Semantic Annotation of Genitive Attributes in a German Treebank

Maya Bangerter
University of Zurich
Institute of Computational Linguistics
E-mail: bangerter@cl.uzh.ch

November 25, 2008

Abstract

German genitive attributes are usually tagged as such in treebanks. However, it is well known that this information is not sufficient for determining the type of relation between head nouns and attributes, as genitive attributes can express many different semantic relations. Various linguistic classifications have been worked out, but to my knowledge, nobody has so far proposed to apply this linguistic knowledge to a corpus. The challenge here is to come up with a classification that is both easy to verify and sufficiently fine-grained. Using earlier linguistic approaches as guidelines, I propose in this paper a detailed annotation scheme for German genitive attributes based on readily identifiable noun features. First insights from its application to the Smultron Treebank show that it is easy to distinguish between the proposed classes and that my classification of genitive attributes can be related to a more general semantic annotation level.

1 Introduction

In recent years a lot of work was done in the improvement of parsers using corpus linguistic resources. Today, semantic annotations are becoming increasingly important. More semantic information would be a great advantage for many NLP applications. The knowledge of implicit semantic relations is, for example, crucial for the quality of question-answering systems. There is no other way to match a potential answer to a question such as in the pair below. [1]

[1] Example sentence from Smultron, Literary Part.

(1) was bedeutet der schwarze Zylinder?
What does the top hat signify?

(2) ... der schwarze Zylinder des Universums ist...
... the top hat of the universe is...

However, the difficulties in relation mining start with determining a classification. In most cases, semantic relations are classified according to a specific domain. There is no consensus among linguists about a general classification of semantic relations. But ad-hoc classifications done in computational linguistics suffer from the lack of linguistic foundation.
In German, genitive attributes are often used to express implicit semantic relations. The genitives extend a core noun phrase. They occur pre- and postnominally, always adjacent to the head noun. Genitives are easily identifiable and therefore it is possible to nest them arbitrarily deep. Nouns can take two genitive attributes in German.
In German language studies, various attempts to classify these attributes have been undertaken. The results are quite different in detail, but the classifications share a large fraction of classes. These can already be very useful to NLP applications.
In the following, I will first discuss advantages and drawbacks of two linguistic classifications, the semantic typology of Helbig/Buscha and the syntactic one of Lindauer. Then I will show how one can arrive at partitioning of similar granularity using exclusively formal criteria and describe first experiences with the annotation of a German corpus. In the last section I will briefly outline how the new distinctions could be annotated automatically using machine learning.

2 Linguistic Approaches

Semantic classification For building their typology of genitive attributes, Helbig and Buscha [6] rely on their *predicative deep structure*, i.e., they paraphrase the attributes as predicates in order to turn implicit semantic relations into explicit ones. According to Helbig/Buscha there are twelve different possibilities; for example, the possessive genitive is based on a relation of ownership ("Haben-Verhältnis") and the defining genitive is based on an is-a relation ("Sein-Verhältnis").[2]

(3) das Haus meines Vaters ← mein Vater hat ein Haus
the house of my father ← my father has a house

(4) die Pflicht der Dankbarkeit ← Dankbarkeit ist eine Pflicht
the duty of gratefulness ← gratefulness is a duty

[2] An exhaustive presentation of all possible classes is given in [6, p. 591]

More specific classes are based on more specific predicates, e.g., the explicative genitive is traced back to a relation of signification ("Bedeuten-Verhältnis") and the 'genitivus auctoris' is based on a relation of creation ("Verhältnis des Schaffens"):

(5) der Strahl der Hoffnung ← der Strahl bedeutet Hoffnung
the ray of hope ← the ray signifies hope

(6) das Werk des Dichters ← der Dichter schuf das Werk
the opus of the poet ← the poet created the opus

There has been much debate on the value of such classifications; Eisenberg [3] , for example, criticizes it as "purely descriptive" and "without explicative value". Because the classification relies on paraphrasing, it is indeed hard to reproduce. This is its major drawback. Nevertheless, the typology of Helbig/Buscha provides a detailed and widely accepted linguistic classification scheme, when it comes to classifying semantic relations between nouns in a more general sense (as, e.g., in SemEval 07 [5]), even though it is constrained to a subclass of relations between nouns.

Syntactic classification Lindauer [7] develops a formal classification for genitive attributes in the context of generative grammar. His goal is a classification based only on morphologic and syntactic criteria. He distinguishes *thematic genitives* from all other genitives. Lindauer's analysis results in three simple syntactic tests, which can be used to positively determine possessive genitives and to determine partitive genitives *ex negativo*. The tests are as follows:

1. Possessive genitives can be replaced by a possessive pronoun.
2. Possessive genitives can be replaced by a prepositional phrase headed by the preposition 'von'.
3. Possessive genitives can occur in prenominal positions.

Lindauer points out that the dependencies between the nouns in partitive constructions are rather unclear. This is reflected in the wide variation between appositive and attributive forms of partitive attributes and verb agreement alternating between the head noun and the dependent noun. For these reasons, partitives certainly have to be considered separately in syntax.

The Duden Grammatik [3] follows Lindauer's new formal classification insofar as it retains a very broad concept of possessives. The group of partitives, however, is subdivided by semantic criteria.

[3] Eisenberg, cited by Lindauer [7, p. 138]

3 Annotation Scheme

A classification which serves as an annotation scheme has to be exhaustive, its classes have to be as selective as possible, it should agree with syntactic analyses and be semantically adequate. Lindauer's purely formal classification is a good starting point for such an approach but it is semantically too coarse and neither compatible with the most important syntactic analyses nor does it fit in a more general classification in the field of relation mining. We will therefore refine it using readily available noun features.

Deverbal nouns Genitive constructions involving deverbal nouns constitute a large fraction of genitive attributes. The genitives fill the argument slots of the deverbal noun which it inherits from the underlying verb. If there is only one attribute attached to the head noun, these constructions are ambiguous. Whether a certain construction is a subject genitive or an object genitive can only be determined through selectional restrictions and the context of a sentence. According to Lindauer, these functional genitives belong to the class of possessives. They certainly pass the tests mentioned above. But there are two arguments for a more fine grained distinction in these cases. First, it is desirable to avoid, if possible, abstract predicates in the semantic annotation. We would therefore prefer the predicate argument structure in 7 over the one in 8 for the noun phrase *Kolumbus Entdeckung Amerikas* 'Columbus' discovery of America' :

(7) COLUMBUS(x) ∧ AMERICA(y) ∧ DISCOVERY(x,y)

(8) COLUMBUS(x) ∧ AMERICA(y) ∧ DISCOVER(z) ∧ POSS(x,z) ∧ POSS(z,y)

Second, Lindauers analysis doesn't go together with a lexical-functional syntactic approach. In the German "f-structure bank"[4], genitive attributes are treated as grammatical functions. In Lexical Functional Grammar, a uniqueness constraint requires that a grammatical function can appear only once for a certain predicate. However, there is, at least for deverbal nouns, nothing special about two genitive attributes occurring in a German noun phrase. One way to solve the problem is to differentiate between left and right attributes, as Forst [4] did. However, we prefer to subdivide possessive genitives in these cases into subject genitives and adnominal genitives according to Chisarik and Payne's LFG work [2].

Relational nouns The semantic annotation should take into account that the grammatical analysis for relational nouns differs from that of other nouns. The favored predicate argument structure for a noun phrase like *Petras Schwester* 'Petra's sister' is different from the one for a noun phrase like *Petras Pferd* 'Petra's

horse'. The semantic relation is given by the head noun and is therefore directly available. Prototypical for these cases is the kinship relation. The genitive attribute can then be considered a true internal argument in the sense of Barker [1]:

(9) SISTER(x,PETRA)
HORSE(x) ∧ POSS(PETRA,x)

Partitive and possessive genitives Partitive genitives are typically defined semantically: They express part-whole-relations. This is the definition used by Duden and Helbig/Buscha. Lindauer, on the other hand, singles out partitives by the tests mentioned above. In his approach, the notion of partitives is very restricted, and does not cover classic part-whole relations as in *die Augen meiner Freundin* 'the eyes of my friend'. Constructions which mention institutions, as in *der Präsident des Bundesrats* 'the president of the federal council', pose problems for Lindauer's test. A closer look shows that in these cases the genitives express both possessive *and* partitive relations: Paraphrasing (in the style of Helbig/Buscha) gives us both *der Bundesrat hat einen Präsidenten* 'the federal council has a president' and *der Präsident ist ein Teil vom Bundesrat* 'the president is part of the federal council'. For compatibility with more general classifications of semantic relations I decided to tag these genitives separately.

Test Result	**Noun Feature**	**Relation**	**Smultron Label**
Positive	Deverbal	Subject/Adnominal	SUBJ/ADN
	Relational	Kinship	KNS
	Deadjectival	Property	PROP
	Body part	Partitive and Possessive	PP
	–	Possessive	POSS
Inconclusive	Institutions	Partitive and Possessive	PP
	Collocations	Adverbial/–	AVG/–
Negative	Quantity	Partitive	PRT
	Abstract concept	Explicative	EXP

Table 1: Proposed classification scheme and mapping to Smultron labels.

Classification My classification, summarized in Table 1, concentrates on Lindauer's first syntactic test, which states that a possessive genitive may be replaced by a possessive pronoun. The class of possessive genitives (in Lindauer's syntactic sense) is subdivided into six categories depending on whether the head noun is deverbal, deadjectival, or relational. A special category is assigned to nouns denoting body parts. In all other cases, the syntactically possessive genitives are also possessives in the semantic sense. The class of syntactically partitive genitives

is subdivided into real partitives – where the head noun denotes a quantity – and explicative genitives in all other cases. There are some cases for which the test cannot give a definitive answer, namely if an expression involves institutions and in the case of collocations.

4 Annotation in Smultron

I have annotated 392 occurences of genitive attributes in the Smultron corpus. Smultron [8] is a parallel treebank consisting of around 1000 German sentences along with parallel Swedish and English texts, extracted from Jostein Gaarder's novel *Sophie's World* (the literary part) and from company annual reports (the economy part). The German syntax trees are annotated according to the TIGER guidelines. My annotation replaces the edge label 'AG' ("genitive attribute") with one of the ten semantic labels listed above. Interestingly, the use of genitive attributes heavily depends on the text type: There are only 0.10 occurrences per sentence in the literary part, compared to 0.65 occurrences per sentence for the economy part. Sentences containing three and more genitive attributes are common in the latter. First parallel annotations by three annotators showed an inter-annotator

Part	SUBJ	ADN	PROP	KNS	POSS	PP	PRT	EXP	AVG	–
Literary	2	3	0	4	12	0	18	4	8	2
Economy	42	82	0	0	77	60	36	13	19	10

Table 2: Distribution of genitive relations in Smultron

agreement of 88 percent, which we consider a very promising result.

5 Conclusions and Future Work

A closer inspection of the various genitive constructions revealed that their semantic classification directly depends on easily determinable morphologic features of the involved nouns. A combination of Lindauer's replacement test and morphologic analysis of the involved nouns thus enables a detailed semantic classification that has the advantage of being both based on transparent criteria and semantically adequate. A first interesting insight from the annotation of the Smultron Treebank is the surprisingly clear correlation between text type and frequency of genitive attributes.

To get a better idea of the applicability and the quality of the classification scheme described in this paper, it is planned to have other annotators apply it to

the same corpus. It is also planned to analyze the TIGER Treebank. The larger amount of data will enable subsequent experiments on automatic annotation using morphologic analysis of nouns and machine learning methods for classification. Corresponding annotation of the English and Swedish subcorpora of the Smultron Treebank are planned as well.

Acknowledgements I would like to thank Michael Piotrowski, Alexandra Bänzli and Étienne Ailloud. The research presented here was funded by the Swiss National Science Foundation (SNSF).

References

[1] Chris Barker. *Possessive Descriptions*. CSLI Publications, Stanford, 1995.

[2] Erika Chisarik and John Payne. Modelling Possessor Constructions in LFG: English and Hungarian. In *Nominals: Inside and out*. CSLI Publications, 2003.

[3] Dudenredaktion. *Duden: Die Grammatik.* Dudenverlag (= Duden, Band 4), 2005.

[4] Martin Forst. Treebank Conversion. Creating a German f-structure bank from the TIGER Corpus. In *Proceedings of the LFG03 Conference*, 2003.

[5] Roxana Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In *Semantic Evaluation Workshop (SemEval) in conjunction with ACL*, 2007.

[6] Gerhard Helbig and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch fÃ¼r den Ausländerunterricht.* VEB Verlag Enzypklopädie Leipzig, 1984.

[7] Thomas Lindauer. *Genitivattribute*. PhD thesis, Universität Zürich, 1995.

[8] Martin Volk and Yvonne Samuelsson. Frame-Semantic Annotation on a Parallel Treebank. In *Proc. of Nodalida Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*, 2007.