Quantifying the Challenges in Parsing Patent Claims

Suzan Verberne*
s.verberne@let.ru.nl

Eva D'hondt†
e.dhondt@let.ru.nl

Nelleke Oostdijk†
n.oostdijk@let.ru.nl

Cornelis H.A Koster‡
kees@cs.ru.nl

ABSTRACT

In this paper, we aim to verify and quantify the challenges of patent claim processing that have been identified in the literature. We focus on the following three challenges that, judging from the numbers of mentions in papers concerning patent analysis and patent retrieval, are central to patent claim processing: (1) The length of sentences is much longer than for general language use; (2) Many novel terms are introduced in patent claims that are difficult to understand; (3) The syntactic structure of patent claims is complex. We find that the challenges of patent claim processing that are related to syntactic structure are much more problematic than the challenges at the vocabulary level. The sentence length issue only causes problems indirectly by resulting in more structural ambiguities for longer noun phrases.

Keywords

Patent Claim Processing, Challenges in Patent Search, Vocabulary Issues, Syntactic Parsing

1. INTRODUCTION

Patent retrieval is a rising research topic in the Information Retrieval (IR) community. One of the most salient search tasks performed on patent databases is prior art retrieval. The task of prior art retrieval is: given a patent application, find existing patent documents that describe inventions which are similar or related to the new application. For every patent application that is filed at the European Patent Office, prior art retrieval is performed by qualified patent examiners. Their goal is to determine whether the claimed invention fulfills the criterion of novelty compared to earlier similar inventions [1].

In its classic set-up, prior art searching involves a large amount of human effort: Through careful examination of potential keywords in the patent application the patent examiner composes a query and retrieves a set of documents. Document by document is then analyzed to judge its relevance. From the relevant documents new keywords are added to the query and the process is repeated until relevant information has been found or the search possibilities have been exhausted. Since professional searchers are expensive, it is worthwhile investigating how the prior art searching process can be facilitated by retrieval engines. Previous work suggests that for prior art search, the claims section is the most informative part of a patent, but it is also the most difficult to parse [12, 25, 14, 13].

Among the language processing tasks that can support the patent search and analysis process are term extraction, summarization and translation [27]. In order to perform these tasks (semi-)automatically, at least sentence splitting and morphological analysis is needed but in many cases also some form of syntactic parsing. Existing natural language parsers may fail to properly analyze patent claims because the language used in patents differs from the 'regular' English language for which the tools have been developed [25].

Patent claims have a fixed structure: They consist of one long sentence, starting with "We claim:" or "What is claimed is:", followed by item lists ('series of specified elements'[1]), which are realized by noun phrases. The terminology used in patent claims is highly dependent on the specific topic domain of the patent (e.g. mechanical engineering).

The challenges related to patent claim processing are identified by a number of researchers in the patent retrieval field (see Section 2) but these studies lack any kind of quantification of the challenges: Most of them do not provide statistics on sentence length, sentence structure, lexical distributions and the differences between the language used in patent claims and the language used in large non-patent corpora.

In this paper, we aim to verify and quantify the challenges of patent claim processing that have been identified in the literature. We focus on the three challenges that are listed in the often-cited paper by Shinmori et al. (2003) [25] about patent claim processing for Japanese:[2]

1. The length of the sentences is much longer than for general language use;

2. Many novel terms are introduced in patent claims that are difficult to understand;

*Information Foraging Lab/Centre for Language and Speech Technology, Radboud University Nijmegen and Research Group in Computational Linguistics, University of Wolverhampton

†Information Foraging Lab/Centre for Language and Speech Technology, Radboud University Nijmegen

‡Information Foraging Lab/Computing Science Institute, Radboud University Nijmegen

1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10), March 28, 2010, Milton Keynes.

[1]http://en.wikipedia.org/wiki/Claim_(patent)

[2]The research on patent processing and retrieval has a somewhat longer history in Japan than in Europe and the U.S. because of the patent retrieval track in the NTCIR evaluation campaign [15].

3. The structure of the patent claims is complex.

Consequently, most syntactic parsers — even those that achieve good results on general language texts — fail to correctly analyze patent claims.

We chose these challenges because we think they are central to patent claim processing, which may be concluded from the frequent mentions of these challenges in other papers concerning patent analysis and patent retrieval (see Section 2). We expect that the challenges that Shinmori et al. found for Japanese will also hold for English patent claims. We will verify this in Section 4. In the same section, we will quantify the challenges related to sentence length, vocabulary issues and syntactic structure, using a number of (patent and non-patent corpora) and NLP tools.

First, in Section 2 we provide a background for the current paper. Then, in Section 3 we describe the data that we used.

2. BACKGROUND: PATENT PROCESSING

In this section, we discuss previous work on patent processing. The papers that we discuss here stress the complexity of the language used in patents, especially in the claims sections. Most of the work is directed at facilitating human patent processing, in many cases by improving the readability of patent texts.

Bonino et al. (2010) explain that in patent searching, both recall and precision are highly important [5]. Because of legal repercussions, no relevant information should be missed. On the other hand, retrieving fewer (irrelevant) documents makes the search process more efficient. In order to have full control over precision and recall, patent search professionals generally employ an iterative search process. This process can be supported by NLP tasks such as query synonym expansion (which is already commonly used in patent text searches), sentence focus identification and machine translation.

Mille and Wanner (2008) stress that of all sections in a patent document, the claims section is the most difficult to read for human readers [22]. This is especially due to the fact that in accordance with international patent writing guidelines, each claim must consist of one single sentence. Mille and Wanner mention similar challenges to the ones listed by Shinmori et al. (2003): sentence length, terminology and syntactic structure. However, they describe the terminology challenge not as an issue of understanding complex terms (as Shinmori does [25]) but as the problem of 'abstract vocabulary', which is not further specified in their paper.

In their introduction to the special issue on patent processing, Fujii et al. (2007) state that from a legal point of view, the claims section of a patent is the most important [12]. They describe the language used in patent claims as a very specific sublanguage and state that specialized NLP methods are needed for analyzing and generating patent claims.

Wanner et al. (2008) describe their advanced patent processing service PATExpert [27]. PATExpert is aimed at facilitating patent analysis by the use of knowledge bases (ontologies) and a set of NLP techniques such as tokenizers, lemmatizers, taggers and syntactic parsers. Moreover, it offers a paraphrasing module which accounts for a two-step simplification of the text: (1) splitting the text in smaller units, taking into account its discourse structure, and (2) transforming the smaller units into easily understandable clauses with the use of 'predefined well-formedness criteria'.

Tseng et al. (2007) experiment with a number of text mining techniques for patent analysis that are related to the analytical procedures applied by professional searchers on patent texts [26]. They perform automatic summarization using text surface features (such as position and title words). Moreover, they extend the porter stemmer algorithm and also an existing stopword word list, both focusing on the specifics of patent language. Tseng et al. identify the extraction of key-phrases as one of the main challenges in patent claim analysis because "single words alone are often too general in meanings or ambiguous to represent a concept". This relates to the 'abstract vocabulary'-problem as identified by Mille and Wanner (see above). Tseng et al. find that multi-word strings that are repeated throughout a patent are good key-phrases and likely to be legal terms.

Finally, Sheremetyeva (2003) uses predicate–argument structures to improve the readability of the claims section [24]. In her system, readability improvement is the first step in a suggested patent summarization method.

All of the papers mentioned in this section use some form of NLP to facilitate patent analysis by humans. In the IR field, however, patent retrieval is generally addressed as a text retrieval task that only uses word level information without deeper linguistic processing. Academic research on patent retrieval has mainly focused on the relative weighing of the index terms and on exploiting the patent document structure to boost retrieval [21]. For an overview of the state of the art in academic and commercial patent retrieval systems, we refer to Bonino et al. (2010) [5].

A small number of approaches to patent retrieval use linguistic processing to improve retrieval. The systems developed by Escora et al. (2008) and Chen et al. (2003) perform a combination of syntactic and semantic analysis on the documents [11, 8]. The work described by Koster et al. (2009) and D'hondt et al. (2010) aims at developing an interactive patent retrieval engine that uses dependency relations as index and search terms [18, 10]. In order to generate these dependency relations, a syntactic parser is developed that is especially adapted to analyzing patent texts. We will come back to this parser in section 3.2.

3. DATA

For the experiments reported in this paper, we use the subset of 400,000 documents of the MAtrixware REsearch Collection (MAREC) that was supplied by MatrixWare[3] for use in the AsPIRe'10 workshop. In the remainder of this paper, we will refer to this corpus of 400,000 patents as the 'MAREC subcorpus'.

3.1 Preprocessing the corpus

Since the aim of the current paper is to quantify the challenges of parsing patent claims, we first extracted the claims sections from the MAREC subcorpus, disregarding the other fields of the XML documents. Moreover, as we are developing techniques for mining English patent texts, we are only interested in those patents that are written in English.

Using a Perl script, we extracted all English claims sections (marked with `<claims lang="EN">`) from the directory tree of the MAREC subcorpus and removed the XML markup. This resulted in 67,292 claims sections[4] with 56,117,443

[3] http://www.matrixware.com/

[4] The other documents in the subcorpus either do not contain

words in total.

Having extracted and cleaned up all claims sections, we used a sentence splitter to split the claims sections in smaller units. As pointed out by [25], sentence splitting for claims sections is not a trivial task. Many sentences have been glued together using semi-colons (;). We therefore decided to not only use full stops as a split characters in our sentence splitter but also semi-colons.

We found that the 67,292 claims sections consist of 1,051,040 sentences.[5]

3.2 Parsing the corpus

In order to assess and quantify the third challenge listed in Section 1 (the complex syntactic structure of patent claims), we need a syntactic analysis of the MAREC subcorpus. To this end, we use the baseline version of the syntactic parser that is under development in the 'Text Mining for Intellectual Property' (TM4IP) project [18]. The aim of this project is to develop an approach to interactive retrieval for patent texts, in which *dependency triplets* instead of single words are used as indexing terms.

In the TM4IP project, a dependency triplet has been defined as two terms that are syntactically related through one of a limited set of relators (SUBJ, OBJ, PRED, MOD, ATTR, ...), where a term is usually the lemma of a content word. [10]. For example, the sentence

"The system consists of four separate modules"

will be analyzed into the following set of dependency triplets:

```
[system,SUBJ,consist] [consist,PREPof, mod-
ule] [module,ATTR,separate] [module, QUANT,four]
```

Using dependency triplets as indexing terms in a classification experiment, Koster and Beney (2009) have recently achieved good results for classifying patent applications in their correct IPC classes [17].

The dependency parser that generates the triplets is called AEGIR ('Accurate English Grammar for Information Retrieval'). In its baseline version, AEGIR is a rule-based dependency parser that combines a set of hand-written rules with an extensive lexicon.

The resolution of lexical ambiguities is guided by lexical frequency information stored in the parser lexicon. These lexical frequencies provide information on the possible parts of speech that can be associated with a particular word form. For example, in general English, we can expect *zone* as a noun to have a higher frequency than *zone* as a verb. For the current paper, we collected lexical frequency information from a number of different sources in order to examine the lexical differences between the English language use in patent claims compared to the language use in difference contexts. We will come back to this in Section 4.2.

For the current paper, we decided to parse 10,000 of the 67,292 English patent claims in the MAREC subcorpus. These 10,000 claims contain a total of 6.9 million words. Sentencing these claims using the sentence splitter described in Section 3.1 results in 207,946 sentences.

a claims section or are in a language other than English.

[5]Recall from Section 1 that patent claims are composed of noun phrases (NPs), not clauses. In the remainder of this paper, we use the word 'sentences' to refer to the units (mostly NPs) that are separated by semicolons and full stops in patent claims. We use the word 'noun phrase (NP)' if we refer to the syntactic characteristics of such units.

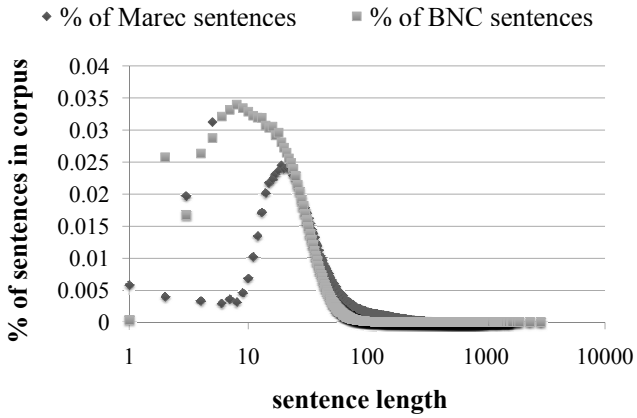

Figure 1: Distribution of sentence lengths in the MAREC subcorpus, compared to the BNC.

4. VERIFYING AND QUANTIFYING PATENT CLAIM CHALLENGES

The three challenges of patent claim processing mentioned in Section 1 are: (1) The length of the sentences is much longer than for general language use; (2) Many novel terms are introduced in patent claims that are difficult to understand; and (3) the structure of the patent claims is complex, as a result of which syntactic parsers fail to correctly analyze patent claims. In the following subsections 4.1, 4.2 and 4.3 we perform a series of analyses and experiments in order to verify and quantify these three challenges.

4.1 Challenge 1: Sentence length

After splitting the MAREC subcorpus into sentences (see Section 3.1), we extracted the following sentence-level statistics from the corpus. As already reported in the previous section, the 67,292 claims sections of the MAREC-400000 subcorpus consist of 1,051,040 sentences. There is much overlap between the sentences: after removing duplicates, 580,866 unique sentences remain. The median sentence length is 22 words; the average length is 53 words.

Binning the sentences from MAREC with the same length together and counting the number of sentences in each group results in a long tail distribution. The peak of the distribution lies around 20 words (25,000 occurrences), with outliers for sentence lengths 3 (20,637 occurrences) and 5 (32,849 occurrences). In Figure 1, the MAREC sentence length distribution is compared to the sentence length distribution of the British National Corpus (BNC) [19], which we preprocessed using the same sentence splitter as we used on the MAREC subcorpus.

Figure 1 shows that sentences in MAREC are, as the literature suggests, longer than the sentences in the BNC (the early peak is the BNC, the later peak is MAREC), even if we use the semi-colon for sentence splitting in addition to the full stop.

4.2 Challenge 2: Vocabulary

Shinmori et al. (2003) state that many novel terms are used in Japanese patent claims. We performed three types of analysis on the vocabulary level to verify this for En-

Table 1: Lexical coverage of the CELEX wordform lexicon on the MAREC subcorpus, both measured strictly and leniently (disregarding single characters, numerals and chemical formulae), and both on the type level and the token level.

CELEX–MAREC strict type coverage	55.3%
CELEX–MAREC lenient type coverage	60.4%
CELEX–MAREC strict token coverage	**95.9%**
CELEX–MAREC lenient token coverage	98.8%

glish patent claims: (1) a lexical coverage test of single-word terms from a lexicon of general English on the MAREC subcorpus, (2) an overview of the most frequent words in the MAREC subcorpus compared to the BNC, (3) frequency counts on ambiguous lexical items (as introduced in Section 3.2) and (4) an analysis of multi-word terms in the MAREC corpus.

The coverage of general English vocabulary

In order to quantify the differences between the vocabulary used in patent claims and general English vocabulary, we performed a lexical coverage test of the CELEX lexical database [2] on the MAREC subcorpus. The CELEX file EMW.CD contains 160,568 English word forms that are supposed to cover general English vocabulary: According to the CELEX readme file[6], the lexicon contains the word forms derived from all lemmata in the Oxford Advanced Learner's Dictionary (1974) and the Longman Dictionary of Contemporary English (1978). The CELEX documentation reports that on the 17.9 million word corpus of Birmingham University/COBUILD, the token coverage of CELEX is 92%.

We measured the coverage of CELEX entries on the MAREC subcorpus using a so-called corpus filter written in AGFL.[7] A corpus filter takes as input a corpus in plain text and a wordform lexicon. The corpus text is split up into tokens. These are matched to the lexicon using a smart form of matching with respect to capitalization: If a word is in the lexicon in *lowercase*, then it may match both an uppercase and a lowercase variant in the corpus. If a word in the lexicon has *one or more uppercase letters*, then it only matches equally uppercased forms in the corpus. This facilitates sentence-initial capitalization in the corpus for lowercase lexicon forms such as *the*, while it prevents proper names from the lexicon to be matched to common nouns in the corpus.

Moreover, the corpus filter allows us to skip over special tokens such as single characters, numerals and formulae. If we disregard these special tokens we get a more lenient lexical coverage measurement. We measured lexical coverage both on the token level (counting duplicate words separately) and the type level (counting duplicate words once). A type-level count always gives a lower lexical coverage because the words that are not covered by the lexicon are generally lower-frequency words. The lexical coverage (both type and token counts) for the CELEX lexicon on the MAREC subcorpus can be found in Table 1.

In Table 1, we marked the strict token coverage in boldface because this is the percentage (95.9%) that can be compared to the token coverage reported by the CELEX documentation on the COBUILD corpus (92%, see above). We can see that these percentages are comparable, the MAREC subcorpus giving a slightly higher coverage than the COBUILD corpus. Unfortunately, we cannot compare the type coverages of the CELEX lexicon for both the corpora because we do not know the type coverage of the CELEX lexicon on the COBUILD corpus.

If we look at the top-frequency tokens from MAREC that are not in the CELEX lexicon, we see that the first 26 of these are numerals (which we excluded in our lenient approach). If we disregard these, the ten most frequent tokens are: *indicia*, *U-shaped*, *cross-section*, *cross-sectional*, *flip-flop*, *L-shaped*, *spaced-apart*, *thyristor*, *cup-shaped*, and *V-shaped*.[8]

The lexical coverage of the CELEX lexicon on the MAREC corpus compared to the COBUILD corpus shows that patent claims do not use many words that are not covered by a lexicon of general English. The next three subsections should make clear what vocabulary differences do exist between patent claims and general English language use.

Frequent words

We extracted a word frequency list from the MAREC subcorpus. An overview of the 20 most frequent words in both the MAREC subcorpus and the BNC already shows remarkable differences (Table 2). The counts are normalized to the relative frequency per 10,000 words of running text. Three lexical items in Table 2 need some explanation:

- In patent claims, *said* is used as a definite determiner referring back to a previously defined entity.[9] In everyday English, it could be rephrased as 'the previously mentioned', e.g. "The condensation nucleus counter of claim 6 wherein said control signal further is a function of the differential of said error signal." *Said* has a strong reference function and can be used for the identification of anaphora in patent texts. The word occurs in 47% of all sentences in the MAREC subcorpus.

- The word *wherein* is used very frequently in claims for the specifications of devices, methods and processes. A brief, prototypical example is "The method of claim 4 wherein n is zero." *Wherein* occurs in 61% of all sentences in the MAREC subcorpus. If we only consider the 122,925 sentences that are around median sentence length (21–25 words), even 71% contains the word *wherein*. The frequent use of *wherein* is strongly connected to the nature and aims of patent claims: to define and specify all characteristics of an invention.

- The same holds for the word *comprising*, which is frequently used to specify a device or method, e.g. "The heat exchanger of claim 1 further comprising metal warp fibers..."

[6] http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html

[7] http://www.agfl.cs.ru.nl/

[8] This small set of terms shows that hyphenation is a productive and frequent phenomenon in patent claims. For that reason, the AEGIR grammar is equipped with a set of rules that accurately analyse different types of compositional hyphenated forms. In this paper, we will not go into specifics on this subject; it will be covered in future work.

[9] AEGIR treats this use of *said* as an adjective, as we will see later in this section.

Table 2: The 20 most frequent tokens in the MAREC subcorpus with their relative frequencies per 10,000 words of patent claims, and the 20 most frequent tokens in the BNC with their relative frequencies per 10,000 words of BNC texts.

MAREC claims		BNC	
Freq. per 10000 words	token	Freq. per 10000 words	token
674	the	715	the
480	a	376	of
457	said	303	and
450	of	266	to
278	and	206	in
261	to	202	a
158	in	129	is
128	claim	120	that
124	wherein	87	it
121	for	86	for
115	is	81	be
102	an	70	on
101	first	68	with
100	means	67	are
90	second	63	by
63	from	62	as
62	with	57	was
57	one	57	this
56	1	55	s
53	comprising	52	I

Table 2 shows a clear difference in the most frequently used words in patent claims (MAREC) compared to general English (the BNC). Thus, when we take into account the frequency of words, the language use in patent texts definitely differs from that found in general English (see the previous subsection).

Lexical frequencies for ambiguous words

As explained in Section 3.2, we consult several resources to obtain lexical frequencies. For the aim of the current paper, it is interesting to analyze the differences between the frequencies obtained from different types of sources. For development and analysis purposes, we obtained lexical frequencies from the following sources: (a) the Penn Treebank [20], (b) the British National Corpus BNC, (c) 79 Million words from the UKWAC webcorpus [3], POS tagged by the treetagger, (d) 7 Million words of patent claims from the CLEF-IP [23] corpus parsed with the Connexor CFG parser [16], and (e) the 6.9 Million words of patent claims from the MAREC corpus parsed with the AEGIR dependency parser (see Section 3.2).

We converted the annotations in each of the corpora to the AEGIR tagset.[10] We extracted from the AEGIR lexicon the 28,917 wordforms that occur in the lexicon with more than one part of speech (POS) and counted the frequencies of the wordforms for each of the POSs that occur in the corpora.

For each wordform w with parts of speech $p_{i..n}$ in source s, we calculated the relative frequency for each POS p as:

$$relfreq_{w,p,s} = \frac{count(w,p,s)}{\sum_{i=0}^{n} count(w,p_i,s)} \quad (1)$$

We took the average relative frequency over the sources $1..m$ as:

$$avgrelfreq_{w,p} = \frac{\sum_{j=0}^{m} relfreq(w,p,s_j)}{m} \quad (2)$$

We calculated the average relative frequency (Equation 2) for two sets of sources: Penn/BNC/UKWAC (PBU) on the one hand (representing general English language use), and MAREC/CLEF-IP (MC) on the other hand. Then we considered wordforms for which

$$avgrelfreq_{w,p,MC} - avgrelfreq_{w,p,PBU} > 0.5 \quad (3)$$

holds to be typical for patent claims.[11]

For example, the wordform *said* with part of speech 'adjective' comes out as being typical for patent language, whereas the same word with the part of speech 'verb' is labeled as atypical for patent language.[12] However, apart from this example it is difficult to draw any conclusions from the output of our lexical frequency analysis. Only 4% of the ambiguous wordforms for which we obtained lexical frequencies are labeled as typical for patent language.

One problem in the identification of typical wordforms is that it is difficult to distinguish between peculiarities caused by a different descriptive model of the parser/tagger used (e.g. one parser may prefer the label 'adjective' over the label 'past participle' for word forms such as *closed* in a phrase such as 'the closed door') and an actual difference in language use in the corpus (e.g. *said* as an adjective vs. *said* as a verb).

Most of the examples in the list of typical wordforms are difficult for us to interpret (e.g. *adhesive* as an adjective is labeled as typical while *adhesive* as a noun is labeled as atypical). Therefore, and because only a fraction (4%) of the words come out as typical for patent language, we consider the lexical frequencies for ambiguous words to be inconclusive. They do not show a clear difference between patent vocabulary and regular English vocabulary.

Multi-word terms

We include the topic of multi-word terms here because in Section 1 we referred to 'novel terms' (following Shinmori [25]) without distinguishing between single-words terms and multi-word terms. Since we found no difference between the single term vocabulary in general English and the English used in patent texts. (see 'The coverage of general English vocabulary'), we hypothesize that the authors of patent claims introduce complex multi-word NPs that constitute new (technical) terms.

To verify this, we make use of the SPECIALIST lexicon [6]. According to the developers this lexicon covers both

[10]For some tags this was not possible, for example where there was a many-to-many match between the labels used in a corpus and the labels used in the AEGIR tagset.

[11]The threshold of 0.5 was chosen because a difference value higher than 0.5 means that in the two text types the other of the two word classes for the same word is the majority word class.

[12]Interestingly, the Connexor CFG parser only labeled 55% of the occurrences of *said* in the CLEF-IP corpus as an adjective, and the other occurrences as a verb. We conjecture that these parsing errors are due to the fact that the Connexor parser was not tuned for patent data but for general English.

commonly occurring English words and biomedical vocabulary discovered in the NLM Test Collection and the UMLS Metathesaurus. By using lexical items from a reliable lexicon, we do not rely on syntactic annotation of the corpus; instead we assume that every occurrence of a word sequence from the lexicon in the corpus is a multi-word term.

The SPECIALIST lexicon contains approximately 200,000 compound nouns consisting of two words, 30,000 nouns consisting of three words, and around 10,000 nouns consisting of four or more words. We used these multi-word terms as input for a corpus filter as described in section 4.2. We found that fewer than 2% of the two-word NPs from SPECIALIST occurs in the MAREC subcorpus. For the three-word NPs, this percentage is lower than 1% and for the longer NPs it is negligible. The ten most frequent multi-word NPs from SPECIALIST in the MAREC corpus are *carbon atoms*, *alkyl group*, *hydrogen atom*, *amino acid*, *molecular weight*, *combustion engine*, *control device*, *nucleic acid*, *semiconductor device* and *storage means*. However, their frequencies are still relatively small. Moreover, the large majority of multi-word terms in patent claims are compositional in the sense that they are formed from two or more lexicon words, combined in one word-form following regular compositional rules. This means that for the purpose of syntactic parsing, it is not necessary to add these multiwords to the parser lexicon.

What does this mean? It seems that lexicalized multi-word NPs (terms from the SPECIALIST lexicon) do not occur very frequently in patent claims. This can be due to the topic domains covered by the MAREC subcorpus being different from the domains included in the SPECIALIST lexicon. However, this is not very likely since we found that on the single-word level the patent domain does not contain many words that are not in the general English vocabulary. We conjecture that patent authors write claims in which they create novel NPs (not captured by terminologies such as SPECIALIST). This is also found by D'hondt (2009), who reports that "these [multi-word] terms are invented and defined ad hoc by the patent writers and will generally not end up in any dictionary or lexicon." [9]. This would confirm the introduction of novel terms by patent authors, but only with respect to multi-word terms.

4.3 Challenge 3: Syntactic structure

According to international patent writing guidelines, patent claims are built out of noun phrases instead of clauses (see Section 2). This can be problematic for patent processing techniques that are based on syntactic analysis. Syntactic parsers are generally designed to analyze clauses, not noun phrases. This means that if there is a possible interpretation of the input string as being a clause, then the parser will try to analyze it as such: In case of lexical ambiguity one of the words will be interpreted as finite verb whereas it should be a noun or participle.

An analysis of the output of the baseline version of the AEGIR parser on a subset of the MAREC corpus can provide insight into the challenges relating to syntactic structures that occur in patent claims. To this end, we created a small sample from the complete set of MAREC sentences: a random sample of 100 sentences that are five to nine words in length. The motivation for this short sentence length in the sample was twofold: On the one hand we wanted to capture most NP constructions that occur in patent claims but at the same time we wanted to minimize structural ambiguity.

Table 3: Evaluation of the baseline AEGIR parser and the state-of-the-art Connexor CFG parser for a set of 100 short (5–9 words) sentences from the MAREC subcorpus.

	AEGIR	Connexor CFG
precision	0.45	0.71
recall	0.50	0.71
F1-score	0.47	0.71
Inter-annotator agreement	0.83	0.83

For evaluation purposes, we manually created ground truth dependency analyses for 100 randomly selected sentences from this set. We found that only 4% of the short sentences are clauses (e.g. "F2 is the preselected operating frequency.").

The ground truth annotations were created by two assessors: both created annotations for 60 sentences, with an overlap of 20 sentences. We measured the inter-annotator agreement by counting the number of identical dependency triplets among the two annotators. Dividing this number by the total number of triplets created by one annotator gives $accuracy_1$, dividing the number by the total number of triplets created by the other annotator gives $accuracy_2$. We take the average accuracy as inter-annotator agreement.[13] This way, we found an inter-annotator agreement of 0.83, which is considered substantial.

For the 20 sentences that were annotated by both the assessors, a consensus annotation was agreed upon with the help from a third (expert) assessor. After that, we adapted the annotations of the 80 sentences that had been annotated by one of the two assessors in accordance with the consensus annotation. This resulted in a consistently annotated set of 100 sentences. We used these annotations to evaluate the baseline version of the AEGIR parser. We calculated precision as the number of correct triplets in the AEGIR output divided by the total number of triplets created by AEGIR, and recall as the number of correct triplets in the AEGIR output divided by the number of triplets created by the human assessor.

In order to compare the baseline version of the AEGIR parser to a state-of-the-art dependency parser, we ran the Connexor CFG parser [16] on the same set of short patent claim sentences. We converted the output of the parser to dependency triplets according to the AEGIR descriptive model[14] and then evaluated it using the same procedure as described for the AEGIR parser above. The results for both AEGIR and the Connexor parser are in Table 3.

Table 3 shows that the performance of the baseline version of the AEGIR parser on short patent sentences is still moderate, and lower than the state-of-the art Connexor parser.

The errors made by AEGIR can provide valuable insights in the peculiarities of patent language. The most frequent parsing mistakes made by AEGIR are (1) the wrong choice

[13]Cohen's Kappa cannot be determined for these data since there exists no chance agreement for the creation of dependency triplets.

[14]A one-to-one conversion was possible to a large extent. The only problematic construction was the phrasal preposition *according to*, which is treated differently by the Connexor parser and the AEGIR descriptive model.

for the head of a dependency relation (e.g. `[9,ATTR,claim]` for "claim 9" and (2) incorrect attachment of postmodifiers in NPs. For example, for the sentence "The method of claim 4 wherein n is zero.", the parser incorrectly generates `[method,PREPof,n]` instead of `[method,PREPof,claim]` and it labels *wherein* as a modifier to *n*: `[n,MOD,X:wherein]`.

The former of these errors is repeated frequently in the data: the regular expression "claim [0-9]+" occurs in 96% of the sentences in the MAREC subcorpus. The latter case (ambiguities caused by postmodifier attachment) is known to be problematic for syntactic parsing. In patent claims, however, the problem is even more frequent than in general English because the NPs in patent claims are often very long (recall the median sentence length of 22 words). This brings us back to the central syntactic challenge mentioned several times in this paper: patent claims are composed of NPs instead of clauses.

In order to find other syntactic differences between patent claims and general English, we plan to evaluate the baseline version of the AEGIR parser on a set of sentences from the BNC and compare the outcome to the results obtained for MAREC sentences (Table 3).[15]

5. CONCLUSIONS AND FUTURE WORK

We have analyzed three challenges of patent claim processing that are mentioned in the literature: (1) The length of the sentences is much longer than for general language use; (2) Many novel terms are introduced in patent claims that are difficult to understand; and (3) the structure of the patent claims is complex, as a result of which syntactic parsers fail to correctly analyze patent claims. Where possible, we supported our analyses with quantifications of the findings, using a number of (patent and non-patent corpora) and NLP tools.

With respect to (1), we verified that sentences in English patent claims are longer than in general English, even if we split the claims not only on full stops but also on semi-colons. The median sentence length in the MAREC subcorpus is 22 words; the average length is 53 words.

With respect to (2), we performed a number of analyses related to the vocabulary of patent claims. We found that at the level of single words, not many novel terms are introduced by patent authors. Instead, they tend to use words from the general English vocabulary, which was demonstrated by a token coverage of 96% of the CELEX lexicon on the MAREC subcorpus. However, the frequency distribution of words in patent claims does differ from that in general English, which can be especially seen from the list of top-frequency words from MAREC and BNC. Moreover, it seems that the authors of patent claims do introduce novel terms, but only at the multi-word level: we found that the lexicalized multi-word terms from the SPECIALIST lexicon have low frequencies in the MAREC subcorpus.

With respect to (3), we parsed 10,000 claims from the MAREC subcorpus using the baseline version of the AEGIR dependency parser and we performed a manual evaluation of the parser output for 100 short sentences from the corpus. We can confirm that syntactic parsing for patent claims is a challenge, especially because the claims consist of sequences of noun phrases instead of clauses while syntactic parsers are designed for analyzing clauses. As a result, the parser will try to label at least one word in the sentence a finite verb.

In conclusion, we can say that the challenges of patent claim processing that are related to syntactic structure are even more problematic than the challenges at the vocabulary level. The sentence length issue only causes problems indirectly by resulting in more structural ambiguities for longer noun phrases.

In the near future, we will further develop the AEGIR dependency parser into a hybrid[16] parser that incorporates information on the frequencies of dependency triplets. These frequencies (which are stored in the triplet database that is connected to AEGIR) guide the resolution of structural ambiguities. For example, the information that 'carbon atoms' is a highly frequent NP with the structure `[atom,ATTR,carbon]` guides the disambiguation of a complex NP such as "cycloalkyl with 5-7 ring carbon atoms substituted by a member selected from the group consisting of amino and sulphoamino" (taken from the MAREC subcorpus), which contains many structural ambiguities. The same holds for the frequent error `[9,ATTR,claim]` that we mentioned in Section 4.3. Given the high frequency of this error type, it is relatively easy to solve using triplet frequencies.

In order to collect reliable frequency information on dependency relations, we use a bootstrap process. As the starting point of the bootstrap we use reliably annotated corpora for general English such as the Penn Treebank [20] and the British National Corpus (BNC) [7]. We then use parts of patent corpora such as MAREC and CLEF-IP [23], which we annotate syntactically using automatic parsers. Moreover, we harvest terminology lists and thesauri such as the biomedical thesaurus UMLS [4], which contain many multi-word NPs and therefore can provide us with reliable ATTR relations (such as `[atom,ATTR,carbon]`).

The addition of this information allows us to tune the AEGIR parser specifically to the language used in patent texts. We expect that a number of the parsing problems described in this paper will be solved by incorporating frequency information that is extracted from patent data. To what extent this will be successful is to be seen from the further development and evaluation of the AEGIR parser.

6. ACKNOWLEDGMENTS

The TM4IP project is funded by Matrixware.

7. REFERENCES

[1] N. Akers. The European Patent System: an introduction for patent searchers. *World Patent Information*, 21(3):135–163, 1999.

[2] R. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA, 1993.

[3] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled

[15] Of course, we can expect some problems when we run a parser that is being developed for patent texts specifically to BNC data, such as the generation of the triplet `[Betty,ATTR,said]` for the last two words of "Oh , that is sad," said Betty.

[16] Hybrid in the sense that it combines rule-based and probabilistic information

corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[4] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.

[5] D. Bonino, A. Ciaramella, and F. Corno. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information*, 32:30–38, 2010.

[6] A. Browne, A. McCray, and S. Srinivasan. The Specialist Lexicon. *National Library of Medicine Technical Reports*, pages 18–21, 2000.

[7] L. Burnard. Users reference guide for the British National Corpus. Technical report, Oxford University Computing Services, 2000.

[8] L. Chen, N. Tokuda, and H. Adachi. A patent document retrieval system addressing both semantic and syntactic properties. In *Proceedings of the ACL-2003 workshop on Patent corpus processing*, pages 1–6, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[9] E. D'hondt. Lexical Issues of a Syntactic Approach to Interactive Patent Retrieval. In *The Proceedings of the 3rd BCSIRSG Symposium on Future Directions in Information Access*, pages 102–109, 2009.

[10] E. D'hondt, S. Verberne, N. Oostdijk, and L. Boves. Re-ranking based on Syntactic Dependencies in Prior-Art Retrieval. In *Proceedings of the Dutch-Belgium Information Retrieval Workshop 2010*, 2010. To appear.

[11] E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, and V. Zervaki. Towards content-oriented patent document processing. *World Patent Information*, 30(1):21–33, 2008.

[12] A. Fujii, M. Iwayama, and N. Kando. Introduction to the special issue on patent processing. *Information Processing and Management*, 43(5):1149–1153, 2007.

[13] E. Graf and L. Azzopardi. A methodology for building a test collection for prior art search. In *Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA)*, pages 60–71, 2008.

[14] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing and Management*, 42(1):207–221, 2006.

[15] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at NTCIR-3. In *Proceedings of the third NTCIR workshop on research in information retrieval, automatic text summarization and question answering*, 2003.

[16] T. Jarvinen and P. Tapanainen. Towards an implementable dependency grammar. In *The Proceedings of COLING-ACL*, volume 98, pages 1–10, 1998.

[17] C. Koster and J. Beney. Phrase-Based Document Categorization Revisited. In *Proceedings 2nd International Workshop on Patent Information Retrieval (PaIR'09)*, 2009.

[18] C. Koster, N. Oostdijk, S. Verberne, and E. D'hondt. Challenges in Professional Search with PHASAR. In *Proceedings of the Dutch-Belgium Information Retrieval workshop*, 2009.

[19] G. Leech. 100 million words of English: the British National Corpus (BNC). *Language Research*, 28(1):1–13, 1992.

[20] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1994.

[21] H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama, and T. Oshio. Proposal of two-stage patent retrieval method considering the claim structure. In *ACM Transactions on Asian Language Information Processing (TALIP)*, volume 4, pages 190–206, 2005.

[22] S. Mille and L. Wanner. Making text resources accessible to the reader: The case of patent claims. In *Proceedings of the International Language Resources and Evaluation Conference (LREC)*, pages 1393–1400, Marrakech, Morocco, 2008.

[23] G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *CLEF working notes 2009*, pages 1–16, 2009.

[24] S. Sheremetyeva. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 workshop on Patent corpus processing*, pages 66–73, 2003.

[25] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Patent claim processing for readability: structure analysis and term explanation. In *Proceedings of the ACL-2003 workshop on Patent corpus processing-Volume 20*, page 65. Association for Computational Linguistics, 2003.

[26] Y. Tseng, C. Lin, and Y. Lin. Text mining techniques for patent analysis. *Information Processing and Management*, 43(5):1216–1247, 2007.

[27] L. Wanner, R. Baeza-Yates, S. Brugmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, et al. Towards content-oriented patent document processing. *World Patent Information*, 30(1):21–33, 2008.