South Africa), the University of Antwerp (Belgium) and Tilburg University (The Netherlands), and is funded through a research grant from the Nederlandse Taalunie (Dutch Language Union) and the South African Department of Arts and Culture (DAC), as well as a grant of the South African National Research Foundation (NRF) (grant number 81794).

We would like to acknowledge the work of Joanie Liversage (North-West University), who attempted a first analysis of XN compounds in her third-year mini-dissertation, as well as the inputs of Walter Daelemans (University of Antwerp).

References

Valerie Adams. 2001. *Complex Words in English*. Longman, Harlow, UK.

Laurie Bauer. 2004. *A Glossary of Morphology*. Edinburgh University Press, Edinburgh, UK.

Laurie Bauer. 2009. IE, Germanic: Danish. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 400–416. Oxford University Press, Oxford, UK.

Geert Booij. 2002. *The Morphology of Dutch*. Oxford University Press, Oxford, UK.

CTexT. 2005. CKarma (C5 KompositumAnaliseerder vir Robuuste Morfologiese Analise). [C5 Compound Analyser for Robust Morphological Analysis]. Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa.

[7]http://tinyurl.com/aucopro

Jan Don. 2009. IE, Germanic: Dutch. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 370–385. Oxford University Press, Oxford, UK.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19:479–496.

Ronald Langacker. 1987. *Foundations of Cognitive Grammar: Volume 1 - Theoretical Prerequisites*. Stanford University Press, Stanford.

Rochelle Lieber and Pavol Štekauer. 2009. Introduction: status and definition of compounding. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 3–18. Oxford University Press, Oxford, UK.

Rochelle Lieber. 2009. IE, Germanic: English. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 357–369. Oxford University Press, Oxford, UK.

Preslav Nakov. 2008. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA08)*.

Martin Neef. 2009. IE, Germanic: German. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 386–399. Oxford University Press, Oxford, UK.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

Ingo Plag. 2003. *Word-Formation in English*. Cambridge University Press, Cambridge, UK.

Sergio Scalise and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Compounding*, pages 34–53. Oxford University Press, Oxford, UK.

Ben Verhoeven, Walter Daelemans, and Gerhard B. van Huyssteen. 2012. Classification of noun-noun compound semantics in Dutch and Afrikaans. In *Proceedings of the Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2012)*, pages 121–125, Pretoria, South Africa.

Ben Verhoeven. 2012. A computational semantic analysis of noun compounds in Dutch. Master's thesis, University of Antwerp, Antwerp, Belgium.