[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Computer-aided analysis of multilingual patent documentation

Ingeborg Blank

CIS - Centre for Information and Language Processing
University of Munich
Oettingenstr. 67, D- 80538 Munich
blank@cis.uni-muenchen.de

Abstract

This paper deals with the processing of a multilingual corpus of technical texts. The aim is to extract special purpose terminology. A semi-automatic tool is developed to help professional translators and terminologists not only to identify technical terms but also to detect possible translation equivalences and typical contexts of terms. Definitions of terminology reported in the literature are discussed. Related studies in multilingual terminology extraction are also considered and the assumptions underlying these studies are examined on the corpus.

1 Introduction

A multilingual text corpus is a collection of texts translated into several languages. It provides a major source of linguistic information useful for translation; according to Isabelle:

Given the staggering volume of translations produced year after year, it is quite obvious that existing translations contain more solutions to more translation problems than any other existing resource. Unfortunately translators can currently derive very little benefit of this fact. (Isabelle, 1992: 8)

This is the idea underlying the present work. Due to the increase of international relations in trade and technology, translation of technical texts, that is texts written in "Languages for Special Purposes" (LSP) as defined by (Picht & Draskau, 1985), is continuously growing. LSP texts are the basic resource for collecting technical terms. The acquisition of bilingual lists of terminology expressions is difficult and time consuming. It is, therefore, worthwhile to investigate methods to compile such lists as automatically as possible. This paper deals with the development of methods to process a multilingual corpus. The developed methods result in a semi-automatic tool that helps professional translators and terminologists not only to identify technical terms, but also to detect possible translation equivalences and typical contexts of terms. This paper is structured as follows: Section 2 identifies the corpus used. Section 3 describes how terminology is defined for German and French. The extraction of German and French candidate terms is dealt with in section 4. Other related methods for terminology extraction are discussed in section 5. Section 6 points out some examples of possible applications of our approach.

2 The corpus

A trilingual (German-English-French) corpus of technical texts comprising about 12 million words was provided by the European Patent Office in Munich (EPO). The corpus includes two subcorpora each one containing a special type of documents.

- The major one is the DBA subcorpus consisting of about 1000 decisions of the boards of appeal (about 10 million words). Each decision is written in one of the three languages and then translated into the other two.
- The other one is the EPC subcorpus which is a collection of the articles and rules governing the European patent system.

The texts contained in the above corpora have legal value. Therefore, the main part of the terminology included therein is juridical and the remaining part is relating to all technical fields mentioned in the International Patent Classification (IPC) system covering all domains of chemistry, mechanics or physics.

The corpus used is particularly suitable for defining and extracting multilingual terminology, for the following reasons:

- it is structured in a very concise, homogenous and uniform manner,
- it is sufficiently big to be statistically relevant and
- the texts are written in a legal context i.e. the translations are of good quality.

For ergonomie reasons, the present study was restricted to German and French texts only. The EPC subcorpus was used for the definition of terms and the part of the DBA subcorpus referring to chemistry (40000 words per language) was used for the extraction of terms.

The experiments were carried out on parallel texts aligned on the sentence level, i.e. texts converted to corresponding segments of one or a few sentences. We used an implementation of the Church-Gale method that had yielded an accuracy of about 95% on a test corpus of about 400 000 words (Blank, 1995).

3 Definition of terminology

The basis for the linguistic definition of terms was literature from terminology, translation science, information retrieval, linguistics and computational linguistics. Terminology science, as founded by Wuester, is an interdisciplinary domain that aims at the definition, collection, storage and diffusion of terminology. Terms are usually defined by semantic criteria according to ISO/DIS 1087 (1988:7):

term: designation of a defined concept in a special language by a linguistic expression.

"Concepts" are also defined by ISO/DIS 1087 (1988:2):
concept: unit of thought constituted by those characteristics which are attributed to an object or a class of objects, note: concepts are not bound to particular languages. They are, however, influenced by the social or cultural background.

Definitions of terms by semantic criteria are, however, not suitable for an automatic procedure. A program for identifying likely terminological units, must take into account the form of terms, i.e. their syntactic and morphological properties. Definitions of that kind can be found in some branches of terminology science and in computational linguistics.

The prescriptive branch of terminology science provides descriptions of the external form of terms as well as "norms" ruling the formation of new terms. In most studies in computational linguistics, technical terms are defined as noun phrases that satisfy a rather restricted set of morpho-syntactic patterns.

Thus, following the above definition, nouns (simple or compound) and noun phrases (built up according to some frequent patterns) are considered to be candidate terms e. g. "Beschwerdeverfahren" in German, "appeals procedure" in English or "procédure de recours" in French.

In order to check the accuracy of this syntactic definition of terminology, parts of the EPC subcorpus were manually parsed in maximal-length noun phrases and, when necessary, segmented in smaller phrases. Such a subcorpus is particularly suited for the detection of terminology. It contains definitions of the basic concepts of the European patent system and the corresponding terms for expressing these concepts. Moreover, in some cases noun phrases are explicitly marked as terms[1].

1. e. g. "Patents granted by virtue of this Convention shall be called European patents" (Art. 2 (1) EPC).

Once the terms detected their morpho-syntactic properties have been determined. It was important to adopt a definition of the notion "term" that facilitates the comparison of terms in both languages.

Word formation is very different in French and German. French compounds consist of orthographically separated elements (e. g. "chambre de recours"), German compounds are often formed by composition of morphemes resulting in one orthographic word (e. g. "Beschwerdekammer"). The recognition of both types of compounding is not trivial in automatic processing (cf sections 3.1 and 3.2 of the present paper).

The definition of candidate terms finally used for the extraction was elaborated and checked on a part of the of EPC subcorpus (5000 words for each language). Candidate terms are defined as noun phrases satisfying a restricted set of **part-of-speech patterns** and are classified by their **length** i.e. the number of nouns, adjectives, verbs, and participles; e. g. the German term "Beschwerdeverfahren" has length 1 whereas the French term "procédure de recours" has length 2.

3.1. Candidate terms in French

Due to the particular properties of the word formation in French, it is sometimes impossible to establish a clear distinction between a free syntagma and a compound. This problem is discussed more in detail in several studies (Daille, 1994; Jacquemin, 1991; Bourigault, 1994). Two compounds can overlap and build a new compound, for instance "procédure de conversion d'hydrocarbures" (length 3) can be considered as a merge of two compounds of length 2, namely "procédure de conversion" and "conversion d'hydrocarbures", both occurring in the corpus. On the other hand, the whole nominal phrase is translated by one compound in German ("Kohlenwasserstoffumwandlungsverfahren").

The automatic recognition and extraction of French compounds is difficult for the following reasons:

- it is impossible to determine whether a morpho-syntactic structure is a sequence of length 2 before the detection of all compounds of length 3 and
- it impossible to determine whether a morpho-syntactic structure is a sequence of length 3 before the detection of all compounds of length 2.

For this reason we decided to adopt a broad definition of potential terms taking into account all noun phrases of length 2 and the most frequent types of noun phrases of length 3. For the unclear cases the denomination "terminological unit" would be more appropriate than the denomination "term".

The types of French terminological units used for the extraction stage are summarized in the following table I[1].

The German translation is also indicated in order to facilitate a comparative evaluation.

type	example in French	translation in German
Length 1:		
N	brevet	Patent
N-	sous-revendication	Unteranspruch
Length2:		
N DE N	demande de brevet	Patentanmeldung
N Adj	brevet européen	europäisches Patent
NAN	stabilisant à la lumière	Lichtstabilisator
N Prep N	protection par brevet	Patentschutz
N N	valeur limite	Grenzwert
Length 3:		
N DE N Adj	groupe d'états contractants	Vertragsstaatengruppe
N DE N DE N	décision de revocation de brevet	Widerrufsentscheidung
N A DE N	sulphoacétate laurique de sodium	Natriumlaurylsulfoacetat

Table I: Linguistic description of French candidate terms

3.2 Candidate terms in German

The major part of German technical terms are compounds. In handbooks of terminology almost all other types of formation are very often considered as just a transitional state for the formation of a "real compound".

The analysis of the German part of the EPC corpus resulted in the following description for German terms.[1]

type	example in German	translation in French
Length 1:		
Ns	Patent	brevet
Comp	Patentinhaber	titulaire de brevet

Table 2: Linguistic description of German candidate terms

1. The following abbreviations are used : N: noun, N-: hyphenated noun, Adj: adjective, Prep: prepositions other than "de" or "à", DE: the regular expression (de+d'+du+de la+de l'+des), A: the regular expression (à+au+à la+à l'+aux).

1. The following abbreviations are used in table 2: Ns: simple noun, Comp: compound noun, N: simple noun or compound noun, Adj:adjective, Prep: preposition, Det-gen: determiner in genitive, N-gen: noun in genitive.

type	example in German	translation in French
Length2:		
Adj N	mündliche Verhandlung	procédure orale
N Det-gen N-gen	Stand der Technik	état de la technique
N Prep (Det) N	Antrag auf Wiedereinsetzung	requête en restitutio in integrum

Table 2: Linguistic description of German candidate terms

4 Extraction of candidate terms

Technical terms are defined by a syntactic form and a semantic function i.e. the representation of a concept. As the grammatical form of terminological units is relatively predictable it is possible to devise an extraction program solely based on syntactic data. It is, however, irrealistic to expect this program to extract only terminological units and nothing else. Since an extraction program cannot capture the semantic function of terms, its output should be considered as candidate terms or likely terminological units.

The extraction task can be considered as a recall and precision problem. The extraction programs takes a text and a morpho-syntactic definition of terms as input and provides candidate terms as output. A very restricted definition of terms yields to a high precision i.e. the probability of getting candidate terms that are really technical terms is growing. With this approach, on the other hand, recall will drop i.e. a part of the technical terms occurring in the text will not be extracted by the program. The extraction program used in this study promotes completeness using the definition of candidate terms presented in the previous section. This approach is justified because it is easier for the terminologist or the translator to eliminate some "likely terminological units" than to find "real terminological units" that escaped detection by the program.

The output of the program must be evaluated in two stages. First it must be checked whether the extracted units are linguistically correct (i.e. well-formed noun phrases) and filter out incorrect sequences. In a second pass it should be judged whether the linguistically correct noun phrases are really domain-specific terminology. However, this question can only be replied by the skilled person in the specific technical field.

4.1 Extraction of French candidate terms

The INTEX system (Silberztein, 1993) is a tool for various lexical tasks, e. g. lemmatization, POS-tagging and search of linguistic patterns. It is based on a complex lexicon system that, initially, was designed for the application on French texts. We used INTEX on the French corpus for the lexical analysis and the search of candidate terms. Patterns were defined as regular expressions formed up by POS categories, a given word or a list of words. INTEX converted

the regular expressions to finite state automata and applied item to the corpus. This resulted in the extraction of sequences that were only considered as candidate terms if they were linguistically correct. INTEX can be tuned for special purposes by means of a user dictionary, a preference lexicon, local grammars etc. (see Blank, 1997). 83.02% of the extracted sequences were linguistically correct. Incorrect sequences were mainly due to disambiguation errors in POS tagging and to segmentation disambiguations of compounds of length 2 and length 3. The correctly extracted sequences can be divided in two subgroups:

- sequences with a number of occurrences higher than 4: 95% were also semantically correct (i.e. domain-specific terminology) and
- sequences with a number of occurrences between 2 and 4: 60-90% were semantically correct.

4.2 Extraction of German candidate terms

The German subcorpus had been lemmatized and annotated with POS categories by means of the CISLEX system. Additionally to this, the system recognizes complex forms and segments them into simple form components with an accuracy of about 98%. Sequences corresponding to the previously defined patterns were extracted by a program written in PERL.

About 98% of the extracted nouns (either simple or compound) were linguistically correct. The extraction of complex noun phrases reached an accuracy of 65%; for this kind of structures a more complex parsing system would be necessary.

The semantic correctness of the correctly extracted sequences varies according to the formation type and the number of occurrences. Among the different types of extracted sequences, the compounds have the highest probability to represent domain-specific terminology. 60-90% of the correctly extracted sequences with a number of occurrences greater or equal to 3 represented domain-specific terminology.

5 Related work

Studies in multilingual terminology extraction concern mainly English-French corpora (Church & Dagan, 1994; Gaussier, 1995; Kupiec, 1993). (Eijk, 1993), based on an English-Dutch corpus, mentioned similar problems in matching translations as those of the present study. This is, probably, due to the fact that word formation in Dutch is similar to German and the word formation in English is similar to French.

The above studies share a similar approach and are based on some assumptions about terminology that are, however, not explicitly stated:

(i) The definition and extraction of terms are based on morpho-syntactic patterns following the assumption that **the formal properties of terminology are relatively predictable.**

This assumption is the condition sine-qua-non for an automatic extraction. This appears in the results reported in section 4.

(ii) The extracted structures are filtered by statistical means or by human revision.

Thus, some studies use statistical measures like mutual information (see Daille, 1994 for a comparative evaluation of statistic filters) attempting to establish the terminological status of an extracted sequence. In the present study we adopted human revision as described by (Church & Dagan, 1994).

(iii) Candidate translations are matched by a statistical framework.

These procedures are based on two assumptions: the structural and the translational equivalence of terminology. **Translational equivalence between terms** means that a term in language A (source language) is translated 1:1 by a term in language B (target language). It refers to the common assumption that the translation of terminology is always standardized. This aspect will be discussed in section 5.1.

Structural equivalences of multilingual terminology means that candidate terms extracted from a corpus in language A according to a set A' of syntactic patterns are translated in a corpus in language B by candidate terms according to a set of syntactic patterns B'. This assumption is discussed in section 5.2.

A sample corpus of three documents was used for the examination of translational equivalences.

5.1 Translational equivalences

The translations of the first 25 most frequent terms of each document were checked. Both translation directions (from French to German and from German to French) were considered. It turned out that, in general, there is a standardized translation. However, 5-15% of the terms had more than one translation.

Some examples:

(1) The standardized translation of "décision de révocation" (N DE N) is "Widerrufsentscheidung" (Comp), but it is also translated by "Entscheidung" (Ns) or "Entscheidung über den Widerruf (N Prep N).

(2) The standardized translation of "maintien du brevet" (N DE N) is "Aufrechterhaltung des Patents" (N Det-gen N-gen) but it is also translated by "Aufrechterhaltung" (Comp) or "das Patent ... aufrechtzuerhalten" (infinitive phrase).

(3) The standardized translation of "rechtliches Gehör" (Adj N) is "principe du contradictoire" (N DE N), but it is

also translated by "droit des parties à être entendues" (complex noun phrase), "qu'elles seront suffisamment entendues" (subordinate clause) or "possibilités suffisantes de se faire entendre" (complex noun phrase).
(4) It seems that for "Einspruchsbeschwerdeverfahren" (Comp) a standardized translation does not exist yet, although it is a domain-specific term. This term was translated by "procédure de recours engagée à l'encontre d'une décision rendue sur opposition" (complex noun phrase) in all three documents.

5.2 Structural équivalences

Some of the above examples showed that French terms, built up according to the predefined set of morpho-syntactic patterns, are not always translated by German terms built up according to the predefined set of syntactic patterns and vice versa. The examination of the structural equivalences in the sample corpus resulted in the distinction of three cases:

(1) Source and target language terms correspond to the set of language-specific predefined patterns (see example (1)). This is valid for 93.5% of the French to German translations and for 67.9% of the German to French translations.

(2) The translation in the target language is a noun phrase but it does not correspond to the predefined set of patterns (see example (4): complex noun phrase).

This is valid for 1% of the French to German translations and for 15% of the German to French translations.

(3) The target-language translation is not a nominal phrase (see example (2): infinitive phrase and (3): subordinate clause).

This is valid for 5.5% of the French to German translations and for 17.1% of the German to French translations.

5.3 Conclusion about equivalences

The examination of translation equivalences reveals that the candidate terms extracted for French were, in general, translated in German by phrases belonging to the set of candidate terms extracted from the German subcorpus. This observation is less frequent the other way round i.e. from German to French. The reason for this is, probably, that the German structures with length 2 or 3 are not so often domain-specific terminology (except the structure Adj N). This fits with the considerations about the formation of German terminology found in handbooks. This means that we extracted more German than French candidate terms as it can be shown by a simple numerical evaluation of the extraction. For this reason we did not propose an automatic matching procedure for translations but we investigated other applications based on the extracted data.

6 Applications

The results are presented in the form of a concordance tool that assists translators and terminologists in constructing glossaries. This tool provides, among others, the following information:
(i) candidate terms and associated concordance lines,
(ii) contextual information for candidate terms and
(iii) grouping of candidate terms with common constituents.

6.1 Candidate terms and associated concordance lines

Each candidate term is presented with the sentence of the source text from which it was extracted and the sentence(s) of the target text aligned with said sentence. One must examine the relevant lines of the text in order to decide whether a candidate term is indeed a term, and to identify the multiword terms that are omitted from the candidate term list. The local and the global frequency are also indicated for each candidate term.

Table 3 gives an example of such a concordance .

French text	**German text**
II. Le 11 août 1982, la requérante a *fait* Opposition à ce *brevet européen*, et en a demandé la révocation pour défaut de nouveauté, en faisant valoir notamment de *nouvelles antériorités.*	II.Gegen diese Erteilung des *europäischen Patents* hat die Einsprechende am 11. August 1982 Einspruch eingelegt und den **Widerruf** des Patents wegen *mangelnder Neuheit* beantragt. Die Begründung wurde unter anderem *auf neue* ***Entgegenhaltungen*** gestützt.
III.Par décision en date du 13 octobre 1983, la **Division d'opposition** a rejeté l'opposition, au motif essentiellement que...	III. Durch Entscheidung vom 13. Oktober 1983 hat die **Einspruchsabteilung** den Einspruch zurückgewiesen. Die **Zurückweisung** wurde im wesentlichen damit begründet, daß...
Rien dans **l'état de la technique** ne permettait d'affirmer que **l'utilisation d'hexaméthylènediamine dans la préparation de zeolites** s'imposait à l'évidence.	Es gebe auch im Stand der Technik keine Anhaltspunkte, die die Verwendung von **Hexamethylendiamin** bei der Herstellung von Zeolithen naheliegend erscheinen lassen.
Le procédé revendiqué dans le brevet en litige permettait de préparer directement une zeolite sans alcali. possibilité qui devait être considérée comme inattendue.	Es sei als überraschend anzusehen, daß durch das Verfahren des **Streitpatents** direkt ein *alkalifreier Zeolith* hergestellt werden kann.

Table 3: Parallel text with candidate terms

French text	German text
Les autres antériorités avaient été publiées durant le **délai de priorité**, et ne pouvaient donc être prises en considération, puisque la priorité avait été revendiquée à juste titre; en effet, le fait que les **résultats d'analyses** ne soient pas identiques dans les exemples du **fascicule de brevet** d'une part et dans le texte du **document de priorité** d'autre part n'entraînait pas la perte du **droit de priorité**.	Die anderen **Entgegenhaltungen** seien im **Prioritätsintervall** veröffentlicht und daher - da die Priorität zu Recht beansprucht sei - nicht zu berücksichtigen; denn **Prioritätsverlust** trete nicht dadurch ein, daß die **Analysenergebnisse** in den Beispielen der **Patentschrift** einerseits und den **Prioritätsunterlagen** andererseits nicht identisch seien.

Table 3: Parallel text with candidate terms

6.2. Contextual information for candidate terms

Verbal, nominal or other contexts in which each term is used are indicated. Language-specific syntagmatic lexical information is very important for translators. Texts that are correct on this level are perceived as fluent and natural. Examples of typical contexts of the French term "procédure orale" are shown in table 4.

Type of context	Example
prepositions	dans une procédure orale avant la procédure orale lors d'une procédure orale
verbs	une procédure orale s'est tenue une procédure orale s'est déroulée la procédure orale a relevé comparaître à une procédure orale être représenté dans une procédure orale prendre part à une procédure orale interrompre une procédure orale demander de recourir à une procédure orale organiser une procédure orale
nominal contexts	la tenue d'une procédure orale l'interruption d'une procédure orale le procès-verbal d'une procédure orale
other contexts	au cours d'une procédure orale après la tenue d'une procédure orale à l'issue d'une procédure orale au terme d'une procédure orale à la suite de la procédure orale

Table 4: Typical contexts of "procédure orale"

Examples of typical contexts of the German term "mündliche Verhandlung", that is the translation equivalent of "procédure orale", are shown in table 5.

Type of contexts	Example
prepositions	in einer mündlichen Verhandlung vor einer mündlichen Verhandlung während einer mündlichen Verhandlung nach einer mündlichen Verhandlung
verbs	eine mündliche Verhandlung findet statt die mündliche Verhandlung hat ergeben zu einer mündlichen Verhandlung erscheinen in einer mündlichen Verhandlung vertreten sein teilnehmen an einer mündlichen Verhandlung eine mündliche Verhandlung unterbrechen eine mündliche Verhandlung beantragen
nominal contexts	die Durchführung einer mündlichen Verhandlung die Beteiligten einer mündlichen Verhandlung die Unterbrechung einer mündlichen Verhandlung
other contexts	die Niederschrift über eine mündliche Verhandlung bei Abschluß einer mündlichen Verhandlung am Schluß einer mündlichen Verhandlung am Ende einer mündlichen Verhandlung im Anschluß an eine mündliche Verhandlung

Table *5:* Typical contexts of "mündliche Verhandlung"

6.3 Grouping of candidate terms with common constituents

All noun phrase terms that have either the same head or other constituents in common are grouped together in a kind of web. Such a grouping of linguistically related terms makes it easier to judge their validity and gives a lexical

1. The following annotations are used: in the French text compounds of type "N DE N" are written in bold, compounds of type "N Adj" in italics, other types of extracted structures are underlined; in the German text nominal compounds are marked in bold, phrases of type "Adj N" in italics and other types of structures are underlined.

overview of the terms of a certain domain (cf annex for further explanation). It would be possible to construct from this grouping a kind of terminological hypertext web as described by (Bourigault, 1994).

From this "terminological web" terms like "brevet", "demande de brevet", "titulaire de brevet", "revocation de brevet", "demande de revocation de brevet", "protection par brevet", "brevet en litige", "brevet litigieux" etc. are grouped together.

7 Conclusion

In this study we developed a method for the extraction of German and French terms from a bilingual corpus of patent documentation. The results are used for designing a concordance tool suitable for translators and terminologists. A linguistic definition of German and French terms was elaborated which was the basis for the extraction algorithm. We checked the assumptions underlying the procedures commonly used for the matching of translations. Due to the particularities of the word formation in each language an automatic matching of translations was not considered. The concordance tool developed seems to be an efficient assistance in terminological work. For the time being it is unknown whether the results of the present study can be repeated for other language couples with closer morphology than German-French (e.g. English-French) and other fields than the patent domain.

Acknowledgment

We would like to thank the European Patent Office for providing a machine readable version of the corpus used in the present study. We would also like to thank the stuff of the Language Service and of the Principal Directorate of Chemistry of the EPO for their contribution in this project.

References

Blank, I. (1995). "Sentence alignment: methods and implementations. In *Traitement automatique des langues* Vol. 36, numéro 1-2, (pp. 81-89).

Blank, I. (1997). Computerlinguistische Analyse mehrsprachiger Fachtexte. Doctoral thesis. University of Munich, Centrum fur Informations- und Sprachverarbeitung, (CIS-Bericht 98-109).

Bourigault, D. (1994). LEXTER, un Logiciel d'EXtraction de TERminologie: Application à l'acquisition des connaissances à partir de textes. Thèse de doctorat. Ecole des Hautes Etudes en Sciences Sociales, Paris.

Church, K. W. & Dagan, I. (1994). Termight: Identifying and translating Technical Terminology. In *Proc. of the 4th Conference on Applied Natural Language Processing* (pp. 34-40), Stuttgart.

Daille, B. (1994). Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques. Thèse de doctorat en informatique fondamentale. Université Paris VIL

Eijk, P. van der (1993). Automating the acquisition of Bilingual Terminology. In *Proc. of the Meeting of the European Chapter of the Association for Computational Linguistics* (pp. 113-119), Utrecht.

Gaussier, E. (1995). Extraction automatique de lexiques bilingues par des méthodes statistiques. Thèse de doctorat en informatique fondamentale. Université Paris VII.

Isabelle, P. (1992). Bi-textual aids for translators. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

Jacquemin, C. (1991). Transformation des noms composés. Thèse de doctorat en Informatique Fondamentale. Université Paris VII.

Kupiec, J. (1993). An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proc. of the 31th Annual Meeting of the Association for Computational Linguistics* (pp. 17-22), Columbus, Ohio.

Maier-Meyer, P. (1995). Lexikon und automatische Lemmatisierung. Doctoral thesis. University of Munich, Centrum fur Informations- und Sprachverarbeitung (CIS-Bericht 95-84), Munich.

Picht, H. & Draskau, J. (1985). Terminology: an introduction. Guilford: The University of Surrey.

Silberztein, M. (1993). Dictionnaires électroniques et reconnaissance lexicale automatique, Paris: Masson.

Sta, J.-D. (1995). Comportement statistique des termes et acquisition terminologique à partir de corpus. In *Traitement automatique des langues* (pp. 119-132), Vol. 36.

Annex

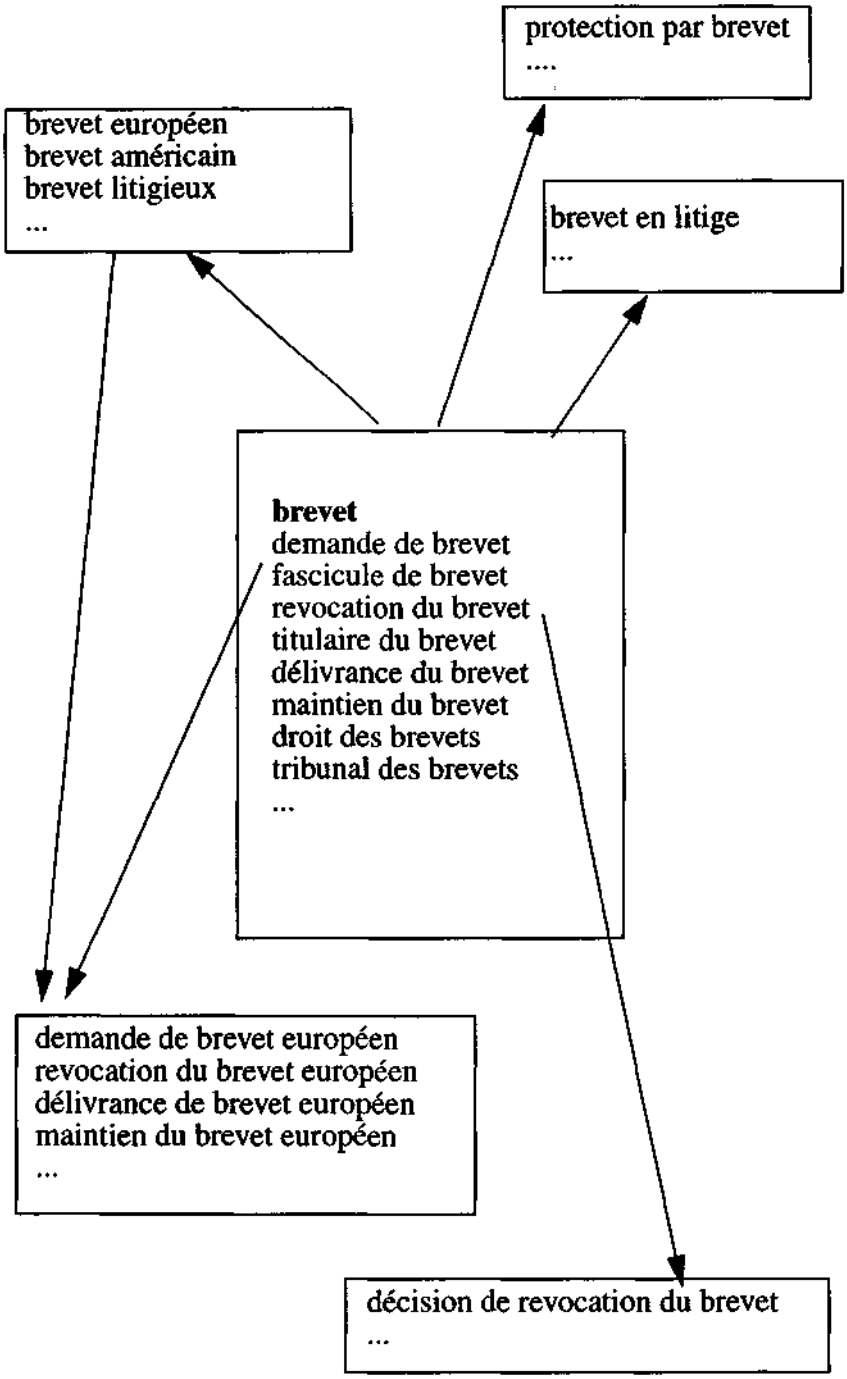

Figure 1: French terms containing the word "brevet" (patent)

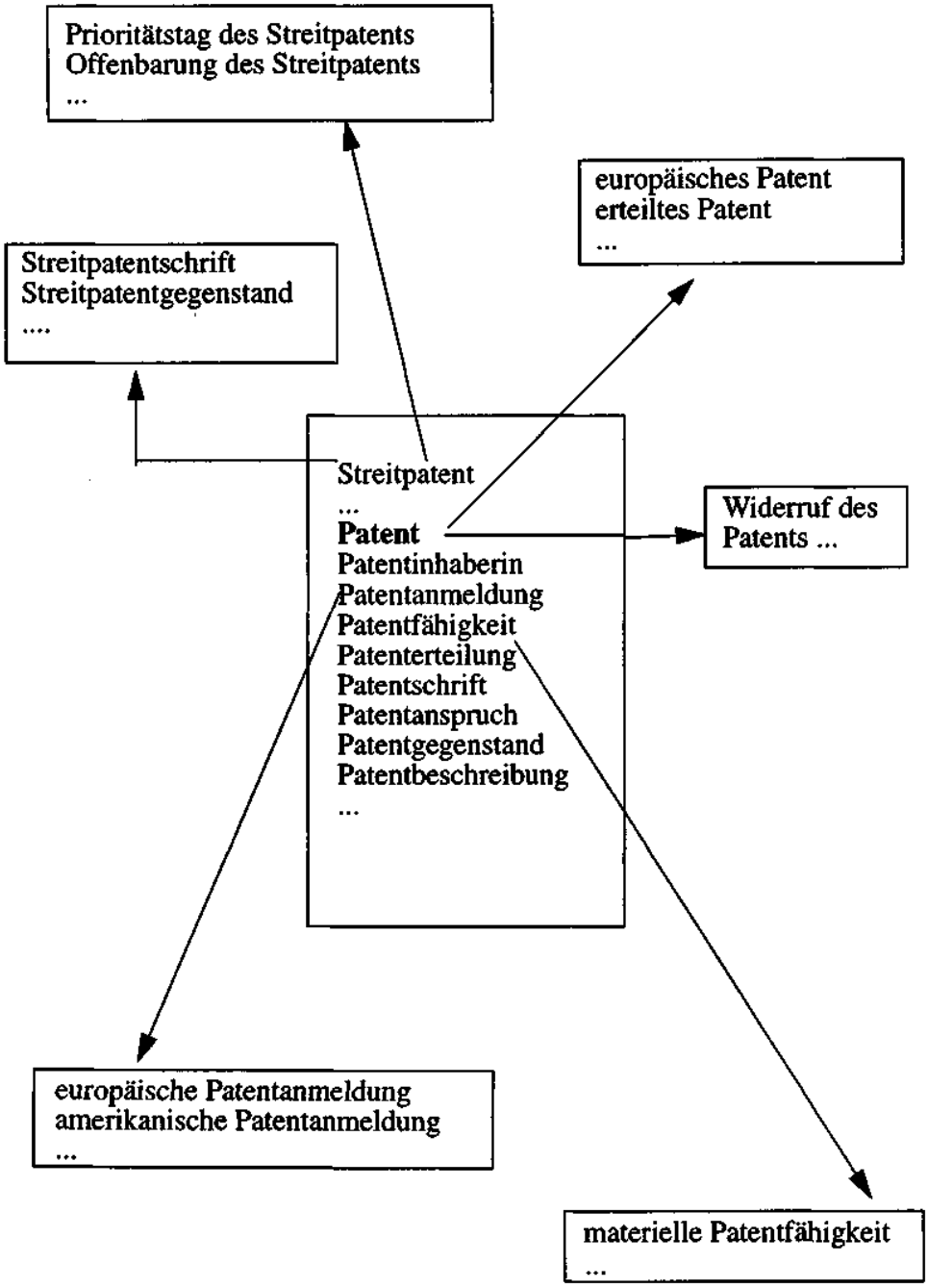

Figure 2: German terms containing the word "Patent" (patent)