Genre and Domain in Patent Texts

Nelleke Oostdijk
CLST & IFL
Radboud University Nijmegen
n.oostdijk@let.ru.nl

Eva D'hondt
CLST & IFL
Radboud University Nijmegen
E.dHondt@let.ru.nl

Hans van Halteren
Dept. of Linguistics
Radboud University Nijmegen
hvh@let.ru.nl

Suzan Verberne
CLST & IFL
Radboud University Nijmegen
s.verberne@let.ru.nl

ABSTRACT

In this paper we investigate the variation in language use within the very broad patent domain. We find that language use (represented by syntactic phrases) not only differs from one patent class to the next, but is also a characteristic that sets apart the four sections of a patent (viz. Title, Abstract, Description and Claims). This lends support to the claim that these sections can be viewed as different text genres.

For the development of a syntactic parser that is trained on patent texts, we quantify the domain and genre differences in terms of the amounts of text needed to train domain-dependent versions of the parser.

Our quantified and exemplified findings on the domain variation in patent data are of interest for the patent retrieval and analysis communities.

Keywords
Parsing, Patent domains, Text mining

1. INTRODUCTION
Keyword retrieval based on a bag-of-words model generally gives a high recall with the price of a possibly low precision. This is especially problematic in collections where subsets of documents share many terms since they cover the same topic domain. For that reason, patent search is a field where finding the relevant documents (prior art) for a given topic (patent application) requires a document representation that is more specific than the bag of words.

To distinguish more clearly between related patents, it can be advantageous to take into account more precise terms in addition to the bag of words. These terms can be statistical collocations or syntactic phrases. The literature shows mixed results on the additional value of such phrases for retrieval and classification tasks (for an overview, see [1]). Recently, good results have been reported on the use of aboutness-based dependency triples for patent classification [2]. An aboutness-based dependency triple is a pair of two content words and the dependency relation between them [3]. In order for a search system to have access to this information, all patent texts must be parsed syntactically.

Parsing patent documents is not an easy task as the language used in patent documents, and especially the claims section, is extremely complex. Previous research has shown that this complexity is largely due to the use of very long and complex grammatical constructions and the abundance of technical terms [4] [5] [6].

General purpose parsers such as the Charniak parser [7] and the Machinese parser [8] appear less suited for handling patent text for a number of reasons: they lack the necessary lexical and terminological resources to cover the wide range of topics addressed in patent documents and they have difficulty coping with the syntactic complexity of the language used, which results in a high degree of structural ambiguity. Add to this the fact that the output is unnecessarily detailed for the purpose of patent search, we find it appropriate to develop a parser specifically aimed at patent texts. This is one of the challenges we have taken up in the Text Mining for Intellectual Property (TM4IP) project.

In this project, we aim at developing a text mining system for interactive patent search. The basis for this text mining system is the AEGIR[1] dependency parser. The AEGIR parser is designed in such a way that it can benefit from previously accumulated knowledge that is available in the form of a database of observed dependency triples and their frequencies. The triple frequencies guide the choice between the many possible analyses of a sentence in the text, yielding the one most probable parse as output.

The accuracy of syntactic parsers is known to highly depend on the text domain they are trained and evaluated on (see for example [9,10]). In the context of the approach taken by the AEGIR parser, a crucial question is to what extent a database of dependency triple observations is domain specific. This question is particularly relevant in the AEGIR context as it will predict the potential effectiveness and reusability of triple databases. For similar reasons, it is worthwhile to investigate the intuitive idea that there are strong differences in language use in the four different sections of a patent (viz. Title, Abstract, Description and Claims). Triple databases that are representative of the language occurring in different domains and sections of patent texts will contribute to a better performance of the parser (also according to [19], p. 243)

With the development of resources for the AEGIR parser as intrinsic motivation for a series of analyses, the goal of this paper is to quantify and exemplify the domain differences in a large

---

[1] AEGIR stands for Accurate English Grammar for Information Retrieval.

patent corpus. We show the patent research community the importance of domain-specific resources for a range of domains.

After a more detailed explanation of our parsing approach (Section 2) and our experimental data (Section 3), we show that each of the four sections in a patent document represents a separate genre (Section 4). We then examine the similarity of triple frequency counts within (Section 5) and between (Section 6) domains and genres. Section 7 shows that different types of dependency triples also behave differently as to their similarity between domains and genres. In Section 8 we present our conclusions and plans for future work.

2. THE AEGIR PARSER IN TM4IP

In the TM4IP project, a text mining system is being implemented for intellectual property search [11]. The system consists of (1) an English hybrid dependency parser (AEGIR) that is especially developed to deal with the syntactic characteristics and vocabulary of technical texts, more specifically patent texts, and (2) a professional interactive search engine that uses structured queries based on dependency relations to find patent documents in a database [12].

The AEGIR parser combines a set of handwritten syntactic rules with an extensive word form lexicon ('the parser lexicon') and information about the frequencies of dependency relations between words. This information is stored in a database of dependency triples ('the triple database') and is consulted during the parsing process.

The grammar uses an aboutness-based dependency model (cf. [13,14]). During parsing, normalization is applied at various levels, viz. the level of typography (normalizing for example the use of upper and lower case, spacing etc.), spelling (e.g. British and American English, hyphenation), morphology (the lemmatization of word forms) and syntax (the standardization of the word order, e.g. active/passive transformation). This makes the parser (and hence the search interface) robust for linguistic variation between the textual content of the relevant documents and the user's query. E.g. "aspirin causes gastric bleeding" gets the same representation as "gastric bleeding can be caused by aspirin".

The parser generates dependency graphs for English input texts. From these graphs, sets of dependency triples (DTs) can be extracted [15]. A DT is represented as $[TERM_1,RELATOR,TERM_2]$ where a *term* is the lemma of a content word and a *relator* one of a fixed inventory of typed relations that holds between the two terms. The most important relators and the relations they denote are presented in Table 1.

Table 1. Types of dependency relations

Relator	Relation
SUBJ	Term 1 is the subject of term 2
OBJ	Term 2 is the object of term 1
PRED	Term 2 is the predicate of term 1
ATTR	Term 2 is the attribute of term 1
MOD	Term 2 modifies term 1
PREP	Term 2 is connected to term 1 through a preposition
THAT	Term 2 is the head of a subordinate clause attached to term 1

The AEGIR parser can be applied using just the rule-based core, but is intended to be used in its hybrid version. The grammar is then linked to a database in which frequency counts of DTs have been stored. During the parsing process, the parser uses the observed relative counts of any potentially present DTs to determine the relative likelihood of the various analyses of structurally ambiguous sentences. For example, in parsing the sentence *We present a new device for the distillation of vinegar,* the ambiguity of the PP-attachment is resolved in favour of 'device for distillation', as the observed frequency of the triple [device,PREPfor,distillation] is higher than that of [present,PREPfor,distillation].

At this moment in time, we are not yet able to implement a sufficiently large experiment for measuring how exactly patent search would improve with an increasingly representative triple database. For this paper, therefore, we measure the quality of a DT database in terms of its representativeness for a target text type. Since a DT Database is a count vector over DTs, and hence comparable to word vectors for documents, and database similarity is comparable to document similarity, we adopt from the field of information retrieval the cosine similarity as a measure for measuring DT database similarity. In order to measure DT database representativeness, we extract two DT databases for two non-overlapping samples of text from the same target text type. If these two DT databases have a cosine similarity of 0.95, then we consider their representativeness – and hence their quality – to be sufficient. We vary the size of the text samples to see what amount of text we need for a sufficiently representative DT database.

3. DATA

3.1 Patent Text

A patent, in the definition of the European Patent Office (EPO), is a legal title granting its holder the right to prevent third parties from commercially exploiting an invention without authorization.[2] Inventions can belong to any field of technology. Patents kept by the EPO are sorted into technical domains which are defined in the International Patent Classification (IPC).[3]

Patent documents are structured according to a prescribed format and follow a set of precise syntactic, lexical and stylistic guidelines, which differ from one legislation to another. According to the EPO guidelines[4] each complete patent application should contain four sections: a title, an abstract, a description and a claims section. Although a title may have up to 250 characters, it must be as short and specific as possible. The abstract points out what is new in the application compared to prior art. It should be in narrative form and generally limited to a single paragraph. An abstract should preferably be no longer than 150 words. However, in our data set, we noticed that quite often patent attorneys exceed this advised limit. The longest abstract in the set contained 394 words. In the description section, the invention must be explained along with the process of making and using the invention in full, clear, concise and abstract terms. Finally, the claims section includes one or more claims which define the scope of protection

---

of the patent in terms of technical features. Each claim takes the form of an extensive noun phrase.

The IPC classes represent what in linguistics are commonly referred to as domains. They are typically subject or topic oriented. Texts occurring within one and the same domain share the same topic which is reflected in the use of the same vocabulary (words, tokens, lemmas, etc). The different sections in a patent can be viewed as different linguistic genres. Each genre is characterized by its style, form and communicative function [16].

3.2 Selection
We selected our data from the MAREC-400,000 Corpus[5]. We only included English language patent documents from the EPO. We chose the intermediate classification level (level 3) in the IPC in order to balance conceptual coherence and number of patents. To go lower in the classification tree (max. level 5) would result in tighter domains but would not yield enough patent documents for our experiments. We selected an equal number of documents (1101, which is the number of documents in the smallest domain) from three different subclasses in the corpus, viz.

1. Semiconductor devices; electric solid state devices not otherwise provided for (IPC subclass H01L);

2. Preparations for medical, dental, or toilet purposes (IPC subclass A61K);

3. Electric digital data processing (IPC subclass F06G).

Only granted patent documents were included so as to avoid duplication and therefore bias in the selected data and to ensure that we included complete documents (containing all four sections).

3.3 Extracting and preparing the data

3.3.1 Data extraction
For each patent document, we extracted the text for the four different sections (using the XML tags in the MAREC Corpus) and the text for each section was saved in a separate file. After removing the remaining XML markup, we also cleaned up certain character conversion errors.

3.3.2 Sentencing
Using the module *Sentence* from the Perl package Lingua::EN[6], we split the text into sentences. The distribution of the number of sentences over the four sections and three domains is shown in Table 2.

Table 2. Number of sentences per section per domain

	H01L	A61K	F06G
Title	1,101	1,101	1,101
Abstract	3,823	2,471	4,542
Description	192,358	311,725	289,201
Claims	18,900	22,884	21,435

---

[5] MAREC stands for Matrixware Research Collection, which is a collection of patent documents. Matrixware supplied 400,000 documents from this collection for use in the AsPIRe'10 workshop.

[6] http://search.cpan.org/~shlomoy/Lingua-EN-Sentence-0.25/lib/Lingua/EN/Sentence.pm

3.3.3 Parsing
Many sentences were found to be extremely long, especially in the descriptions and claims sections where sentences up to 3684 and 5089 words occurred. Therefore, in a second preprocessing step, sentences containing semicolons were split on the semicolon, thereby reducing the average length of the input units for the parser. Table 2 represents the statistics for full sentences, before further splitting, as do all further tables below.

For the purpose of the present paper, the data was parsed by means of a development version of the AEGIR parser, using only the rule-based core. In case of multiple parses, only the first parse was taken into consideration. The first parse results from an intuitive ordering of the rules in the grammar underlying the parser and is overall already of quite good quality.

4. GENRES IN PATENT TEXTS
In Table 2, we already listed the number of sentences for the various sections of the patents in the three domains represented in our dataset. In this section we aim to prove that these sections can actually be considered different genres, using measures commonly used in corpus linguistic research. If patent sections can indeed be considered different genres then these genre differences deserve attention in the development of linguistic tools (such as syntactic parsers) for text mining.

Previous research (e.g. [17]) has identified three variables for genre characteristics: sentence length, type/token ratio (TTR, the ratio between the number of unique terms and the total number of terms in a corpus) and hapax ratio (HR, the proportion of terms that only occur once in a corpus [18]). It has been suggested that in formal written genres where there is a high density of information, texts are characterized by relatively long sentences and high TTRs and HRs, reflecting careful phrasing which involves precise word choice and an exact presentation of informational content. While the high TTRs are indicative of the use of many different words, a high proportion of unique words (HR) points to a lack of repetition.

For our data we calculated the mean sentence length for the sentences as they were obtained from sentencing. Both the TTR and HR were calculated, taking the DTs as basic units instead of words.

Table 3 shows the mean sentence length and Tables 4 and 5 the TTR and HR for the different genres and domains. The differences between the genres are significant in all three tables ($p < 0.05$, ANOVA). This confirms our assumption concerning the existence of four different genres.

Table 3. Mean sentence length in number of words and DTs

	H01L		A61K		F06G	
	wrds	DTs	wrds	DTs	wrds	DTs
Title	7.99	5.24	7.76	4.94	7.98	5.05
Abstract	36.12	26.07	36.24	24.61	32.24	22.26
Description	32.29	21.70	30.38	18.62	29.96	21.14
Claims	47.03	29.80	37.02	23.25	59.11	37.47

Table 4. Triple type/token ratio (TTR)

	H01L	A61K	F06G
Title	0.74	0.82	0.80
Abstract	0.54	0.58	0.58
Description	0.24	0.27	0.25
Claims	0.24	0.23	0.23

Table 5. Triple hapax ratio (HR)

	H01L	A61K	F06G
Title	88.56	91.91	90.25
Abstract	78.99	81.69	79.31
Description	66.90	70.16	65.98
Claims	56.27	59.30	52.88

While the TTR is useful for measuring the triple coverage, in the light of the issues addressed in this paper, it is not the best measure: what a measure such as TTR fails to bring out is the more specific distribution, which especially for the high frequency items is of particular interest in view of the role envisaged for the triple count database. Therefore, we also calculated the triple entropy (cf. Table 6).

Table 6. Triple entropy

	H01L	A61K	F06G
Title	12.3	12.2	12.2
Abstract	16.0	15.4	16.0
Description	20.0	20.8	20.4
Claims	17.0	16.8	17.2

As is apparent from Table 6, the Description is by far the most varied. The other sections are clearly more restricted in language use, as can be expected given their prescribed role in the patent.

# 5. AMOUNT OF TEXT NEEDED FOR REPRESENTATIVENESS

In this section, we investigate how many words of patent text would be needed to yield a DT Database that is stable enough to be considered representative of its text type (see Section 2). We will do this for domain H01L and separately for each genre G.[7] For N varying from 1 to 512, we repeatedly (100 times) draw two random patent sets PSx and PSy of N patents such that, a) PSx and PSy contain only text from H01L and b) PSx and PSy are disjoint. Then we derive DT Databases DBx and DBy from those sections of PSx and PSy that belong to genre G, and measure the cosine similarity between DBx and DBy. We calculate the similarity at N as the mean over the 100 measurements. If this number is higher than 0.95 for a given N and G, we assume that DT Databases are stable enough at N that they can be seen to represent the text in genre G.

We present our measurements in Figure 1, plotting the cosine similarity for the individual genres. The x-axis does not show N, but rather the number of DTs represented in the various databases, since the amount of text in a patent in the four genres differs considerably.

---

[7] In view of the genre differentiation (cf. Section 4), we will treat each genre independently.

We observe that, with the amount of data we have available, only the description genre shows a similarity over the 0.95 threshold at N=512 and we do apparently not have enough patents to reach this convergence for the other genres. This means that, for these other genres, in any comparisons below we should only consider relative values and not overall sufficiency. However, for the current question we can observe that the curves in Figure 1 are almost perfectly parallel and, on the basis of this observation, we can extrapolate the leftmost three curves to the point where the cosine would reach 0.95. This allows us to roughly estimate the number of DTs (and hence patents/words) that would be needed for representativeness. These estimates are listed in Table 7. Obviously, these estimates may vary from domain to domain but we expect H01L to be fairly representative.

Table 7. Estimated amount of data needed to build reliable domain specific DT Databases

	# PATENTS	# WORDS	# DTS
Title	4,000	30,000	20,000
Abstract	1,700	375,000	150,000
Description	400	3,750,000	1,500,000
Claims	800	600,000	400,000
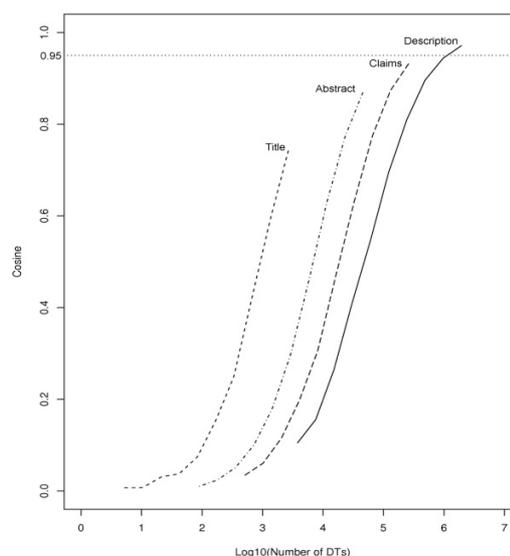
Figure 1. Convergence of the within-genre DT Database similarity in D1 when increasing the number of patents from which the databases are derived from N=1 to 512

6. SPECIFICITY OF DT COUNT DATABASES

In this section, we investigate to what extent DT Databases are specific for given domains (Section 6.1) and genres (Section 6.2), by measuring the similarity between two such databases based on the same or different text types.

6.1 Differences between Domains

We first examine to which degree DT Databases differ between various patent domains. In order to measure the similarity of

databases for genre G for domain sets DSx and DSy,[8] we adapt the procedure described in Section 5. We repeatedly (100 times) draw two random patent sets PSx and PSy of N patents such that, a) PSx contains only text from DSx and PSy contains only text from DSy and b) PSx and PSy are disjoint. Then we derive DT databases DBx and DBy from those sections of PSx and PSy that belong to genre G, and measure the cosine between DBx and DBy. We again characterize the similarity between DSx and DSy with the mean over the 100 measurements. For the comparisons between various DSx and DSy, we set N at 512.

In Table 8 we show the average within- and cross-domain-set similarities (and their standard deviation between brackets) for DS={H01L} and DS ={F06G}. As we already observed in Section 5, only the Description genre reaches our chosen target similarity of 0.95 and the differences in amounts of DTs in the various sections prevent cross-genre comparisons. However, looking at the relative values within each genre, we can see that cross-domain similarity is far below within-domain similarity, indicating that there are indeed substantial differences between domains.

Table 8. Mean (top number) and standard deviation (bottom number, between brackets) for cosine similarity within and between specific domain (D) databases

	H01L / H01L	H01L / F06G	F06G / F06G
Title	0.744 (0.009)	0.102 (0.016)	0.487 (0.014)
Abstract	0.872 (0.004)	0.331 (0.011)	0.815 (0.004)
Description	0.971 (0.002)	0.603 (0.011)	0.966 (0.002)
Claims	0.932 (0.004)	0.596 (0.012)	0.925 (0.004)

We also repeated our procedure with a specialized DS={H01L} and a generic DS={H01L,A61K,F06G}, with N again set at 512. Note that H01L is itself included in the generic set, with only two other domains, so that the similarity between H01L and Generic will be a rather high estimate of the similarity between H01L and the combination of all 629 patent subclasses. The results are shown in Table 9. Again, the cross-DS similarity is far below the within-type similarity.[9]

The difference between domains can be exemplified with the most frequent content triples. Table 10 on the last page shows the ten most frequent triples from H01L containing only content words, together with their frequency in domain F06G. For most of these triples, namely those referring to concepts from the invention domain H01L, the rank and count in H01L differs considerably from the rank and count in F06G, as can be expected. However, we also observe that the top ten contains four triples (1, 3, 6 and 9, printed in boldface) that are representative of patents text in general rather than for the specific domain, and therefore rank similarly high in F06G.

Table 9. Mean (top number) and StdDev (bottom number, between brackets) for cosine similarity between mixed and specific domain databases

	H01L / H01L	Generic / H01L	Generic / Generic
Title	0.744 (0.009)	0.435 (0.040)	0.389 (0.029)
Abstract	0.872 (0.004)	0.655 (0.022)	0.770 (0.008)
Description	0.971 (0.002)	0.819 (0.019)	0.952 (0.005)
Claims	0.932 (0.004)	0.779 (0.024)	0.878 (0.031)

6.2 Differences between Genres

In a similar fashion, we investigated the differences between genres. We repeated our procedure again, with N at 512, but this time keeping the domain constant and varying the genre. As we already saw from the measurements for Title and Abstract the constructed databases are (due to their smaller sizes) not representative enough to be compared to each other (see Section 5), we restricted ourselves to the Description (DSC) and Claims (CLM) genres. However, we measured within and across genre similarity for these two genres for all three domains. The results are shown in Table 11. As we can see, the differences between genres are even larger than those between domains.

Table 11. Mean (top number) and StdDev (bottom number, between brackets) for cosine similarity within and between specific Genre databases

	DSC-DSC	DSC-CLM	CLM-CLM
H01L	0.971 (0.002)	0.454 (0.008)	0.932 (0.004)
A61K	0.967 (0.006)	0.342 (0.014)	0.899 (0.013)
F06G	0.966 (0.002)	0.467 (0.009)	0.925 (0.004)

Again, we exemplify the differences with the most frequent content triples. Tables 12 and 13 on the last page list the top-ten triples in the description and claims genres with the corresponding ranks and relative counts in both genres. The picture that emerges is slightly more subtle than the one observed between domains. Two triples in the top 10 occur just as frequently in both genres in the domain. These are the domain-dependent, genre-independent ones that make up the terminology of the domain. Interestingly, we also find partial domain-dependent DTs, that is DTs which denote the same content but differ in the way the content matter is addressed. In the description, the triples directly represent the content, e.g. *reaction mixture*. In the claims, we see an overlay of claim specific language use, e.g. *said mixture*. Finally, there are the almost completely genre-determined triples, such as *embodiment of invention, preferred embodiment* and *shown in fig(ure)* in the description, which rank very low or do not even appear in the claims and, and *claim 1, method of claim*, etc. in the claims, which are similarly not frequently present in the description sections.

7. DIFFERENTIATION OF DTS

In our measurements in the previous sections we considered all the DTs to be of equal importance, except that more frequent DTs probably provided a larger contribution to the cosine. However, we already saw in the examples in Tables 10, 12 and 13 that there is a differentiation in DTs, and that the differences between text types are much smaller for some types of DTs than for others.

7.1 Different Relations

We investigated whether some relation types (ATTR, SUBJ, OBJ, etc.) are more frequent in one domain or genre than other. We found that there are no large differences between the occurrences of relation types among domain or genres: Apart from the titles, all genres and domains show a similar pattern of relation use (Tables 14 and 15). Therefore, we expect that differentiation according to the dependency relation in the triple database does not have much effect.

Table 14. Percentages of different relations in the various genres (Ttl=Title, Abs=Abstract, Dsc=Description, Clm=Claims)

	TTL	ABS	DSC	CLM
SUBJ	15.6	19.5	18.6	17.6
OBJ	15.8	13.1	11.7	11.6
PRED	0.5	4.0	2.9	3.1
ATTR	41.7	22.3	24.2	25.1
MOD	2.2	4.9	5.7	5.9
PREP	15.0	14.5	15.3	16.0
THAT	0.2	0.9	1.0	1.0

Table 15. Percentages of different relations in the various domains

	H01L	A61K	F06G
SUBJ	18.0	18.0	19.4
OBJ	11.5	10.8	12.7
PRED	2.4	4.1	2.3
ATTR	24.8	24.2	24.0
MOD	6.0	5.7	5.3
PREP	15.4	16.5	14.3
THAT	0.9	0.6	1.3

7.2 More and Less Common DTs

A more useful differentiation is to categorize the DTs in terms of the fraction of patents they occur in. We have already seen in the examples above that some DTs behave in a rather similar manner across domains. We therefore again borrow from the field of IR and use the IDF (inverse document frequency) measure, in our case inverse patent frequency (IPF) of DTs. We created IPF bands by using only the number before the decimal point, e.g. [film,ATTR,insulating] has an IPF of 4.27 and we assign it to IPF 4. We then repeated the calculation for Table 8 for each IPF band individually (only for descriptions). The results are shown in Table 16. As we see, the between-domain similarity decreases sharply when going to higher IPF bands. Especially IPF 1 (i.e. DTs occurring in more than 25% of the patents) shows a remarkably

high similarity, which suggests that the corresponding DTs can be considered to be a sort of 'universal'[10] DTs.

Table 16. Mean and StdDev for cosine similarity for the description genre within and between specific domain databases, differentiated as to IPF band

	H01L - H01L	H01L - F06G	F06G - F06G
IPF 1	0.997 (0.001)	0.910 (0.007)	0.996 (0.002)
IPF 2	0.988 (0.004)	0.374 (0.015)	0.988 (0.002)
IPF 3	0.960 (0.007)	0.310 (0.011)	0.968 (0.005)
IPF 4	0.897 (0.011)	0.247 (0.011)	0.915 (0.009)
IPF 5	0.798 (0.011)	0.177 (0.009)	0.803 (0.011)
IPF 6	0.566 (0.012)	0.117 (0.006)	0.612 (0.013)
IPF 7	0.292 (0.011)	0.062 (0.005)	0.321 (0.010)
IPF 8	0.082 (0.004)	0.018 (0.001)	0.094 (0.018)

Closer inspection reveals a set of 45 DTs which all appear with relatively high frequencies in the abstracts, claims and descriptions in all domains. They seldom appear in titles. Examples are [invention,ATTR,present], [embodiment,PREPof, invention], [be,MOD,for_example] and [#IT#,SUBJ,comprise][11] Although universal DTs constitute a small set, their contribution in the parsing process may be relatively high in view of their high frequencies.

The type of DTs we see in the set of universal DTs also leads us to speculate that there are DTs which are universal, except for the fact that one of the two terms that participate in the relation is domain- or even patent-specific. We have already seen an example of this in the relation [*,ATTR,said] which is found very frequently in the claims. One could imagine, then, that if we could determine that a term belongs to a recognized syntactic-semantic class (e.g. 'invention' or 'method') we could make use of this information and create a kind of abstracted (underspecified) DTs, whose counts can be used as a fallback option for fully specified DTs which are rare or absent.

8. CONCLUSIONS AND FUTURE WORK

In the current research project, we aim at developing a text mining system for interactive patent search. The core of the text mining system is a robust, aboutness-based dependency parser that is especially geared towards patent texts. The parser combines a rule-based component with frequency information on the dependency relations between words (dependency triples).

---

[10]This term is used here to refer to DTs that occur in patent texts irrespective of domain.

[11] #IT# is a placeholder term for the implied agent in passive constructions.

Given the known influence of domain-dependency on syntactic parsing, our parser will give better performance when featured with a triple database that is optimally representative for the texts to be parsed. Since the patent domain is known to be very diverse, we have investigated the variation in language use within the patent domain. Our findings on the domain variation in patent data is of interest for the patent retrieval and analysis communities.

In Section 4, we found that, apart from the domains already identified by the EPO/IPC subclasses, we should also distinguish between four genres, connected to the four prescribed sections of patent documents (title, abstract, description, claims).

We experimented with triple databases based on different amounts of text (Section 5). We found that, since the average lengths of the texts highly differs between the four sections, very different amounts of documents are needed for creating a triple database that is representative for a genre.

In Section 6, we measured the similarity between triple databases extracted from different patent subclasses and we found that the between-domain similarity is far below the within-domain similarity for disjoint text samples. Thus, we expect our parser to be helped by domain-specific and genre-specific databases. This presents a potential problem, since the number of IPC subclasses is 629. Therefore, we will in the near future investigate the re-use of triple databases among subclasses from the same main class.

Another way to overcome the problem of hundreds of domain- and genre-specific triple databases, is the use of a generic database that is restricted to triples that show similar frequency over all domains and all genres, possibly extended by abstracted forms of triples that also have this property (Section 7).

Currently, we are working on a reference set of patent sentences for parser evaluation. In future work, we will use this reference set for evaluating the quality of our parser with different triple databases. As baselines, we consider first the rule-based parser without triple database, and second the parser together with a database of generic triples. Such an experiment will show the impact of triple frequency differences on the parsing process.

In general, for future work in linguistically-motivated approaches to patent search, it is important to take into account the large differences between domains and genres in patent corpora.

9. ACKNOWLEDGEMENT

10. REFERENCES

[1] Cornelis Koster and Jean Beney. 2009. Phrase-based Document Categorization Revisited. In *Proceedings CIKM 2009*, pages 49-55.

[2] Cornelis Koster and Jean Beney. 2009. Phrase-based Document Categorization Revisited. In *Proceedings CIKM 2009*, pages 49-55.

[3] Eva D'hondt, Suzan Verberne, Nelleke Oostdijk, and Lou Boves. 2010. Re-ranking based on Syntactic Dependencies in Prior-Art Retrieval. In *Proceedings of the Dutch-Belgium Information Retrieval Workshop* 2010.

[4] Noriko Kando. 2000. The NTCIR Workshop: an evaluation of Asian language information retrieval. Presented at RIAO '2000.

[5] Svetlana Sheremetyeva 2003. Natural Language Analysis of Patent Claims. *Proceedings of the workshop "Patent Corpus Processing" in conjunction with 41st Annual Meeting of the Association for Computational Linguistics* (ACL 2003), Sapporo. Japan, July 7-12.

[6] Suzan Verberne, Eva D'hondt, Nelleke Oostdijk, Cornelis Koster. 2010. Quantifying the Challenges in Parsing Patent Claims. *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010)*, pp. 14-21.

[7] Eugene Charniak. 1995. Parsing with Context-Free Grammars and Word Statistics. *Technical Report* CS-95-28, Dept. of Computer Science, Brown University.

[8] *http://www.connexor.eu/technology/machinese/demo/syntax/*

[9] Satoshi Sekine. The domain dependence of parsing. In *Proceedings of the Ninth Conference on Innovative Application of Artificial Intelligence (IAAI-97)*, Providence, RI, pages 96-102. AAAI Press/The MIT Press, July 1997.

[10] Pyysalo S, Ginter F, Pahikkala T, Boberg J, Järvinen J, Salakoski T: Evaluation of Two Dependency Parsers on Biomedical Corpus Targeted at Protein-Protein Interactions. *Recent Advances in Natural Language Processing for Biomedical Applications*, special issue of *the International Journal of Medical Informatics* 2006 , 75(6):430-442.

[11] Cornelis Koster, Nelleke Oostdijk, Suzan Verberne, and Eva D'hondt. 2009. Challenges in Professional Search with PHASAR. *Proceedings of DIR 2009*: 101-102.

[12] Cornelis Koster, Marc Seutter, and Olaf Seibert. 2006. The Phasar Search Engine. In *Proceedings NLDB 2006.* Springer LNCS 3999: 141-152.

[13] Igor Mel'čuk. 2009. Dependency in Natural Language. In Alain Polguère and Igor Mel'čuk (Eds.), *Dependency in Linguistic Description*: 1-110. Amsterdam/ Philadelphia: John Benjamins.

[14] Peter Bruza and Theo Huibers. 1996. A Study of Aboutness in Information Retrieval. *Artificial Intelligence Review*, 10: 1-27

[15] Cornelis Koster, Marc Seutter, and Olaf Seibert. 2007. Parsing the Medline Corpus. In *Proceedings RANLP 2007:* 325-329.

[16] Guillaume Cleuziou and Céline Poudat. 2007. On the Impact of Lexical and Linguistic Features in Genre- and Domain-Based Categorization. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing* (Mexico City, Mexico, February 18-24, 2007). A. Gelbukh. Ed. Lecture Notes in Computer Science, 4394. Springer-Verlag, Berlin, Heidelberg, 599-610.

[17] Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

[18] Michael Oakes. 1998. *Statistics for Corpus Linguistics. Edinburgh*. Edinburgh University Press.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8, 243-257.

Table 10. Rank of the top ten content DTs in a DT Database for domain H01L, over all genres, and the count per one million words

Rank in H01L	Frequency of occurrence per Mw in H01L	DT	Rank in F06G	Frequency of occurrence per Mw in F06G
1	**3119.9**	**[invention,ATTR,present]**	**1**	**1997.7**
2	1319.1	[device,ATTR,semiconductor]	321	60.9
3	**953.7**	**[embodiment,PREPof,invention]**	**2**	**807.4**
4	694.6	[form,OBJ,film]	9288	6.9
5	600.4	[film,ATTR,thin]	10779	6.3
6	**567.8**	**[show,PREPin,fig]**	**18**	**246.9**
7	536.6	[substrate,ATTR,semiconductor]	20216	3.8
8	526.0	[electrode,ATTR,gate]	11178	6.1
9	**477.6**	**[claim,ATTR,1]**	**12**	**294.3**
10	470.9	[surface,PREPof,substrate]	9685	6.7

Table 12. Rank of the top ten content DTs in a DT database for Description, and the count per one million words

Rank in DSC	Frequency of occurrence per Mw in DSC	DT	Rank in CLM	Frequency of occurrence per Mw in CLM
1	2794.3	[invention,ATTR,present]	5947	11.6
2	726.2	[embodiment,PREPof,invention]	258092	0.5
3	377.7	[embodiment,ATTR,preferred]	—	NOT PRESENT
4	**303.9**	**[compound,PREPof,formula]**	**8**	**800.5**
5	**303.1**	**[device,ATTR,semiconductor]**	**7**	**845.3**
6	289.8	[temperature,ATTR,room]	3202	17.9
7	288.1	[show,PREPin,fig]	91354	1.0
8	265.6	[acid,ATTR,amino]	65	319.1
9	239.9	[invention,SUBJ,provide]	31281	3.2
10	238.7	[mixture,ATTR,reaction]	1974	25.8

Table 13. Rank of the top ten content DTs in a DT Database for Claims, and the count per one million words

Rank in CLM	Frequency of occurrence per Mw in CLM	DT	Rank in DSC	Frequency of occurrence per Mw in DSC
1	4034.6	[claim,ATTR,1]	5686	7.8
2	1789.5	[comprise,MOD,further]	240	60.6
3	1618.9	[method,PREPof,claim]	62656	1.2
4	1390.7	[comprise,OBJ,step]	195	68.7
5	891.1	[group,SUBJ,consist]	81	108.7
6	875.3	[acceptable,MOD,pharmaceutically]	21	179.6
7	**845.3**	**[device,ATTR,semiconductor]**	**5**	**303.1**
8	**800.5**	**[compound,PREPof,formula]**	**4**	**303.9**
9	771.5	[select,PREPfrom,group]	103	95.5
10	35.8	[mixture,ATTR,said]	35273	1.9