A linguistic bootstrapping approach to the extraction of term candidates from German text

Ulrich Heid

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart
heid@ims.uni-stuttgart.de

version December 1998 – Draft

Contents

This paper deals with computational linguistic tools and methods for the extraction of raw material for terminological glossaries from machine-readable text[1] . We concentrate on monolingual German term candidates, and only briefly hint at tools and procedures for the creation of bilingual glossaries.

Most of the examples we use to illustrate methods and results of our work come from technical texts provided by the translation services of DaimlerChrysler AG[2] and from legal texts made available by the European Academy in Bozen, Südtirol. The Academy is working on translations of legal documents for bilingual South Tyrol, and, in this context, on the creation, upgrading and maintenance of terminological resources[3].

1 Context and objectives

1.1 Tool context

One of the underlying assumptions of our work is that the identification of terminologically relevant linguistic material can make use of the same type of computational linguistic corpus processing tools as any other exercise in the retrieval of linguistic information, also for general language. Thus, many of the tools we use were not originally developed exclusively with terminology in mind, but rather as general purpose tools for computational linguistic corpus exploration: these include low-level corpus pre-processing, such as tokenization (to identify word and sentence boundaries), part-of-speech tagging and lemmatization (to assign word class and lemma tags to inflected forms), as well as partial parsing for the identification of relevant patterns in the text material under analysis. Such tools are needed as basic building blocks of most types of linguistically oriented text exploration systems. For German, other than for English, no corpus parsing tools are yet available that would provide wide enough coverage to be directly usable for term candidate identification; consequently, partial parsing by means of a pattern-matching-based approach still seems to be the most viable procedure for practically usable term extraction[4].

[1] It is in part the result of joint work with Arne FITSCHEN, Stefan EVERT and Judith ECKLE-KOHLER, whom I would like to thank very much for their contributions. The responsibility for any misconceptions or infelicities in this presentation is of course my own.

[2] The material comprises 3.2 million words from maintenance manuals for cars, the material basis for the current joint project CONTRANS, which, among others, aims at the preparation of a controlled terminology for the maintenance literature. CONTRANS involves DaimlerChrysler language services, DaimlerChrysler research centre (Ulm) and the IMS of Universität Stuttgart. The project is part of development work for DaimlerChrysler language services.

[3] We whould like to thank Johann GAMPER and Felix MAYER of the European Academy for making two law texts available to us, for experimentation, along with the terminological expertation horizon worked out by the European Academy. The laws concern public construction work and urbanism.

[4] Cf. [Heid et al. 1996]: the article focusses on the usability of standard corpus processing tools for term extraction. For English, existing chunk parsers may be used more efficiently than pattern-based approaches, for the identification of multiword terms. On the other hand, work by BOURIGAULT shows, at least for French, that a pattern-based approach can lead to good results (cf. [Bourigault *et al.* (1996)]). Note that BOURIGAULT identifies *boundaries* of multiword term candidates, rather than trying to identify the candidates themselves by means of patterns. As far as we can see from experiments with German, a boundary-based approach seems to be more effective for configurational than for non-configurational languages.

The actual terminological dimension of the extraction work comes into play when the extraction tools are parameterized in such a way as to provide as much relevant material and as little "noise" as possible; or, in terms of information retrieval: maximizing recall and precision is a matter of tuning the extraction tools to the specific needs of terminology; mostly to the needs of the terminology of the domain under analysis. The idea is to use a bootstrapping approach: first, identify single-word items relevant to the sublanguage under study, then reuse these candidates to find phraseological material, compounds, complex noun groups, etc.

1.2 Terminological working context

The acquisition of term candidates is relevant for several types of applications. An evident one is the computational treatment of specialized texts, such as machine translation, indexing and information retrieval or information extraction.

On the other hand, there is also a need for term extraction in a much more widespread scenario, namely that of human translation and technical writing, be it computationally supported or not. Most translators use some sort of glossary or terminological database. Many translation bureaux or translation departments in companies have terminological data collections. We are contemplating, in this article, situations where computational support is sought to construct, update, maintain or enhance such glossaries. Although this work may be carried out by terminologists in large companies and administrations, more often than not, time pressure and costs make it difficult to get the glossary building task done fully (or mainly) manually; thus, there is a pressing need for tool support. The type of term bank we wish to support is not extremely sophisticated; it contains, for each language, terms and their basic linguistic description: word class, grammatical usage specificities (e.g. plural-only use), collocations and contexts. We do not aim, here, at acquiring relational knowledge (e.g. on nominalizations vs. verbs or adjectives, as in the case of JACQUEMIN's work) or definitional knowledge (as with [Pearson (1998)]). See also below (section 1.3).

A precondition for the tools to work is the availability of machine-readable text; ideally, both source and target language texts should be available in electronic form; here, we concentrate on the creation of raw material for the German part of the glossaries, thus using essentially German texts.

A typical usage situation for our term extraction tools is the terminological exploration of completed translation jobs: before the jobs are archived, the relevant terminology is extracted and stored in a term bank. The same procedures may be used in the preparation of a larger translation job: then, the goal is to work through the source texts, in order to retain the terminology already available and, conversely, to identify whatever needs special terminological attention.

For the terminological exploration of translation jobs during archiving, translation memory systems tend to be used rather frequently; however, in our view, they are only a suboptimal solution. A translation memory will store any sentence pairs (source language plus target language) or any pairs of expressions, without any notion of terminological relevance. Over time, a translation memory may thus grow considerably, contain a lot of potentially relevant material, but as well a lot of terminologically uninteresting sentence pairs or word sequence pairs. Term identification and translation memory fulfil different purposes; term extraction may be used to provide terminologically relevant input for the actual dictionary function of a translation memory. And, vice versa, material from a translation memory can be used as a sentence-aligned bilingual corpus for further terminological exploration.

1.3 Objectives: the targeted glossaries

Many terminological glossaries are not very rich in linguistic information: they contain equivalent source and target language items, occasionally along with domain labels and indications of the source from where the terms have been gathered. In many cases, this structure is sufficient for the translator, who mainly wishes to verify his or her intuitive judgements.

However, slightly more usage information may be useful both for human users of a glossary, and for its use as a resource for Natural Language Processing; we suggest the inclusion of the following additional data categories for usage information:

- Source and target language sentence pairs, as illustrative examples;
- definitions or definitional elements;
- labels or decorations concerning the usage of the terms, for example by areas (e.g. Portugal vs. Brasil), by clients, by (sub-)domains (e.g. motor vs. gearbox), by text types (e.g. jurisprudence vs. laws) etc.

To cover a relevant portion of contextual and phraseological information, a glossary should not only contain single word terms, but also multiword terms. We adopt the point of view, here, that many of the categories of multiword lexemes in use in general language lexicography can be used as well in the description of terminological multiword expressions. So, in our view, next to single word terms[5] (see example 1), also noun-noun-collocations (usually nouns plus genitive (see example 2) or noun plus prepositional group), noun-adjective-collocations (see example 3) and noun-verb-collocations (see example 4) need to be entered into a terminological database.

We do not enter into a discussion, here, about the status of such units (should they be considered "terminological units" in their own right, or are they a sort of "decoration" of nominal terms? Where is the separating line between sublanguage collocations and multi-word terms? Do we need such a separation?): we set ourselves the goal to extract such items from text material. Their further sub-classification and structured storage is a topic outside the scope of this paper (see however [L'Homme (1998)] for a discussion of terminological collocations, also in comparison with general language collocations).

[5] We are aware that we adopt a very mechanistic view of "single word terms", here: a single word (form) is taken to be a string of characters between two blanks. For our work on German, we thus take noun compounds (written in one string) as "single word terms"; yet. We make use of information about components of compounds, wherever necessary. But in a first, broad approach, we use the tokenizing output as a definition of "single word" vs. "multiword" terms.

(1) *Zuschlag, Angebot, Bieter, Landesbauhof, Auftragsvergabe*

(2) *Erstellung eines Bauleitplanes,*
Erteilung einer (neuen) Konzession,
Führung eines (landwirtschaftlichen) Betriebes

(3) *bestehende Baumasse, geltende Fassung, öffentliche Arbeiten*

(4) *im Bauleitplan (etw.) festschreiben,*
(eine) Geldbuße verhängen,
(etw.) zur Nutznießung besitzen

Along with the kinds of data described above, a few more linguistic information types may be necessary for the creation of terminological glossaries for use in Natural Language Processing; some of this is rarely found in term banks or glossaries, so far, and we thus aim at extracting evidence from where to derive such information:

- word class information;
- information about the syntactic subcategorization of verbs, adjectives and nouns (e.g. *etwas auf etwas überprüfen*);
- information about preferred prepositional adjuncts used along with the items under description (e.g. *etwas als Sondermüll entsorgen*, or, in a chemical context: *etwas erwärmen + bei kleiner Flamme + auf ... Temperatur + unter Umrühren*, etc.).

2 Extraction methods and tools

In this section, we sketch the computational linguistic extraction methods and tools used to gather term candidates from German text material.

As the objective is to extract as much glossary-relevant material as possible, from a given text, and especially to extract not only single word terms, but also collocations and noun groups, we follow a stepwise procedure, similar to bootstrapping: we first identify relevant morphological components of term candidates and single word term candidates, then, in a second step, we extract collocations and other multiword term candidates. This procedure seams also most efficient from the point of view of the overall architecture (see also below, section 2.2.2).

The extraction tools combine linguistic procedures based on pattern matching via regular expressions with the well-known statistical approach of relative frequency comparison[6], used here as a device for the acquisition of morphological components of term candidates. The procedures are briefly summarized in the following.

[6]See e.g. [Ahmad *et al.* (1992)] for a description of the relative frequency based method of term candidate identification. The principle is simple: a specialized language text (to be searched for term candidates) and a general language text (corpus) are compared. Term candidates are usually comparatively more frequent in the sublanguage text than in the general language text. Thus, we build a word form or lemma list from the specialized text, and compare the relative frequency of each list item in the specialized text with the relative frequency in a general text. Items which have much higher relative frequency in the specialized text than expected from general language are taken to be term candidates.

2.1 The linguistic infrastructure

The extraction tools operate on pre-processed corpora which contain the results of tokenizing (word and sentence boundaries identified), word class annotation (POS-tagging) and lemmatization[7]. Figure 1 contains an annotated sentence (word forms, part-of-speech tags, lemmas) from one of the laws of the Land Südtirol (*Bauaufträge über zwei Milliarden Lire, ohne Mehrwertsteuer, werden außerdem an der Amtstafel der Landesregierung veröffentlicht.*).

```
<s>
Bauaufträge/NN/Bauauftrag
über/APPR/über
zwei/CARD/zwei Milliarden/NN/Milliarde Lire/NN/Lira ,/$,/,
ohne/APPR/ohne MwSt/NE/Mehrwertsteuer ,/$,/,
werden/VAFIN/werden außerdem/ADV/außerdem
an/APPR/an der/ART/d <Amtstafel/NN/Amtstafel>
der/ART/d Landesregierung/NN/Landesregierung
veröffentlicht/VVPP/veröffentlichen
./$./.
</s>
```

Figure 1: Sample annotated text: Word forms – category tags (POS) – lemmas

The retrieval tools operate on word forms, part-of-speech tags and lemma annotations, in any combination, making use of lexical data (e.g. lists of grammatical words) and of sequence information. They are implemented as regular expressions over character sequences, word forms and word (or lemma, or POS) sequences, processed by the CQP corpus query processor[8], usually in the form of query templates for batch processing[9].

The schema in figure 2 summarizes the two steps of pre-processing and query. The extraction is automatic, once the appropriate query templates have been designed.

[7] Other than for English (and similarly for Romance and Slavonic languages), lemmatization is of major importance for the computational treatment of German texts, to relate inflected forms to lemmas. We use a statistical tagger-lemmatizer based on the STTS tagset, an EAGLES-conformant part-of-speech tagset for German with 54 categorially and distributionally defined tags. Cf. [Schmid (1994a)], [Schmid (1994b)], [Teufel 1995], etc.

[8] Cf. [Christ (1994b)], [Christ/Schulze (1996)].

[9] The macroprocessor allows to automate the application of queries. Since query results can form subcorpora on which additional queries can operate, the queries can be cascaded, and the text corpus partitioned into subsets, according to certain types of phenomena.

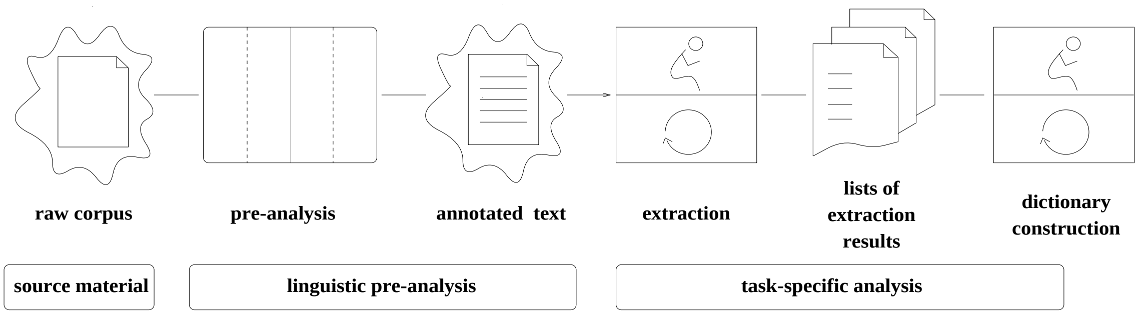

Figure 2: Simplified schema of pre-processing and pattern-matching based construction of raw material for a glossary

2.2 Procedures for term candidate extraction

2.2.1 Extraction of term candidates with regular patterns

The extraction routines used in our work operate at three different levels:

- Regular expressions over characters: for the identification of special word forms, abbreviations[10], etc.;
- regular expressions over morphemes for the identification of single word term candidates;
- regular expressions over sequences of positions, where positions are word forms and their lemma and/or part-of-speech annotations (POS-shapes), for the identification of multiword term candidates.

Certain technical terms have morphological properties which distinguish them from general language vocabulary. So, for example, [Reinhardt *et al.* (1992)] have identified a number of (mainly latinate) affixes which are more frequent in technical vocabulary than in general language. By extracting all words from a given text that contain any of these prefixes and/or suffixes, one may thus get access to a certain amount of term candidates. Examples of such affixes are given in (5) and (6), below.

(5) *ab.+*, *auf.+*, *ent.+*, *anti.+*, *bi.+*, *mega.+*, *mikro.+*, *multi.+*, *radial.+*, *semi.+*, *ad.+*, *ex.+*, *in.+*, *ko.+*, *pro.+*, [...]

(6) *.+grad*, *.+heit*, *.+nis*, *.+schaft*, *.+tum*, *.+ial*, *.+gramm*, *.+graph*, *.+id*, *.+ik*, *.+tion*, *.+tät*, *.+um*, [...], *.+ator*.

[10] Note that in many technical texts, abbreviations, pointers, names of objects given by means of identification numbers follow a strict "grammar", which can easily be modeled in regular expressions and then used for specific extraction purposes. For a general identification of these "non-lexical words", a set of simple regular expressions (identifying items containing special symbols ("-", "/", etc.), numbers, capitals within word forms, etc.) is sufficient. A more detailed "regular grammar" is used only where the amount of text to be processed and/or the need to identify certain types of "non-lexical words" justify this.

The examples (5) and (6) contain regular expressions[11] over morphemes; of course, not all elements in (5) and (6) are (derivational) morphemes in the strict sense, as not all term candidates are derivatives. We combine the search for derivatives with a search for compounds, and with a search for relevant mono-morphemic items.

Similarly, certain kinds of syntagmatic groups are more typical for technical vocabulary than for general language; examples are nouns followed by genitives or prepositional groups (with or without article), which may denote (complex) objects or procedures and thus may have term status.

However, the regular extraction patterns cover several types of word combinations which it may be useful to keep separate, at least conceptually: complex noun groups[12] (`N + N (genitive)`, `N + Prp (+ Det) + N`), adjective-noun-collocations (`A + N`) noun-verb-collocations, nouns followed by adverbs or uninflected adjectives (`N + A`, `N + Adv`), etc.

These word combinations are of different types:

- Sublanguage- or text type-specific vs. general: `N + A` and `N + Adv` are rather uncommon in general language and typical for certain types of technical prose only: we found many examples in the car maintenance manuals, to denote specific objects; these forms may be good candidates for a terminological glossary; whether they should be catalogued in this specific form (with the noun preceding the modifying adjective or adverb) has to be decided by the terminologist.

- Lexically determined vs. free: not all combinations of e.g. nouns and adjectives are terminologically relevant; if we extract all combinations of a certain POS-shape, additional filters are needed afterwards, to identify those combinations which are either collocational (i.e. lexically determined) for the sublanguage under analysis or otherwise relevant for the glossary, e.g. because they denote specific subtypes of an object, a device or a technical process.

Thus, the use of regular expressions over categorially annotated word forms and/or lemma instances is only a first step in the process of extraction of multiword term candidates; it produces considerable noise with respect to terminologists' expectations and thus requires the application of subsequent filters, e.g. against general language collocational lexicons. However, it is an indispensable intermediate step in the extraction process.

Typical examples of terminological adjective-noun collocations extracted by means of POS-shapes, from maintenance manuals for cars, are given in 7, below:

(7)	Fensterheber	elektrisch	375
	Fensterheber	mechanisch	10
	Getriebe	automatisch	1113
	Getriebe	mechanisch	293

[11] Notation: the dot (".") stands for an arbitrary character; we use Kleene's "+" and "*" symbols to denote "one or more" or "zero or more" repetitions of an element, respectively.

[12] We give in parentheses a "part-of-speech shape" of the items in question. Here "+" means "followed by". The actual query patterns may be more complex.

Leuchtweitenregulierung	automatisch	8
Leuchtweitenregulierung	elektrisch	58
Leuchtweitenregulierung	pneumatisch	32
Mantelrohr	starr	13
Mantelrohr	verstellbar	43

2.2.2 Combining regular patterns with statistical term candidate search

The three kinds of regular expressions described in the preceding section are formulated too loosely for being usable for the extraction of term candidates: they would extract any word form containing the right affix, or any sequence of words satisfying the criteria stated by the respective "POS-shape". To facilitate terminology extraction, the amount of noise in the result should however be as low as possible, with no or as little as possible silence[13].

In experiments with press releases from Daimler Benz, we found that the use of domain-specific morphemes (components of the stems of candidate terms) drastically reduces noise in the extraction results obtained for single word nouns (cf. [Heid et al. 1996]): on a text of 35.500 word forms, analyzed both automatically and, beforehand, manually (as a "gold standard" against which to compare the automatic results), the noise rate of the extraction routines for nouns went down from 13.5% (when using only prefixes) and 8% (when using only suffixes) to only 2% when using domain-specific morphemes. Combining suffixes (which proved more selective than prefixes) with domain-specific morphemes does not further improve precision, but may even have a negative impact on recall (over-determination of the queries). Typical examples of domain-specific morphemes for the car manufacturing press releases, are given in (8):

(8) *.*fahr.**, *.*motor.**, *.*trieb.**, *.*bau.**, *.*stoff.**, *.*elektr.**, *.*system.**, *.*auto.**, *.*techn.**, *.*filt.**, *.*kanal.**, *.*tank.**, *.*kabel.**, [...], *.*park.**.

Against these results, we have worked out our term extraction strategy as a bootstrapping exercise: we first need to acquire domain-specific morphemes, before doing single word and multiword term extraction. To do so, we use the method of relative frequency comparison, as well as a morphological compound analyzer.

The relative frequency comparison lists single word term candidates, on the basis of information about how much more frequent the candidates are in the specialized text than in a general language text. To provide satisfactory results for German, a few specific preprocessing steps are needed, before the relative frequency comparison can be applied (see below, section 2.3.2).

We use the statistically identified term candidates in two ways: as a first, still noisy, set of single word term candidates, and, secondly, as an input to the search with domain-specific morphemes.

[13] The practical trade-off is about the amount of human intervention (e.g. to remove noise) needed: the more noise, the more time-consuming is the task of "cleaning up" candidate lists. On the other hand, most translators would rather accept a small amount of noise than silence.

2.3 Acquiring domain-specific morphemes

2.3.1 A "morpheme collector"

To identify domain-specific components, a tool for the morphological analysis of base nouns and compounds (and, to some extent, derivatives) is used: all nouns, compound participles (e.g. *energieaufnehmend, lichtreflektierend, hochspannungsführend, verbleit*) and adjectives (e.g. *betriebswarm, blasenfrei*) are analyzed and their component morphemes are collected. Example (9) contains two sample analyses[14] of German compounds, for the nouns *Achsgetriebeöl* and *Ansauglufttemperatur*:

(9)
```
Achsgetriebeöl
1. Achs=Getriebe=Öl+NN.Neut.Akk.Sg
2. Achs=Getriebe=Öl+NN.Neut.Nom.Sg
3. Achs=Getriebe=Öl+NN.Neut.Dat.Sg

Ansauglufttemperatur
1. Ansaug=Luft=Temperatur+NN.Fem.Akk.Sg [...]
```

Those component morphemes which have been used most (threshold!) are used as domain-specific morphemes in the subsequent regular queries. The lists below, in (10) and (11), contain some of the more frequent domain-specific morphemes identified in the legal texts from Südtirol:

(10) *Zuschlag.*, Vertrag.*, Land.*, Bau.*, Unternehmen.*, Ausführung.*, Akkord.*, Wild.*, Wettbewerb.**

Ersatz., Bereitstellung.*, Beratung.*, Benützung.*, Ausschreibung.*, Ausschließlichkeit.*, Arbeit.*, Amt.*, Abnahme.**

(11) *.*preis, .*entwurf, .*vergütung, .*genehmigung, .*erklärung, .*empfänger, .*auftrag, .*stellen, .*recht, .*prozentuell, .*neunt,*

*.*maßnahme, .*linie, .*konzession, .*kasse, .*inhaber, .*wettbewerb, .*haushalt, .*gericht, .*erteilung, .*betrieb, .*betrag*

Many of the items in (10) and (11) appear outside compounds as well and are then also considered as term candidates (*Ausschreibung, Vergütung, Konzession*, etc.), but some may not themselves be term candidates for the (legal) texts we analyze, such as *Amt*, *Preis*, *Arbeit*, *Linie*; these latter ones enter however into terminologically relevant compounds. They can be removed from the term candidate lists by means of the relative frequency comparison.

To feed the regular extraction patterns for single word term candidates, we first extract potential domain-specific morphemes from a given text, by means of the morphological

[14] The "morpheme collector" has been developed by Arne FITSCHEN, on the basis of the two-level morphology system DMOR (cf. [Schiller (1994)]).

analysis of base forms, compounds and derivatives, and by listing the morphemes thus identified by order of frequency.

We can run this “morpheme collector” over the whole text, or we may use it only on a list of single word term candidates, identified by means of the relative frequency comparison. The morpheme collector does not take frequency of occurrence of individual word forms into account, but only collects (and counts) how many different types of compounds could be analyzed. Collecting morphemes only from single word term candidates identified through their frequency may in some cases lead to less noise[15].

The extraction of domain-specific morphemes has a further advantage for term candidate extraction: no German lexicon will be complete with respect to compounding and/or derivation in specialized texts; thus, it is hard to formulate any detailed lexical expectations for term candidates (e.g. which components show up as heads, which ones as modifiers, etc.); the same even holds for a tagging lexicon. An unspecific tool for breaking up compounds and a subsequent (again unspecific) regular query will however at least be able to find and identify all compounds contained in the text: for example, the participle *achsaufnehmend* is found due to its component *achs-* (from *Achse*).

2.3.2 Adapting the relative frequency comparison tools for German

For the extraction of single-word term candidates, [Ahmad *et al.* (1992)] propose procedures based on a comparison of the relative frequency of word forms in specialized and in general language texts. The underlying assumption is that a domain-specific text will feature those words as more frequent than expectable from general language which are most relevant for the topic of the text (see footnote (6) above, in section 2, for more details).

To run the tool in practice, a frequency count is performed for each word form from the specialized text, in both the specialized and a general language (or “domain-unspecific”) text (collection) used for comparison. For English, a corpus like BNC (the *British National Corpus*), with its broad coverage of different fields of knowledge and its 100 million occurrences, can be used as a general language comparison corpus.

For German, no such “balanced” corpus exists, and thus journalistic corpora needed to be used, in our experiments, although they might indeed blur the results in certain cases (e.g. the terminology of economy or sports).

Moreover, the figures obtained for German are less clearly interpretable than those produced for English, unless a number of preprocessing steps are carried out:

- Sentence-initial (and complete) capitalization has to be neutralized, since otherwise capitalized forms of non-nouns may be treated separately from “all-lowercase” forms, leading to noise in the candidate set.
- Inflected forms need to be lemmatized and relative frequency figures calculated on lemma frequencies rather than word form frequencies. The existence of several inflected forms tends to dilute the relative frequency figures; in addition, if run on word

[15] More comparative experiments still need to be made.

forms, either the result lists need to be lemmatized, or several forms of the same lemma show up in the same list[16].

- Only those categories need to be kept as term candidates which usually have term status: noun, verbs and adjectives; the part-of-speech data from the tagged corpus thus can be used for selection purposes.

Our version of the relative frequency comparison tools keeps track of the above specificities; its output are single word term candidates which are subsequently analyzed by means of the "morpheme collector", to provide input to symbolic extraction routines.

2.4 Bootstrapping term candidate extraction – architectural considerations

So far, we have discussed different tool components which serve to identify term candidates and/or domain-specific morphemes. Experiments are under way to assess which combination of these components is optimal for our purposes, and to find out whether for different text types, specific different combinations (or a different sequencing) of the tools are required.

The following stepwise procedure has proven quite successful so far:

1. Use relative frequency to identify domain-specific morphemes;
2. Use the domain-specific morphemes from (1) to identify single-word term candidates, by means of regular expression queries over affixes and domain-specific morphemes;
3. Use the single word term candidates from (2) as a relevance filter for the results of queries for multiword term candidates, extracted by means of regular expression queries over part-of-speech shapes and lemma information.

A conservative approach is to combine the data extracted by means of different tool components and to compare the results; term candidates fulfilling criteria of several kinds (e.g. those which contain domain-specific morphemes and have a significant relative frequency score) may then be presented to the user as "best candidates".

In our terminology extraction work on car maintenance manuals, we could use an existing terminological glossary as a relevance filter for multiword term candidates: the glossary contains only very few multiword terms, but it includes many single word terms. Below, in sections 3.3 and 3.2, we will report on their use to filter multiword term candidates according to their likely usefulness for a glossary.

Table 1 summarizes the steps and the pertaining tool components:

[16] Not much information is lost because of lemmatization: only in rare cases, forms distinguish term candidates from non-terminological material: *technische Unterlagen* (plural) may be considered as a term, whereas (*standfeste, harte, weiche*) *Unterlage* may not, or may be seen as a different term.

Step	Operation	Tool
1a	Find single-word term candidates	Relative Frequency
1b	Find relevant morphemes in single-word terms (including compounds)	Morphological Analyzer
2	Find all compounds with relevant morphemes	Corpus Query
3	Find multiword terms	Corpus Query
4	Apply filters for "term status"	Lexical Lists Termbank

Table 1: Acquisition of term candidates by bootstrapping: steps and tool components

3 Results

In the following, we briefly comment on some of the results of our approach to term candidate extraction; we focus on multiword term candidates.

3.1 Single word term candidates

The relative frequency-based search for candidates produces several types of noise which can be removed to a large extent by use of filters (see above: lemma frequencies, part-of-speech filtering, etc.).

We combine the results obtained via relative frequency with those produced by means of symbolic queries. There is a large zone of overlap between both, but the symbolic procedure leads to less noise than the statistical one, and it also retrieves items which are terminologically relevant but not particularly frequent in the specialized text[17].

3.2 Noun-Noun collocations

We have extracted and analyzed noun-noun collocations of two types:

- Nouns with genitive NPs
  (e.g. *Ablauf der Frist, Erstellung eines Bauleitpans*);
- Nouns with prepositional groups
  (e.g. *Prüfkupplung für Diagnose, Infrarotfernbedienung für Zentralverriegelung*).

[17] An assessment, on the basis of a small-scale test against manual selection results, has been given in [Heid et al. 1996].

Both types are only weakly determined by (morpho-)syntactic patterns, and – contrary to what [Pearson (1998)] observed in English corpora – determination cannot be used as a clue, at least not in texts of the type we used (maintenance manuals). Having a list of single word term candidates, we can however rather easily filter the noun-noun collocation candidates produced by the system, according to whether the first or the second, or both nouns are assigned term status independently; the necessary knowledge either comes from a termbank (in case just multiword terms need to be added to an otherwise existing terminological data collection), or from the first step of our extraction procedures.

From data extracted from the 3.2 million words of the car maintenance corpus, it seems that indeed only the combinations of two nouns which both have term status independently, are again terminologically relevant (see the most frequent examples with absolute frequency figures in the maintenance manual corpus, in (12), below), whereas cases with only the first element as a term are marginal and cases with only the second item considered to be a term have rather to do with actions performed with objects denoted by the term.

(12)
```
720  Prüfkupplung für Diagnose
305  Infrarotfernbedienung für Zentralverriegelung
273  CD-Spieler mit Wechsler
166  Spule für Transponder
160  Fahrzeug mit Fremdzündungsmotor
127  Kopfstützenabsenkung im Fond
```

As with adjective-noun collocations (cf. the examples in (7) above), the complex noun groups give information on subtypes of certain concrete objects: for example, next to *automatische Kindersitzerkennung*, we also have *Kindersitzerkennung im Beifahrersitz*, and types of motors are classified according to certain of their components (see (13)), expressed by means of *mit*-phrases (examples are given with their absolute frequency):

(13)
```
 6  Motoren mit Aluminiumsaugrohr
 7  Motoren mit Kunststoffsaugrohr

 6  Motoren mit Reiheneinspritzpumpe
 5  Motoren mit Verteilereinspritzpumpe

 6  Motoren mit Kältekompressor
10  Motoren mit Kompressor

 8  Motoren mit Viscolüfterkupplung
```

3.3 Adjective-Noun collocations

The extraction of adjective-noun-collocations has so far mainly been based on attributive adjectives; it can however also make use of predicative adjectives, although predicative contexts provide much less terminologically relevant collocational information (the adjective *möglich*, alone, makes up for about 10 to 15 % of the predicative contexts, in our manual corpus).

More examples[18] of subtype-denoting adjective-noun-collocations can be found in (14) below. Note that in the examples given in (14), both the noun and the adjective are considered as terms, according to the relative frequency-based extraction procedure and/or the existing terminological database.

Typically, adjective-noun-collocations the noun component of which is considered as a term, whereas the adjective is not, tend to be much less relevant for a terminological data collection. There are two types of such cases: the adjective belongs to general language and is used to denote positions, sizes etc. of objects (see the examples in 15), or the adjectives denote general or normal cases, or they are used to make reference to an object introduced in the "discourse" of the manual previously. Examples of this latter group can be found in (16) and (17). For practical purposes, we have used a "stop list" of such "generic" adjectives, to remove any occurrences of adjective-noun-collocations of this type from our term candidate set. In the car manual corpus, these "generic" collocations amount to ca. 25% of all N/A-collocation types found.

(14)	Außenspiegel	abblendbar	7
	Außenspiegel	abklappbar	5
	Außenspiegel	verstellbar	40
	Außenspiegelposition	gespeichert	17
	Außenspiegelverstellung	elektrisch	12
	Einbruchdiebstahlwarnanlage	aktiviert	12
	Einbruchdiebstahlwarnanlage	deaktiviert	8

(15) *vorder, hinter, links, rechts, ober, unter, inner, äušer, mittler, groš, klein.*

(16) *gesamt, ander(e), allgemein, angegeben, normal.*

(17)	Bauteil	betreffend	6
	Bauteil	folgend	69
	Bauteil	genannt	5
	Bauteil	restlich	7
	Betriebszustand	außergewöhnlich	8
	Betriebszustand	jeweilig	17
	Betriebszustand	nachfolgend	5

Noun-adjective-combinations the noun component of which is not considered as a term (or at least not as a term of the relevant domain) tend to be general language collocations (*unsachgemäße Behandlung, offenes Feuer*, etc.) or collocations from fields of specialization outside the main topic area of the text (e.g. *anerkannter Sachverständiger* in the maintenance manuals); examples can be found in (18).

(18)	Sachverständiger	anerkannt	11
	Tuch	fusselfrei	10
	Behandlung	unsachgemäß	5

[18] Format: Noun lemma – adjective lemma – frequency of cooccurrence in the texts.

Feuer	offen	36
Erwärmung	spürbar	8
Finger	bloß	5
Bewegung	kreisend	7

To identify these collocations, a collection of general language adjective-noun collocations (e.g. from newspaper texts) would be very useful as a filter. If we have identified these collocations already in general language text, we have independent evidence for their status as "non-terminological" unit. This procedure is necessary for work in domains where nominal term candidates may be lexemes which have both a terminological and a non-terminological meaning, and where collocations of both meanings exist.

Adjective-noun collocations retrieved from the text material can also be used to support the overall goal of setting up a terminologically controlled, synonym-free controlled language. If the extraction results are listed by order of nouns (all adjectives along with a given noun), it is relatively easy for terminologists to identify those adjective-noun combinations which are likely to be redundant synonyms. A few examples are indicated below, along with the number of occurrences identified in the text of the car maintenance manuals.

(19)
- Drehzahl: *nieder* (6), *niedrig* (29);
- Fehler: *abgespeichert* (41), *gespeichert* (43);
- Heckscheibe: *heizbar* (549), *beheizbar* (72);
- Kraftstoff: *bleifrei* (16), *unverbleit* (19);
- Kühlmittel: *warm* (8), *erwärmt* (8), *aufgeheizt* (6);
- Sitz (des Sicherheitsgurts): *gut* (10), *korrekt* (55), *richtig* (187), *einwandfrei* (11).

4 Conclusion

This paper summarizes some of our term candidate extraction work currently under way for German. The overall approach is a bootstrapping exercise: first single word term candidates are extracted, then collocational evidence is searched, for the single word term candidates identified beforehand.

The procedures rely on standard low-level corpus processing technology: tokenization, POS-tagging, lemmatization and regular queries. Since it is not easy (if at all possible) to give a definition of what makes a "good term candidate", it is correspondingly difficult to assess the tools quantitatively. In parallel, strategies to remove non-terminological candidates have to rely on frequency figures and on a comparison with existing data from "general language" corpora or from an existing terminological glossary. A qualitative assessment, however, provides at least some indications about possible types of noise, and, consequently, about possibilities to exclude the noise.

The tool components used for term candidate extraction are still under development. They are a part of a more general toolbox for the analysis of German text corpora. In the near future, more work on the extraction of those types of grammatical and morphosyntactic information is planned which are necessary for a more refined automatic processing of the texts, by means of a parser: agreement features, syntactic subcategorization of predicates, specific syntactic and distributional properties, etc. Along with this, work on terminological variation and attempts towards the use of parallel corpora for the extraction of candidate equivalents are under way.

The approach has been tried on automotive maintenance manuals, so far, as well as on the legal texts from the Land Südtirol, and on a few smaller text collections. Engineering texts from different technical fields as well as administrative terminology will be the next domains to be used for texts.

References

[Ahmad *et al.* (1992)] Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1992). What is a term? The semi-automatic extraction of terms from text. In M. Snell-Hornby *et al.*, eds., *Translation Studies - an interdiscipline*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

[Bourigault *et al.* (1996)] Bourigault, D., Gonzalez-Mullier, I., and Gros, C. LEXTER, a Natural Language Processing Tool for Terminlogy Extraction. In *Proc. Seventh EURALEX International Congress on Lexicography.* Göteborg, Sweden.

[Christ (1994a)] Christ, O. (1994a). A Modular and Flexible Architecture for an Integrated Corpus Query System. In F. Kiefer, G. Kiss, and J. Pajzs, eds., *Papers in Computational Lexicography, COMPLEX '94*, 22–32. COMPLEX 94, Budapest.

[Christ (1994b)] Christ, O. (1994b). *The Xkwic User Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[Christ/Schulze (1996)] Christ, O. and Schulze, B. Ein flexibles und modulares Anfragesystem für Textcorpora. In H. Feldweg and E. Hinrichs, eds. *Lexikon und Text.* 121-133. Max Niemeyer, Tübingen.

[Heid (1996)] Heid, U. (1996). Computerunterstützter Aufbau von Glossaren. In Arnold/Lee-Jahnke, ed., *Équivalences '95 – Les Actes*, Lausanne 1996.

[Heid et al. 1996] Ulrich Heid, Susanne Jauß, Katja Krüger, Andrea Hohmann: "Term extraction with standard tools for corpus exploration - Experience from German", in: *Proceedings of TKE '96*, (Frankfurt: Indeks), 1996

[Jauß(1996)] Jauß, S. (1996). *Erkennung und corpusbasierte Extraktion von Fachterminologie unter Einbeziehung der Terminologiedatenbank* Interfass *der Mercedes-Benz AG*. Master's thesis, Institut für maschinelle Sprachverarbeitung - Computerlinguistik, Universität Stuttgart.

[Krüger (1996)] Krüger, K. (1996). *Corpusbasierter Aufbau von Glossaren – ein integriertes System zur Extraktion von Termkandidaten aus deutschen und englischen Fachtexten*. Master's thesis, Institut für maschinelle Sprachverarbeitung - Computerlinguistik, Universität Stuttgart. To appear.

[L'Homme (1998)] L'Homme, M.-C. Caractérisation des combinaisons lexicales spécialisés par rapport aux collocations de langue générale. In *Proc. Eighth EURALEX International Congress on Lexicography 1998.* Liège, Belgium.

[Otman (1991)] Otman, G. (1991). Des Ambitions et des performances d'un système de depouillement terminologique assisté par ordinateur. *La banque des mots, revue semestrielle de terminologie française publiée par le Conseil international de la langue française*, 59–96.

[Pearson (1998)] Pearson, J. (1998). *Terms in Context.* John Benjamins, Amsterdam/Philadelphia.

[Reinhardt *et al.* (1992)] Reinhardt, W., Köhler, K., and Neubert, G. (1992). *Deutsche Fachsprache der Technik*. Georg Olms Verlag, Hildesheim.

[Reinke (1996)] Reinke, U. (1996). Der Einflußintegrierter Übersetzungssyteme auf den Übersetzungsprozeß. In *Sprache und Wirtschaft in der europäischen Informationsgesellschaft*. 12. - 15.10.1995, Humboldt-Universität Berlin. To appear.

[Rooth/Carroll 1996] Mats Rooth, Glenn Carroll: "Valence Induction with a Head-Lexicalized CFG", (Stuttgart: IMS), ms. 1996

[Schiller (1994)] Schiller, A. Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO. In R. Hausser, ed., *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*, 37-52. Max Niemeyer, Tübingen.

[Schmid (1994a)] Schmid, H. Part-of-Speech Tagging with Neural Networks. In *Proc. 15th International Conference on Computational Linguistics (Coling)*. Kyoto, Japan.

[Schmid (1994b)] Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. International Conference on New Methods in Language Processing (NeMLaP)*. Manchester, UK.

[Schulze *et al.* (1994)] Schulze, B. M., Heid, U., Gaschler, J., Grefenstette, G., Rooth, M., Schiller, A., Schmid, H., and Teufel, S. (1994). Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-oriented Tools. Decide, deliverable d–1b i, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

[Schulze 1996] Bruno Maximilian Schulze: *MP user manual*, Stuttgart: IMS, 1996.

[Teufel 1995] Simone Teufel: *ELM-DE: A typed incarnation for German of the EAGLES Standard Proposal for Morphosyntactic Annotation – Lexical Specification and Classification Guidelines*, (Stuttgart/Pisa: IMS/EAGLES) 1995, ms. 172 pp. See also the electronic version, on the URL of the EAGLES project: `http://www.ilc.pi.cnr.it/EAGLES/home.html`

[Weber (1994)] Weber, N. (1994). Maschinelle Hilfen bei der Herstellung, Verwaltung und Überarbeitung von Fachwörterbüchern.