Automated detection and annotation of term definitions in German text corpora

Angelika Storrer, Sandra Wellinghoff

University of Dortmund; Faculty of Cultural Studies
D- 44221 Dortmund
angelika.storrer@uni-dortmund.de, wellinghoff@hytex.info

Abstract

We describe an approach to automatically detect and annotate definitions for technical terms in German text corpora. This approach focuses on verbs that typically appear in definitions (= definitor verbs). We specify search patterns based on the valency frames of these definitor verbs and use them (1) to detect and delimit text segments containing definitions and (2) to annotate their main functional components: the definiendum (the term that is defined) and the definiens (meaning postulates for this term). On the basis of these annotations we aim at automatically extracting WordNet-style semantic relations that hold between the head nouns of the definiendum and the head nouns of the definiens. In this paper, we will describe our annotation scheme for definitions and report on two studies: (1) a pilot study that evaluates our definition extraction approach using a German corpus with manually annotated definitions as a gold standard. (2) A feasibility study that evaluates the possibility to extract hypernym, hyponym and holonym relations from these annotated definitions.

1. Goals and project framework

In our paper we will describe methods to automatically detect and annotate definitions for technical terms in written text corpora. These methods are developed within the framework of the project HyTex[1]. In this project we create and evaluate strategies for automatically generating hypertext views and link structures that support both the selective reading and browsing of technical documents.

Two strategies that we implement in this context are:
(1) We link occurrences of technical terms in the documents with a ranked list of text segments, in which these technical terms are explicitly defined. The ranking of the definitions is based on a typology of definition types and on the position of the definition in the document (cf. Beisswenger et al. 2002).
(2) We create glossary views, in which the technical terms are displayed within the context of related terms (near-synonyms, hyperonyms, antonyms etc.). With these glossary views, definitions for all related terms can be retrieved and displayed in a new window. The glossary views are generated on the basis of a WordNet-style semantic net, in which technical terms are represented by using XML Topic Maps (cf. Lenz/Storrer 2002).

In the first phase of our project, the hypertextualization strategies were implemented and tested using a corpus with German technical texts. However, the linguistic basis of these automatic processes – the annotated definitions and the semantic net – was developed by hand-coding. In the second phase of our project we want to automate these linguistic preprocessing steps by developing and evaluating two approaches:

(1) The *DefTagtiv* approach detects definitions of technical terms in documents and annotates its components according to the annotation scheme we developed in the first phase of our project. In the following section we will describe our annotation scheme for definitions, specify our detection strategies and compare our approach to related work. In section 3 we discuss the results of a pilot study in which we evaluate this detection method on a German text corpus with manually annotated definitions.
(2) The *NetExpander* approach uses these annotated definitions to extract WordNet style semantic relations that will be used to enrich and expand the semantic net developed in the first phase of our project. In section 4 we will explain the results of a feasibility study on the possibility to extract semantic relations from definitions annotated according to our scheme.

In the framework of the HyTex project, both approaches are important if we want to apply our hypertext linking strategies to arbitrary technical domains. Furthermore, the detection and annotation of definitions may be useful in the context of terminology work and in computational lexicography: the DefTagTiv approach could support the lexicographers in writing adequate sense definitions in terminology databases and digital dictionaries. In addition, one could establish links between entries in digital lexicons and their definitions occurring in corpus documents. This may support the dictionary users in getting a better understanding of how terms are used in different schools or approaches of a scientific domain.

2. Annotation scheme for definitions and guidelines of the extraction approach

The main purpose of a definition is to explicitly ascertain the meaning in which a word is used in a technical or scientific document. Definitions typically consist of three functional components: the *Definiendum* (the term to be defined), the *Definiens* (meaning postulates for the term)

[1] The project HyTex (Hypertextualization based on textgrammatical annotations, cf. www.hytex.info.) is part of the research group "Texttechnologische Informationsmodellierung" funded by the German Science Foundation (DFG).

and the *Definitor* (the verb which relates the definiens component to the definiendum component). In the first phase of our project, we developed an annotation scheme with specific mark-up for these three main components. This annotation scheme may be illustrated by the following example:

Example (1):
Software, die dem Nutzer Orientierungs- und Navigationswerkzeuge für die interaktiven hypertextspezifischen Rezeptionsformen bereitstellt, bezeichnet man als Browser.
(Engl.: *Software that provides the user with orientation and navigation tools for the interactive hypertext-specific forms of reception is called a browser*).

The three main components of this definition will be annotated according to this schema in the following way:

```
<defSegment>
<def>
   <definiens>
   Software, die dem Nutzer Orientierungs- und
   Navigationswerkzeuge für die interaktiven
   hypertextspezifischen Rezeptionsformen bereitstellt
   </definiens>
   <dfnSegment> bezeichnet </dfnsegment>
   man
   <dfnSegment> als </dfnsegment>
   <definiendum> Browser </definiendum>.
</def>
</defSegment>
```

In the first project phase we manually annotated definitions of a test corpus according to this scheme. This corpus comprises 20 technical documents (103.805 words) from the domains of text technology and hypertext research. We developed an annotation guideline in which we specified the characteristic properties and patterns of definition segments. This guideline concentrates on definition patterns that correspond to the Aristotelian definition schema of *genus proximum* and *differentiae specificae* in their definiens. These definitions correspond fairly well to *formal definitions* in the typology of Trimble (1985,74f) und Flowerdew (1992, 209f). On the basis of this guideline we manually annotated 174 definitions in the corpus according to our scheme. These manually annotated definitions are used (1) as the gold standard in the pilot study on definition detection (cf. section 4), and (2) as the empirical basis in our feasibility study on extracting semantic relations from definitions (cf. section 5).

We tested our guideline on interoperability with two students. This study revealed that, in some cases, one needs to be familiar with the domains of text technology and hypertext research in order to decide whether a term occurs as a definiendum or is just used in a more general way. Since the annotated definitions are used as the gold standard in our evaluation study, we needed to obtain a reliable basis. Thus, the final decision whether a text segment was annotated as a definition or not was made by a researcher familiar to the domains of text technology and hypertext research.

3. Related Work

In comparison to other approaches on finding definitions (e.g. Saggion 2004, Klavans/Muresan 2001, Muresam/Klavans 2002), our understanding of "definition" and “term” is more narrow: technical terms in our approach are linguistic expressions, the technical meaning of which is explicitly defined in our corpus. The term “definition” is used to refer to text segments that contain the three main structural components of the definition schema: definiendum, definens, and definitor. Different definition patterns with these components are specified in our annotation guideline.

Definition detection approaches developed in the context of question-answering-tasks (e.g. Saggion 2004) are definiendum-centered, i.e. they search for definitions with a given term. Our approach, in contrast, is definitor-centered, i.e. we search for verbs that typically appear in definitions with the aim of finding the complete list of all definitions in a corpus independently of the defined terms. In our search patterns we define valency frames for such characteristic verbs like "bezeichnen als" (= to refer to as), "definieren" (= to define), "verstehen unter" (= to mean by). These frames specifiy the syntactic slots for the definiens and the definiendum components. In our project framework this approach has several benefits: (1) It facilitates the elimination of polysemous occurences of these verbs (such as “jmd verstehen” in the sense of “to understand s.o.”). (2) It is a good basis for annotating the internal structure of the definition in order to automatically extract semantic relations. (3) It helps to cope with variable word order in German sentences (a definiendum slot may occur on different positions).

In our approach definitions without a definitor-verb, as in example (2), have to be treated as special cases:

Example (2):
Homepage dtsch. Leitseite . Eingangs- oder Startseite eines Hypertext-Clusters.
(Engl: *Homepage Germ. Homepage . Introductory or starting page of a hypertext cluster*)

Such examples are typical for glossary or dictionary entries, but we also found them in ordered and unordered lists of our corpus documents. In these types of definition (henceforth called "glossary definitions") the components always appear in the same order: the definiendum component in the first position, followed by different types of separators, followed by the definiens component. Glossary definitions are hard to detect with our definitor-based method, since a multitude of different separators exists. In some cases there is no separator at all; instead, the definiendum is separated from the definiens by a different font or a different type face.

Definitions occurring in glossaries and online dictionaries are the primary source of the *Google* Glossary search function[2] which offers possibilities to display definitions to search terms. By contrast, the focus of our approach is on definitions that occur in the text body of technical and scientific documents.

[2] Cf. http://www.googleguide.com/glossary.html .

4. Pilot study on detecting definitions

In a pilot study, we evaluated our definitor-centered approach using the Insight Discoverer™ Extractor from the TEMIS Group.[3] This information extraction technology allows one to specify search and extraction patterns on different levels of analysis in so-called Skill-Cartridges™. With this technology we defined general concepts for the main components of our definition analysis – definiendum, definitor, definiens – and then specified for each definitor the valency slot for the definiens as well as for the definiendum. Additional constructs are introduced to cope with German word order alternatives.

We specified frames for 19 definitors (see table 1) in a Skill-Cartridge™. We applied these extraction patterns to our corpus and evaluated precision and recall using the definitions that we had manually coded according to the guidelines developed in the first project phase (see above). Table 1 shows the results of this evaluation. The figures in parenthesis (behind the definitor verbs) correspond to the number of definitions that we found in our corpus.

definitor	precision	recall
sein (80)	31%	83%
bezeichnen als (16)	43%	75%
verstehen unter (13)	100%	85%
nennen (10)	100%	20%
bestehen aus (7)	41%	100%
spezifizieren als (4)	100%	100%
heißen (3)	50%	100%
verwenden als (3)	9%	100%
bedeuten (2)	11%	100%
beschreiben (2)	33%	100%
begreifen als (1)	100%	100%
benennen (1)	100%	100%
charakterisieren als (1)	100%	100%
definieren als (1)	100%	100%
gebrauchen (1)	50%	100%
sprechen von (1)	50%	100%
Terminus einführen (1)	100%	100%
vorstellen als (1)	100%	100%
bekannt als (1)	50%	100%
total	**34%**	**70%**

Table 1: Results of the evaluation study

The results show, that precision and recall are both highly dependent on the definitor. Recall values are considerably high when the definiens occurs with a characteristic preposition that is specified in the valency frame of the definitor (as in "verstehen unter" or "spezifizieren als"). For the definitors that occur only once, the values are not significant and have to be evaluated with larger corpora. The precision value for the definitor "sein" (= to be) is especially problematic. Although the part of speech tag set used in the Insight Discoverer™ Extractor allows one to differentiate between the possessive pronoun "sein", the main verb "sein", and the auxiliary "sein", there are still many examples that satisfy all characteristic properties but that, nevertheless, are not definitions. Such an example is:

(3) Visualisierung ist eine gute Möglichkeit, Anwenderprobleme bei der Suche abzumildern.
(Engl.: *Visualization is a good possibility for downplaying user problems in a search*.)

Since "sein" is the most common definitor, the problems associated with this definitor have an impact on all of the recall and precision values. In addition, we do not yet account for glossary definitions in our search patterns for the reasons explained in the previous section. Since 25 definitions in our corpus are glossary definitions, this has, of course, negative effects on the evaluation of recall.

5. Feasabilty study on extracting semantic relations from annotated definitions

In the NetExpander approach we want to exploit our annotated definitions for an automatic extension of our semantic net. This net uses an extended inventory of the semantic relations that are specified in the Princeton WordNet (Fellbaum 1996). The idea of the NetExpander approach is that we use the annotated definitions and additional pattern matching rules to extract semantic relations that occur between the head noun of the definiendum and the head noun of the definiens. If the definiens follows the classical scheme "genus proximum + differencia specifica", the following extraction rule should apply: the definiendum head noun is a hyponym (a subclass) of the definiens head noun. An example that confirms this rule is our definition example (1) (cf. section 2). In this definition the head noun of the definiendum "Browser" is a hyponym of the head noun of the definens "Software" which is the hypernym (the superclass). In our feasibility study we checked for all definitions, whether this rule can be applied and whether there are other types of relations that may be systematically extracted.

The results of this study are shown in Table 2. It is obvious that in a considerable number of cases, the rule that the head noun of the definiendum is a hyperonym of the head noun of the definiens proved to be valid. This encouraging result shows that it is worthwhile not only to detect definitions but also to annotate their internal structure. However, the study also revealed that there are exceptions to this rule. In one of these exceptions, the relation between the definiendum and the definiens seems to be conversely specified. An example is definition (4):

(4) XML ist der **Oberbegriff** für die Regeln, die beim Definieren von Datenformaten angewendet werden.
(Engl.: *XML is the **generic term** for the rules that are used when defining data formats*.)

In this example, the definiendum XML is explicitly stated as being a hypernym (a superclass) of the subclasses denoted in the definens: the German word "Oberbegriff" (= generic term) is synonymous to "hypernym". In most of the definitions in which the relation between the definiendum and the definiens is conversely defined, we find such characteristic head nouns ("Oberbegriff" or "Klasse" (= class) in the definiens. In fact, if our main rule

[3] For our study we used the Insight Discoverer™ Extractor Version 2.1. (cf. http://www.temis-group.com/). We thank the TEMIS group for kindly permitting us to use this technology in the framework of our project.

is correctly applied to these cases, the definition states that the definiendum "XML" is a hyponym (a subclass) of the superclass "Oberbegriff". Since such relations on a meta-linguistic level are not relevant for our semantic net, we extract the converse relation between the superclass in the definiendum (in our example "XML") and the subclass in the head noun of the prepositional modifier (in our example "Regeln").

definitor	total number of definitions	definiendum is the **subclass**	definiendum is the **superclass**	definiendum is **holonym (whole)**	others / unclear
bedeuten	2	2	-	-	-
begreifen	1	1	-	-	-
bekannt als	1	1	-	-	-
benennen	1	1	-	-	-
beschreiben	2	1	1	-	-
bestehen aus	7	-	-	7	-
bezeichnen als	16	13	1	-	2
charakterisieren als	1	1	-	-	-
definieren als	1	1	-	-	-
einführen (Terminus)	1	1	-	-	-
gebrauchen	1	1	-	-	-
heißen	3	1	-	-	2
nennen	10	7	-	-	3
sein	80	73	5	-	2
spezifizieren als	4	4	-	-	-
sprechen von	1	1	-	-	-
verstehen unter	13	11	1	1	-
verwenden als	3	3	-	-	-
vorstellen als	1	1	-	-	-
glossary definitions	25	17	1	-	7
total	**174**	**141**	**10**	**7**	**16**

Table 2: Results of the feasibility study

Some definitors indicate part-whole relations rather than superclass-subclass relations. The head noun of the definiendum that occurs in definitions with the definitor "bestehen aus (consist of)" is typically a holonym (the whole) of the head noun of the definiens. An example is definition (5):

(5) Links **bestehen aus** einem oder mehreren Ankern, die in Ressourcen verankert sind.
(Engl.: *Links **are made up** of one or more anchors, which are fixed to resources*)

But we also found cases with other definitors that indicate part-whole relations. An example is definition (6):

(6) Unter einem Datenformat versteht man **die Gesamtheit** der Richtlinien, die für jedes Dokument dieses Typs gelten.
(Engl.: *Under 'data format' one understands **the totality of** principles that hold for every document of this type*.)

These cases are regularly indicated by characteristic head nouns in the definiens like "Gesamtheit" (= totality). Typical for the converse relation of meronymy are nouns like "Teil" (= part) or "Bestandteil" (= component). It should, thus, be feasible to sort out these exceptions from the general hyponymy rule and determine holonymy or meronymy correctly.

6. Further work and outlook

Future work aims to improve precision and recall of the extraction patterns in the following way: we want to evaluate and optimize the search patterns specified for our definitors on the basis of a large German text corpus, namely the DWDS core corpus[4]. In this step we will pay special attention to "booster words" and typical constructions indicating that polysemeous verbs (like "sein" or "bedeuten") are used as definitors. Furthermore, we want to enhance our list of definitors by examining more documents from various technical and scientific domains. This will be done semi-automatically using TEMIS extraction technology. In order to improve our recall values, we want to include search patterns for glossary definitions occuring in the text body of technical and scientific documents. Since the results of our feasability study on the extraction of WordNet-style relations are quite encouraging, we want to implement these extraction methods and evaluate them using definitions in corpora of different scientific domains.

7. References

Beißwenger, M, Lenz, E., Storrer, A. (2002): Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen. In: Busemann, S. (ed.): *Proceedings of KONVENS 2002*.

Fellbaum, C. (ed.) (1998): WORDNET: an electronic lexical database. London.

Flowerdew, J. (1992). Definitions in Science Lectures. In: *Applied Linguistics* (13)2, 202-221.

Geyken, A. (in press): The DWDS corpus: a reference corpus for the German language of the 20th century. Fellbaum, C. (ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press.

Klavans, J., Muresan, S. (2001). Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text. In: Proceedings of the 1st JCDL 2001. Roanoke, USA.

Lenz, E., Storrer, A. (2002): *Converting a corpus into a hypertext: An approach using XML topic maps and XSLT.* In: Proceedings of the Language Resources and Evaluation Conference. (LREC 2002).

Muresan, S., Klavans, J. (2002): A Method for Automatically Building and Evaluating Dictionary Resources. In: Proceedings of the Language Resources and Evaluation Conference. (LREC 2002).

Saggion, H. (2004): Identifying Definitions in Text Collections for Question Answering. In: Proceedings of the Language Resources and Evaluation Conference. (LREC 2004).

Trimble, J. (1985): English for Science and Technology. A Discourse Approach. Cambridge University Press.

[4] This corpus was constructed at the Berlin-Brandenburg Academy of Sciences (BBAW); it comprises more than 100 million word tokens balanced chronologically and by text genre (cf. Geyken (in press)). The corpus is POS-tagged, lemmatized and can be queried online via a linguistic search engine: http://www.dwds-corpus.de .