Originalveröffentlichung in: Gippert, Jost/ Gehrke, Ralf (Hrsg.):
Historical Corpora. Challenges and Perspectives.–
Tübingen: Narr, 2015. S. 91-99. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5)

GAYE DETMOLD / HELMUT WEISS[1]

Historical corpora and word formation

How to annotate a corpus to facilitate automatic analyses of noun-noun compounds

Abstract

In this paper we present some preliminary considerations concerning the possibility of automatic parsing an annotated corpus for N-N compounds. This should in principle be possible at least for relational and stereotype compounds, if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). These lexical entries then supply the necessary information about the argument structure of a relational noun or about the stereotypical purpose associated with the noun's referent which can be used to establish a relation between the first and the head constituent of the compound.

1. Introduction

In our paper we present the outline of a research project on noun-noun (N-N) compounds in Old High German (OHG). The main focus of the paper lies on topics of corpus annotation with respect to word formation – a topic hardly ever addressed before to our knowledge.

The project is mainly concerned with three issues:

1) Structural aspects

2) Interpretation of N-N compounds

3) Corpus-linguistic aspects

In order to achieve our goals, we proceed in two steps:

First we build a data base which contains the N-N compounds listed in the OHG dictionary of Jochen Splett (1993). In a second step we try to investigate the question of which kind of corpus annotation is necessary to facilitate a quasi-automatic analysis of N-N compounds. The data base should deliver the empirical basis for the investigation of the structural aspects and the internal semantics of N-N compounds. It will contain all the relevant information for

[1] We want to thank two anonymous reviewers for helpful comments.

this purpose such as, e.g., the internal structure or the type of compound. To prepare a corpus annotation we try to investigate which and how much information must be given in the annotation to enable more or less automatic semantic analyses of N-N compounds in annotated corpora.

2. Classification of N-N compounds

In our project we make use of an approved classification of N-N compounds (cf. Olsen 2000, Meibauer et al. 2002) to analyze the OHG N-N-compounds listed in the OHG dictionary of Splett. We compile a data-base which contains all N-N-compounds and classifies each of these compounds according to this classification, as far as this is possible. We expect that this classification will be sufficient to analyze a great deal of the OHG compounds but have to be refined in some way (for example with respect to coordinative compounds, cf. Çinkılıç/Weiß 2012). We will only consider non-lexicalized forms, whose meaning is built up in a compositional way.

In this paper, we just give a short impression of the classification we are using (for further details on this topic, cf. Çinkılıç/Weiß 2013). Compounds could be subdivided into the following subclasses (of which only determinative compounds will be considered in the remainder of the paper):

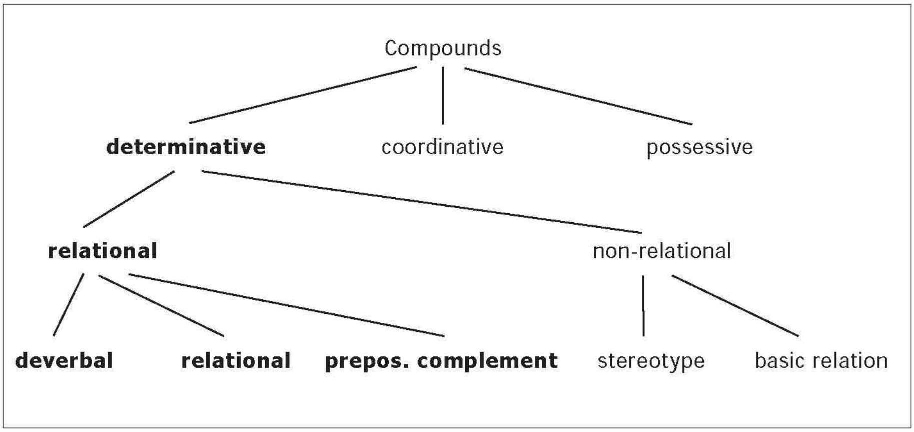

Figure 1: Subclasses of compounds

2.1 Relational compounds

Relational compounds are defined as compounds where the head is a relational noun and the first constituent is its argument. There are three subtypes:

Deverbal head

The first type includes relational compounds with a deverbal head, i.e. the noun is derived from a verb and the first constituent qualifies as an internal argument.

Examples:[2]

Hûs-eigo	‘house lord’	< *eigan* ‘to possess’
Wind-fanga	‘porch’	< *fâhan* ‘to grip, catch’
Obaz-traga	‘fruit carrier’	< *tragan* ‘to carry’

Noun with prepositional complement as head

The head of the second type is a noun which can take a prepositional complement. Note that the relevant examples *minna* and *lust* are not derived from a verb.

Examples:

minna	‘love of/for’
heim-minna	= ‘love of home, patriotism’
mâg-minna	= ‘love of relatives’
lust	‘desire for’
minna-lust	= ‘desire for love’
weralt-lust	= ‘desire for the world’

Relational noun as head

The third type includes compounds with an inherently relational noun as head.

Examples:

sun	‘son’
basun-sun	= ‘son of the aunt’
brouder-/swester-sun	= ‘nephew’
huorûn-sun	= ‘sun of a whore’

[2] All examples in this section are taken from Çinkılıç/Weiß (2013).

A noun like *son* is inherently relational because it denotes a relation between two individuals (A is the son of B). The first constituent of the compound can function as member of such a relation, so a *basun-sun* is the son of the aunt and *swester-sun* denotes the son of the sister.

2.2 Non-relational compounds

Non-relational compounds are ones where the relation between first and head constituent which underlies the interpretation is not evident by the argument structure or the lexical meaning of the head, but comes from elsewhere. There are two types of non-relational compounds:

Stereotype

Because of our world knowledge we have built some stereotypes in our minds. Such stereotypes comprise, for example, the knowledge that artifacts are made for certain purposes, and this knowledge can be the basis of the interpretation, cf.:

teig-/wazzartroc	'dough/water trough'

It is a stereotypical aspect of the lexical meaning of 'trough' that it is made to contain something – and this can be named as the first constituent.

bouhscrîni	'bookcase'

A cupboard (= *scrîni*) is a piece of furniture used for storage, in this case for the storage of books.

Basic relations

The interpretational relations between the first and the head constituent can be basic relations which "are part of the mental algebra that processes meanings combined with other meanings" (Olsen 2000: 909). We use basic relations everywhere to classify things according to their shape, consistence, composition, purpose, or their tempo-spatial localization – and such basic relations can be inferred as holding between the constituents of an N-N compound. So a *Holzhaus* is a 'house made of wood' or the *Mittagessen* 'the meal consumed at noon'.

As for OHG, there are numerous examples in Splett's dictionary for most of the basic relations, especially for LOC, TEMP and CONST (cf. Çinkılıç/Weiß 2013 for a more explicit description of the basic relations):

LOC	*hals-adra* ‘carotid (artery)’ – *herz-âdra* ‘aorta’
TEMP	*âband-muos* ‘dinner’ – *âband-sterno* ‘evening star’
PURP	*ambaht-hûs* ‘workshop’
CONST	*agat-stein* ‘agate stone’ – *stein-ofen* ‘stone oven’ – *fîg-flado* ‘cake made with figs’
INSTR	*scif-wîg* ‘battle with ships’ – *scilt-spîl* ‘fight with shields’
CAUSE	*wurm-âz* ‘damage caused by worms’
THEME	*êwabuoh* ‘law book’
SIM	*goldamaro* ‘yellowhammer’
PART	*swert-helza* ‘hilt of a sword’ – *huot-snuora* ‘hat string’

3. Corpus annotation

Apart from the data base and the qualitative (and quantitative) analysis of OHG N-N compounds, a further focus of the project lies on topics of corpus annotation. We try to elaborate which and how much information must be given in the annotation to enable more or less automatic semantic analyses of N-N compounds in annotated corpora. Questions to address here are, e.g., how to integrate context information or information from the lexical entry of the head constituent regarding its argument structure, its relational nature, etc. The challenge is, among others, to make visible information concerning the internal relation of the constituents (i.e. information below the actual ‘word boundaries’) and to guarantee access to it. To develop a special design for digital corpus studies of word formation is thus a further main object of investigation. In the following, we will present some preliminary considerations concerning this task.

As is known (cf. Linde 2011), annotating linguistic corpora involves morphological tagging concerning various kinds of information, e.g., parts of speech (POS) tagging, grammatical feature encoding, morphological segmentation, and morpheme type encoding. With this information at hand, it is already possible to automatically detect N-N compounds in an annotated text. If information about the type of the morphemes a complex word consists of is supplied, it is possible to automatically identify compounds – as well as any other kind of complex word.

However, this information does not suffice to identify the ‘semantic’ type of a N-N compound. As described in section 2, the types of N-N compounds are distinguished according to the semantic relation between the head and the first constituent. The first type is established by relational compounds where the head is a relational noun and the first constituent represents an argument of the head. As mentioned above, the head can be a deverbal noun, a noun with a prepositional complement, or an inherently relational noun. So we need access to two types of information: (i) information whether the head noun is of such a relational kind, and (ii) information whether the first constituent can serve as an argument of the head noun.

At least the first kind of information can easily be retrieved if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). Such lexical entries store information concerning the phonology, morphology, syntax, semantics, and pragmatics. For our purpose, it is crucial that lexical entries contain syntactic and semantic information of the necessary kind.

We will present a concrete example for how this automatic analysis could probably work (cf. Çinkılıç/Weiß 2013). The OHG lexicon of Splett (1993: vol 2, 1008) lists several compounds with *tragâri* ‘carrier’ as their head noun, e.g.:

lioht-tragâri	‘candle carrier’
spera-tragâri	‘spear carrier’
stank-tragâri	‘scent carrier’
swert-tragâri	‘sword carrier’
wazzar-tragâri	‘water carrier’
wolla-tragâri	‘wool carrier’

The head noun *tragâri* is a deverbal noun: it consists of the verbal root *trag* ‘to carry’ and a nominal suffix, -*âri*, which derives nouns from verbs:

$$[\mathit{trag}]_{\mathrm{v}} + [\mathit{\hat{a}ri}]_{\mathrm{Nsuf}}$$

Now, it is crucial that both morphemes of the head noun are listed separately in the lexicon, i.e., that the derivational suffixes also get lexical entries. The lexical entries of -*âri* and *trag*- would look like this:

Lexeme	*-âri*
PHON	/ɑ:rɪ/
MORPH	Masculin
SYN	N_{af} [V _]
SEM	AGENT or INSTRUMENT which executes the V-action
PRAG	

Table 1: Lexical entry of *-âri*

Lexeme	*trag-*
PHON	/trag/
MORPH	Strong inflection
SYN	V [NP_{nom}1, NP_{akk}2]
SEM	Activity TRAG (x1, x2) x1: AGENS, x2: THEME
PRAG	neutral register

Table 2: Lexical entry of *trag-*

From the lexical entries we can see that *tragâri* is a deverbal noun. In addition to that, the lexical entry of the verb *tragan* provides the necessary information about its argument structure: it takes two arguments and one of them is a theme argument. So compounds with the head -*tragâri* could in principle be relational compounds – if the first constituent is a possible theme of *tragan*. However, the final decision would then require to look at the lexical entry of the first constituent – in our examples the lexical entries of 'candle', 'scent', 'sword', 'water', and 'wool' – to see whether they qualify as possible themes for the verb *tragan* or not.

However, at the moment it is hard to imagine how the information whether the first constituent is a possible argument of the relational head noun or not can be specified in the lexical entry. In addition, there are some further principal problems. One problem is that words can be used non-literarily; e.g., one cannot only carry material things like the ones in the examples mentioned above but also immaterial ones (cf. *lugi-tragâri* lit. 'lie carrier', i.e. 'someone who spreads lies'). Another problem arises from the general possibility to interpret relational compounds as non-relational ones; so a *Steinträger*, lit. 'stone-carrier', can denote a person who carries stones (as a profession), or it denotes a pillar made of stone. The second meaning is based on the basic relation CONST ('X consists of Y'). Which interpretation is meant can only be decided on the basis of information coming from the co(n)text. Whether these problems can ever be solved in a satisfactory way remains open for further investigation.

Nonetheless the way sketched above is a conceivable way, at least to decide whether a given compound could be a relational compound or not.[3] In contrast to this, compounds with a basic relation would be harder – if not impossible – to analyze automatically, since basic relations are independent of the lexical meaning of words (Olsen 2000: 909), so access to lexical entries would not help. Here, one may use some kind of exclusion rule like: if the head is not a relational noun, infer some basic relation for the interpretation of the relation between the first constituent and the head noun.

However, stereotype compounds may be treated in a similar way as relational compounds, because one can still use lexical information in a broader sense to interpret their semantics. Stereotypes arise from opinions about typical properties associated with the referents of the words (Olsen 2000: 910). This kind of 'wor(l)d knowledge' may be part of the description of the meaning of a word in the mental lexicon or at least connected with it (i.e., be part of a neural network representing the word meaning). However that may be realized in our brains need not concern us here. The important thing is that it is possible to integrate such information into lexical entries.

In section 2.2 we already mentioned examples of stereotype compounds from OHG: *teig-/wazzartroc* 'dough/water trough', *bouhscrîni* 'bookcase'. A *troc* as in *teig-/wazzartroc* for instance is a trough where you can put something into – like dough or water. Troughs are artifacts and we know (whether a priori or from experience does not matter in our context) that artifacts are constructed to fulfill a certain purpose. This is one of the main differences to natural objects like trees or cats. It is obvious to think that the purpose an artifact is made for is part of the lexical meaning of the word denoting it. If this is the case, the lexical entry of *troc* 'dough' will contain a description of its possible purposes,

[3] The other two subtypes of relational compounds can be treated in the same way. As shown in section 2.1, the meaning of an inherently relational noun like *son* contains per definitionem a relation, because it denotes a male human offspring of *somebody*. The same holds for nouns with a prepositional complement.

and access to this information can be used to establish a relation between the head and the first constituent. So we can apply the same procedure for stereotype and relational compounds.

4. Conclusions

In this paper, we have mainly presented some preliminary considerations concerning the possibility of automatic parsing an annotated corpus for N-N compounds. This should in principle be possible at least for relational and stereotype compounds, if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). These lexical entries then supply the necessary information about the argument structure of a relational noun or about the stereotypical purpose associated with the noun's referent, which can be used to establish a relation between the first and the head constituent of the compound.

References

Çinkılıç, Gaye/Weiß, Helmut (2012): Kopulativkomposita. In: Linguistische Berichte 232: 417-435.

Çinkılıç, Gaye/Weiß, Helmut (2013): Historical word formation in German. On the interpretation of N-N compounds. In: Fisiak, Jacek/Bator, Magdalena (eds.): Historical English word formation and semantics. Frankfurt a.M./New York: Peter Lang, 211-227.

Höhle, Tilman N. (1982): Über Komposition und Derivation: zur Konstituentenstruktur von Wortbildungsprodukten im Deutschen. In: Zeitschrift für Sprachwissenschaft 1: 76-112.

Linde, Sonja (2011): Referenzkorpus Altdeutsch. Kurzbeschreibung. www.deutschdiachrondigital.de/data/home/manual/dateien/Manual.pdf (last accessed: January 29, 2015).

Meibauer, Jörg/Demske, Ulrike/Geilfuß-Wolfgang, Jochen/Pafel, Jürgen/Ramers, Karl-Heinz/Rothweiler, Monika/Steinbach, Markus (2002): Einführung in die germanistische Linguistik. Stuttgart/Weimar: Metzler.

Olsen, Susan (2000): Composition. In: Booij, Geert/Lehmann, Christian/Mugdan, Joachim (eds.): Morphology. A Handbook of Inflection and Word Formation. Berlin/New York: de Gruyter, 897-916.

Splett, Jochen (1993): Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 2 vols. Berlin/New York: de Gruyter.