Finding General-Language Definitions in Corpora: Conceptual Design and Annotation

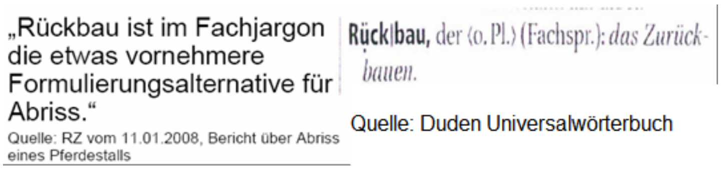
Irene Cramer

Abstract. Definitions mentioned in general-language text documents can be used as supplements to dictionary entries and thus help users to understand the meaning of general-language words or phrases. In this paper we present studies conducted in order to shape a conceptual design of what types of text segments are to be considered. Since it is a time-consuming and demanding task to find definitions in corpora, we constructed annotation guidelines and accordingly annotated a corpus sample, which can be used to develop an extraction system.

1 Motivation

Most of us use dictionaries when we come across words or phrases which we do not know or understand (i.e. in specific context). But what about our understanding of words or phrases not mentioned in these dictionaries, such as Ballonleuchte (Engl. powermoon)[1]? And what happens if we simply do not get on with the definitions given? There are (many) definitional segments in (general language) corpora compiled by language researchers and computational linguists, and, potentially, these might even provide a better support to grasp the meaning of a word or phrase than (some) dictionary entries do (in some cases). Unfortunately, finding helpful and/or interesting segments is a tedious and time-consuming work when performed manually. In recent years, researchers have started to develop methods for the automatic extraction of definitions from corpora (Walter and Pinkal (2006); Storrer and Wellinghoff (2006); Westerhout and Monachesi (2007)). However, to our knowledge all of these publications focus on specialized text. In contrast, in this paper we address the extraction and annotation of general-language definitions in general-language documents. In order to illustrate the application scenario as well as the purpose of our work, Figure 1 opposes a dictionary entry and a definitional text segment both dealing with Rückbau (Engl. renaturation or renaturization). As a matter of course, we could criticize the quality of the given dictionary entry, but even (almost) perfect ones might miss aspects in which a certain user is especially interested. We therefore regard the definitional text segments as being a supplement (rather than

[1] Cp. http://www.sfs.uni-tuebingen.de/ lothar/nw/Archiv/Datum/d080204.html



„Rückbau ist im Fachjargon die etwas vornehmere Formulierungsalternative für Abriss.“

Quelle: RZ vom 11.01.2008, Bericht über Abriss eines Pferdestalls

Rück|bau, der ⟨o. Pl.⟩ (Fachspr.): *das Zurückbauen.*

Quelle: Duden Universalwörterbuch

Figure 1. Dictionary Entry vs. Definitional Text Segment (Engl. `Rückbau` (renaturation or renaturization), in lingo, is the more elegant way of putting `Abriss` (demolition). Source: Rhein-Zeitung (Newspaper) on 01/22/08; report on the demolition of a horse stable)

being a substitution). In this paper we present two aspects of our work on extracting general-language definitions: Firstly, we describe two studies which helped us to understand what types of segments were judged as being relevant and valuable by potential users. Secondly, we developed annotation guidelines and started to find and mark-up definitional segments accordingly in two corpus samples.

Paper plan: The remainder of this paper is structured as follows: In Section 2, we describe two studies we conducted in order to determine (in a user-driven manner) the types and main features of general-language definitions. In Section 3, we summarize the annotation criteria which we (among others) derived from these studies and present an annotation case study as well as some quantitative aspects. In Section 4 we outline the plans which will guide our future work.

2 User Requirements

In order to identify the requirements implied by potential users, two questions have to be examined:

1. Which types of information do users retrieve? Or, in other words, what do users want to know?

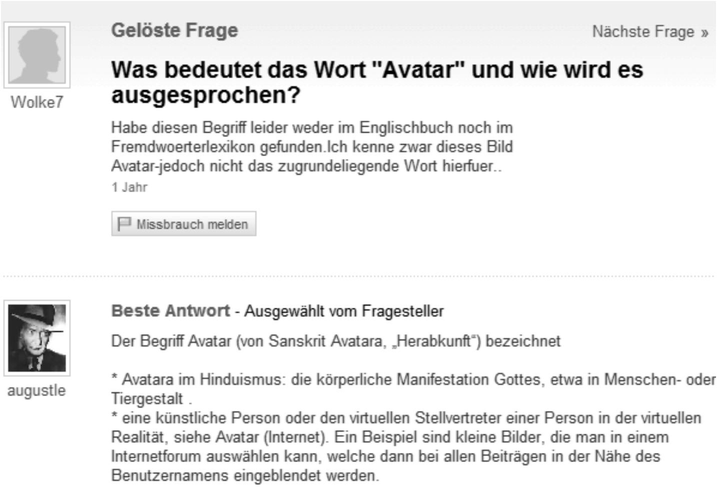
2. Which (types of) text segments satisfy their information requirements? Or, in other words, what can count as a definition from a user's perspective?

Both aspects can in principle be approached in a theoretical and in an empirical way. Unfortunately, only few researchers have investigated general-language definitions in corpora up to date, therefore, the theoretical basis is not yet of sound standing[2]. In

[2] Certainly, the theoretical basis is sound standing from a lexicographer's point of view, however, we do not investigate dictionary entries and, thus, the types of definitions usually used in dictionaries but we are rather interested in definitional text segments mentioned in corpora.

this paper we focus on the empirical method since this also provides an interesting insight into the underlying application scenario.

In order to answer the above mentioned questions we analyzed a randomly selected sample of 90 question-answer-pairs extracted from `Yahoo!Clever`[3]. We only considered questions asking for a definition or explanation of a word/phrase and respective answers given by non-experts. An example is shown in Figure 2.



Gelöste Frage

Was bedeutet das Wort "Avatar" und wie wird es ausgesprochen?

Habe diesen Begriff leider weder im Englischbuch noch im Fremdwoerterlexikon gefunden.Ich kenne zwar dieses Bild Avatar-jedoch nicht das zugrundeliegende Wort hierfuer..

Beste Antwort - Ausgewählt vom Fragesteller

Der Begriff Avatar (von Sanskrit Avatara, „Herabkunft") bezeichnet

* Avatara im Hinduismus: die körperliche Manifestation Gottes, etwa in Menschen- oder Tiergestalt .
* eine künstliche Person oder den virtuellen Stellvertreter einer Person in der virtuellen Realität, siehe Avatar (Internet). Ein Beispiel sind kleine Bilder, die man in einem Internetforum auswählen kann, welche dann bei allen Beiträgen in der Nähe des Benutzernamens eingeblendet werden.

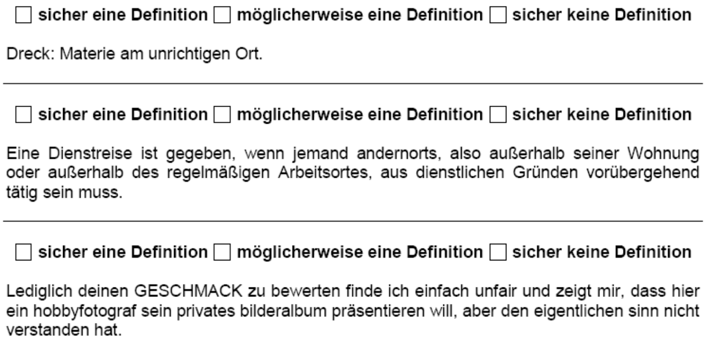
Figure 2. Examples of `Yahoo!Clever` Question-Answer-Pairs (**Question:** What does the word avatar mean and how is it pronounced? **Answer:** The term avatar (derived from Avatara meaning origin/parentage) denotes ...)

In addition, we asked approximately 50 subjects to fill in a questionnaire. The questionnaire contained a short instruction and 20 text segments. The subjects were asked to specify for each text segment whether (in their opinion) it consists of a definition or not. An example is shown in Figure 3. All text segments were randomly selected from a sample of the Wacky corpus (Baroni and Bernardini 2006).

2.1 Types of Information

The 90 question-answer-pairs were selected as a resource for our study because the `Yahoo!Clever` setting resembles the one we aim at. In both cases the users are willing to rely on the linguistic competence of non-experts and do not necessarily

[3] Cp. http://de.answers.yahoo.com/



□ **sicher eine Definition** □ **möglicherweise eine Definition** □ **sicher keine Definition**

Dreck: Materie am unrichtigen Ort.

□ **sicher eine Definition** □ **möglicherweise eine Definition** □ **sicher keine Definition**

Eine Dienstreise ist gegeben, wenn jemand andernorts, also außerhalb seiner Wohnung oder außerhalb des regelmäßigen Arbeitsortes, aus dienstlichen Gründen vorübergehend tätig sein muss.

□ **sicher eine Definition** □ **möglicherweise eine Definition** □ **sicher keine Definition**

Lediglich deinen GESCHMACK zu bewerten finde ich einfach unfair und zeigt mir, dass hier ein hobbyfotograf sein privates bilderalbum präsentieren will, aber den eigentlichen sinn nicht verstanden hat.

Figure 3. Extract of the Questionnaire (Options to be ticked: certainly a definition; possibly a definition; certainly not a definition. Extracts to be assessed: a) filth: material in the wrong place b) it is called business trip, if someone needs to be preliminary occupied in a different location, i.e. outside of his habitation or outside of his regular working place, for official reasons c) I regard it to be unfair to judge your taste only and it demonstrates to me that this is an amateur photographer intending to present his private photo album here, but who has not grasped the actual sense)

Table 1. Categorization of `Yahoo!Clever` Questions (Nouns: 10%; Verbs: 10%; Adjectives: 45%)

Category (Need of ...)	**Approx. Proportion**
rationale or explanation	34.1%
example or illustration	9.9%
specific subject area	29.7%
contrasting term / juxtaposition	5.5%
short form	33.0%
failed understanding / no entry in dictionary	5.5 %

search for information about terminology but are often interested in general-language words/phrases. Therefore, we examined the 90 question-answer-pairs with respect to several aspects. Firstly, the questions were analyzed in order to determine the users' information requirements. The results are summarized in Table 1. Secondly, the answers were explored in order to find the typical structures used by non-experts to define (or explain) words/phrases. The results of this analysis are discussed in Section 2.2. Obviously, in almost two-thirds of the cases the questioners give reasons for their query or state the situation/context of the word/phrase usage. Probably, they thus want to ensure that the answers given really meet their information need. In contrast, one-third of the questions are short forms such as: `Definition devot`

Table 2. Principle Definition Types

Type of Definition	**Approx. Proportion**
genus proximum + differentia specifica	18.1%
lexical / lexical semantic localization	20.1%
operational definition / genetic definition	21.2%
naming of a synonym, translation, class, example	19.4%
	∑ 78.7%

(Engl. Definition (of the word) devot) or `Was bedeutet devot` (Engl. What does (the word) devot mean?), containing no further specification of the precise information requirement. Interestingly, in approximately 5% of the cases the questioners explicitly indicate that they were not able to find the word/phrase in a dictionary or that they could not understand the dictionary entry.

Certainly, these observations have to be compared with the scholarly discussion in lexicography. This comparison will hopefully give us an idea of the information requirements on the one hand of users in a setting like `Yahoo!Clever`, which is–as mentioned above–similar to our application scenario, and typical users of dictionaries on the other. We are planning on dedicated future work to such comparisons.

2.2 Types of Text Segments

Two resources were used to determine the principle types of definition: the above mentioned questionnaire and again the 90 `Yahoo!Clever` question-answer-pairs. I.e. with between 7 and 8 answers per question, the 90 `Yahoo!Clever` entries contained 670 answers. The analysis of the questionnaires showed that especially text segments featuring one of three characteristic and basic structure types, namely, *genus proximus et differentia specifica*, itemizations of constituents or examples, and *if-and-only-if-variants*, were most consistently judged as definitions. The analysis of the above mentioned 670 `Yahoo!Clever` answers confirmed this observation. The results of this analysis are summarized in Table 2. Especially, given the analysis of the 670 `Yahoo!Clever` answers, we assume that there are four basic definition types, most likely (in 80% of the cases one of these four types can be observed, cp. Table 2) to be used in definitional text segments and recognized as definitions. Examples of the four basic types are shown in Table 3. This result is, in our opinion, interesting and relevant at the same time since almost all researchers investigating definitions from a corpus-based or computational linguistics point of view have until now considered the genus-et-differentia type as being the most prominent and therefore (almost exclusively) focussed on this type. Furthermore, other bits of information

Table 3. Principle Definition Types

Question	Answer
Was bedeutet das Wort Avatar und wie wird es ausgesprochen? [...]	**definition type: genus et differentia** Ein Avatar ist eine künstliche Person oder ein grafischer Stellvertreter einer echten Person in der virtuellen Welt, beispielsweise [...]
What does the word avatar mean and how is it pronounced?	An avatar is an artificial person or a graphic representative of a real person in the virtual world, e.g. [...]
Definition anlernen?	**definition type: lexical semantic localization** anlernen: anleiten, einweisen, unterweisen, einführen, kurzzeitige Unterweisung.
Definition of to show so. the ropes	instruct, admit so. to sth. , train, introduce, short-term instructions
Was heisst magensaftresistent? also also was heisst es z.B. wenn auf einer tablettenpackung [...]	**definition type: conditional construction** wenn das medikament erst im darm wirken, also aufgenommen werden soll, weil da die die wirkstoffe hingehören oder weil du möglicherweise einen empfindlichen magen hast [...]
What does enteric-coated mean? What does it e.g. mean when it says enteric-coated on a package of pills?	If the medicine is supposed to take effect only in the bowels, because the agents are needed there–or to protect a sensitive stomach [...]
Was bedeutet das Wort DEVOT?	**definition type: naming synonym** Unterwürfig!
What does the word submissive mean?	Abject/obedient/servile

such as notes on the etymology or connotation as well as examples of the usage are consistently grouped around these basic types. That is to say, most of the answers use one of the four types as core of their definition and add more or less additional bits of information as extra. Thus, answers without one of the above mentioned cores only stating extras are extremely rare and in the setting of `Yahoo!Clever` only occurred when there already were answers by other users (these again featuring a core structure).

3 Conceptual Design

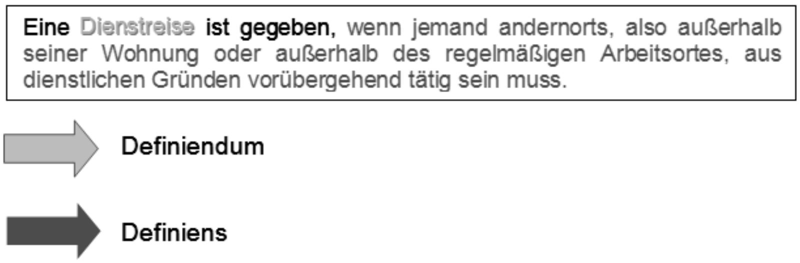
On the basis of these studies, among others, annotation guidelines were developed. In this section the main features and principles of the annotation are discussed (Section 3.1), and an annotation case study is presented (Section 3.2)[4].

3.1 Conceptual Design

Previous case studies with different annotation principles have showed that the inter-annotator agreement and confidence of the annotators with respect to their task is best if the following constraints are considered:

1. The annotation guidelines are organized according to the principle components of a definitional text segment: definiendum, definiens and definitor (cp. Beisswenger (2004); Wellinghoff (2006)).
2. The annotation guidelines state easily checkable–that means determinable without any particular linguistic knowledge–features of the three principle components.

In the following the conceptual design of our annotation guidelines is sketched. The complete guidelines will be freely available for download in autumn 2008. In order to illustrate the principle definitional components, in the example shown in Figure 4, definiendum, definiens, and definitor are marked. **Definiendum:** The definiendum represents the word or phrase that is to be defined. Some examples are shown in Table 4. In contrast to previous work[5], we consider nouns as well as verbs and adjectives. **Definiens:** The definiens represents the phrase or clause that defines or explains the



Eine Dienstreise ist gegeben, wenn jemand andernorts, also außerhalb seiner Wohnung oder außerhalb des regelmäßigen Arbeitsortes, aus dienstlichen Gründen vorübergehend tätig sein muss.

Figure 4. Principle Components of a Typical Definition

[4] ... which is work in progress.

[5] In the studies by Walter and Pinkal (2006) / Storrer and Wellinghoff (2006) / Westerhout and Monachesi (2007) only terminology was considered.

Table 4. Definiendum Types–Examples

Definiendum	**Definitional Text Segment**
Noun: Gesellschaft (society)	In der Regel bezeichnet der Begriff Gesellschaft die Gesamtheit der Menschen eines Landes bzw. einer Nation und ihre Beziehungen zueinander [...] Normally, the term society denotes the entirety of the people of a country or nation respectively as well as their relations to each other [...]
Adjektive: weise (wise)	Erfahren, besonnen, gescheit, verständig, wissend, gedankenversunken - all dies fassen wir unter dem Begriff weise zusammen. Experienced, considerate, clever, prudential, knowing, absorbed in thought–all of this we subsume under the term wise.
Verb: putzen (clean)	... putzen, auch reinigen, sauber machen, bedeutet in seiner allgemeinen Verwendung Schmutz entfernen [...] clean, also expurgate, cleanse, means–in its general use–to remove dirt [...]

definiendum. This is–at least from the annotator's point of view–the most critical component of definitions. We therefore assembled a list of semantic/informational structures potentially present in the definiens. An extract of the complete list of semantic structures is illustrated in Table 5. **Definitor:** The definitor, finally, represents the connector between definiendum and definiens. Some examples are shown in Table 6. In most cases, the definitor constitutes a verbal phrase. However, it also can be represented by e.g. typographic features. In order to facilitate the annotation, we also collected a list of potential pitfalls, such as anaphoric structures, short forms, inaccurate (or even incorrect) definitions, twist-definitions, and included in the guidelines a detailed discussion addressing all of these problematic cases, more than 50 mostly positive (and some negative) example definitions as well as an FAQ.

3.2 Annotation Case Study

In a subsequent annotation case study, three subjects were asked to annotate definitions; two samples of segments containing the most relevant definitor verbs were

Table 5. Definiens Structures–Examples

Type of Information	**Example**
Origin	[...] auch durch bestimmte Schmerzmittel selbst können unter Umständen Schmerzen ausgelöst werden (sog. medikamentinduzierter Kopfschmerz) [...] [...] even certain painkillers may under certain circumstances trigger pain themselves (so-called physic-induced headache) [...]
material, principle components, characteristics	... ein Schimmelpilz besteht aus einem Geflecht von Zellfäden, den sogenannten Hyphen [..] [...] a mold consists of a net of cell fibers, the so-called hyphae [...]

Table 6. Definitor Verbs–Examples

Definitor	**Example**	**Definitor**	**Example**
sein	[...] Rückbau ist im Fachjargon die etwas vornehmere Formulierungsvariante für Abriss [...]	nennen	[...] der Vorgang wird Elektrolyse genannt [...]
to be	Rückbau (renaturation / renaturization), in lingo, is the more elegant way of putting Abriss (demolition)	to call	[...] the procedure is called electrolysis [...]
bedeuten	[...] operativ bedeutet hierbei, dass [...]	bezeichnen (als)	[...] als cache bezeichnet man einen Speicherbereich, der [..]
to mean	in this context, functional means that [...]	to refer to as	[...] a memory area (buffer) which [...] is referred to as cache [...]

extracted, one from the DWDS[6] (Geyken 2007) and one from the Wacky corpus.

[6] We would like to thank Alexander Geyken for his help.

Table 7. Overview Results of Annotation Case Study (Wrt. Random Selection of Segments of Wacky Corpus Containing Approximately 200,000 Tokens)

Subject	**Time**	**not DEF**	**DEF**
subject 1	approx. 32 hours	approx. 2,500	approx. 500
subject 2	approx. 37 hours		
subject 3	approx. 37 hours		

Because of the disappointing yield of positive entries in the DWDS sample, we decided to concentrate on the Wacky corpus. We asked the subjects to log the amount of data annotated and the time needed. The results of this case study are summarized in Table 7. The inter-annotator agreement for this setting ranges between 65% and 80%[7]. The table illustrates that the subjects were able to annotate approximately 100 segments per hour. This very much stresses the need of automatic means to support the extraction of definitional text segments. We also found that in order to accurately annotate text segments with respect to definitions a rather substantial training period seems to be necessary. We currently experiment on the inter-annotator agreement and the training necessary in order to annotate high quality data with a group of 7-10 subjects.

4 Conclusions and Future Work

The annotation of definitions is a time-consuming and demanding task. However, in order to develop a tool able to automatically extract definitional text segments from corpora, annotated data is needed. We have summarized several studies which we conducted in order to determine the principle structures and types of definitions. On this basis, we developed annotation guidelines. At present, we are running a (compared to the above mentioned case study) more extensive inter-annotator agreement test and at the same time extending the amount of annotated data. The main objective of our research is to develop a tool able to automatically find and extract large amounts of definitional text segments, which can be used as supplements to dictionary entries. As our next step, we therefore plan to analyze the annotated data in view of discriminant features of definitional text segments. By this means, we hope to identify relevant syntactic and semantic structures, which can be utilized by the extraction system. We also plan to scrutinize the differences between dictionary definitions and the ones present in corpora of general-language documents.

[7] The sample was split into parts; every fraction was annotated by two of the three subjects.

References

Baroni, M. and S. Bernardini (eds.) (2006). *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Beisswenger, M. (2004). Annotation definitorischer Textsegmente und terminologiesensitives Linking. Technical report, University of Dortmund, Germany.

Geyken, A. (2007). The dwds corpus: a reference corpus for the german language of the 20th century. In Christiane Fellbaum (ed.), *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*, Continuum Press, London.

Storrer, A. and S. Wellinghoff (2006). Automated detection and annotation of term definitions in german text corpora. In *Proceedings of the LREC 2006*.

Walter, S. and M. Pinkal (2006). Automatic extraction of definitions from german court decisions. In *Proceedings of the COLING-ACL 2006 Workshop Information Extraction Beyond the Document*.

Wellinghoff, S. (2006). Manuelle Annotation definitorischer Textsegmente incl. Guidlines Phase I und II. Technical report, University of Dortmund, Germany.

Westerhout, E. and P. Monachesi (2007). Extraction of dutch definitory contexts for elearning purposes. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands*.