French and German Corpora for Audience-based Text Type Classification

A. Todirascu*, S. Padó†, J. Krisch‡, M. Kisselew‡, U. Heid‡

*LiLPA, Université de Strasbourg, † ICL, Universität Heidelberg, ‡ IMS, Universität Stuttgart
todiras@unistra.fr, pado@cl.uni-heidelberg.de, {krischjr,kisselmx,heid}@ims.uni-stuttgart.de

Abstract

This paper presents some of the results of the CLASSYN project which investigated the classification of text according to audience-related text types. We describe the design principles and the properties of the French and German linguistically annotated corpora that we have created. We report on tools used to collect the data and on the quality of the syntactic annotation. The CLASSYN corpora comprise two text collections to investigate general text types difference between scientific and popular science text on the two domains of medical and computer science.

Keywords: audience-based text type, features for text categorization, text extraction

1. Introduction

Studies in text classification tend to concentrate on topics and domains (Sebastiani, 2005), while there is less work on text type (Santini, 2007; Poudat et al., 2006) or audience-related classification. Some genre-specific corpora exist, mostly for English, including chats, blogs, tweets or emails, but to our knowledge, there are no corpus resources for German in this area. For French, we mention Scientext corpus (Tutin, 2010), composed exclusively of scientific texts.

Many NLP applications (such as text simplification, search engines or web content generation) have a need for text type classification with regard to audience, genre, or text type. Another example is the provision of specialized terminology, where typically texts of a high degree of specialization are more adequate than e.g. popular science texts. Although there are focused crawlers such as de Groc (2011) which collect relevant texts based on domain-specific seed words, such tools do not include any text type-related filtering.

The CLASSYN project has investigated the task of text classification with regard to these parameters rather than the traditional topic or domain-based distinctions. We assess to what extent texts aimed at certain audiences and for certain functions are characterized by (morpho-)syntactic properties that distinguish them from texts for another audience and/or for another function. Typical examples of such text types are scientific as opposed to popular science articles. Examples of such differences are the complexity of subjects and objects (higher for scientific text) or the frequency of second person pronouns (higher for popular science text). Some research projects focus on the use of some linguistic information (POS tags or lemmas) to improve text type classification (Charnois et al., 2008), (Stamatatos et al., 2000), (Karlgren and Cutting, 1994), but few systems exploit morpho-syntactic features in detail.

We are also interested in the extent to which such (morphosyntactic) properties are stable across domains and possibly even across languages: do e.g. popular science articles from different domains share properties (such as the predominance of second person pronouns) that allow us to recognize them and to distinguish them from scientific writing for specialists? This would constitute a clear advantage over the type of features usually employed for topic-based text classification, namely n-grams, are lexical and thus domain- and language-specific (Sebastiani, 2005).

To investigate these issues, we have (i) collected and analysed corpora; and (ii) conducted classification experiments. This paper concentrates on the first of the two activities.

Section 2 discusses the concept of "text type" and presents the analysis framework. Section 3 explains the design principles underlying the CLASSYN corpora. Section 4 presents the tools used for corpus gathering and discusses the quality of dependency parsing obtained on the different parts of the corpus. Section 5 analyses the actual text type differences we find in each language and domain, and Section 6 compares the results for German and French.

2. Text type – genre – audience

There is an active debate in linguistics concerning the notions of text types, genre or register (Halliday and Hasan, 1985; Swales, 1990; Biber and Conrad, 2009). Genre corresponds to conventional text patterns, recognized by a discourse community (Swales, 1990). Texts represent a channel for sharing and building knowledge across the community. In addition, community members use these text patterns when producing documents. This is a mean of proving their integration in the community, by adopting similar communication practices. For example, people from computer science community write and structure scientific articles in a similar manner (following the main structure : motivation, approach, methodology, evaluation and discussion of the results). Scientific articles from linguistic area will not have an evaluation section, but a section presenting data analysis. Argumentation sequences present some common domain-independent features : cue markers expressing arguments, hedges to express author's point of view with respect to the results presented in the article, modal verbs expressing possibility of a hypothesis that should be validated by the scientific community.

Text patterns are actually realized by structural elements (introductory and final formulas, recognized by the community members), but also by specific linguistic parameters, characterizing genre or text types (Biber and Conrad, 2009). Indeed, linguistic features are chosen for specific communication goal and for communication situation. For example, an author of a textbook adressed to Bachelor students care-

fully explains the terms used in the book: he gives complete definitions. He/she avoids ambiguous words and illustrates these notions with explanatory sequences. These procedures are applied because the targeted audience is not able to interact directly with the author. A teacher preparing an electronic document for a lecture knows that the students might ask clarification questions about the terms, so he provides short definitions, he prefer simple syntactic structures. During the oral presentation, he uses deictic pronouns to explain the adressed notions. In addition, the teacher invites his students to ask clarification questions. An author of a popular science article illustrates the main concepts with simple examples, taken from the real life. He proposes simplified syntax and definitions, to be understood by the target audience. The hypothesis and research results are presented as sure. In all the situations presented here, the text author have a marked preference to some linguistic phenomena, adapted to audience type. If some of these features are unique and characterized by their position in the document (at the end or at the beginning of it), we are mainly interested in frequent features.

Biber and Conrad (2009) propose a complete framework to genre or text type analysis. Their study proposes a set of 64 morpho-syntactic parameters interesting for characterizing genres or text types. These features include simple POS tags (the frequency of content words), but also complex features as relative clauses, subject type or complexity of noun and of prepositional phrases. Thus, textbooks are characterized by complex subjects or objects and explanatory sequences, while lecturer's documents used to illustrate the oral presentations are characterized mainly by personal pronouns, short definitions, rhetorical questions adressed to the audience, the preference for imperatives or question marks. Frequent simple noun phrases and simple objects are prefered by popular science articles. The presentation of the main findings are presented as sure, so modality verbs are quite rare in popular science.

We adopt Biber and Conrad's (2009) point of view: text types/genres/registers are identified by linguistic features selected by speakers to fit their communicative purpose and the profile of their target audience (its level of knowledge and language proficiency, among others).

To identify the most relevant linguistic features for our scientific vs. popular science corpora from several domains, we have identified linguistic markers for both categories, starting out from basic assumptions and extending them by conspicuous corpus patterns. For scientific texts, we start from Swales (1990) and Tutin (2010) by identifying specific patterns (argumentation patterns, author's point of view, explanation sequences). For popular science with its frequent educational purpose, we start from Hyland's analysis (Hyland, 2009), assuming a high prominence of rhetorical questions and of definitional patterns.

3. Corpus Design Principles

3.1. Selecting Scientific and Popular Science texts

The corpora created for work on the differences between scientific writing and popular science cover both French and German, within two domains, medicine and computer science. These two rather distant domains were selected on purpose, to avoid overlaps (as would occur, e.g. between medicine and biology), and to test the generalizability of the (morpho-)syntactic cues for the text type.

	Type	Publication/Source	Words	Format
FR	SCI	Revue de Rhumathologie	104k	PDF
		Médecine/sciences	209k	HTML
		Scientext	201k	Text
	POP	Patient sites	238k	HTML
		www.futura-science.fr	158k	HTML
		Le guide santé	106k	HTML
DE	SCI	Ärzteblatt	407k	HTML
		Arzneimittelbrief	829k	HTML
	POP	Diabetes-Ratgeber	336k	HTML
		Senioren-Ratgeber	180k	HTML
		TV-Gesund	18k	HTML

Table 1: Sources of medical texts for French and German, scientific (SCI) vs. popular science (POP), text type, size in words, and original format

We gathered corpora for these two domains by browsing the web and manually identifying relevant, publicly accessible publications. The results for the medical domain are shown in Table 1. The German publications *Deutsches Ärzteblatt* and *Arzneimittelbrief* target general practicioners and aim at keeping them up to date on new medical developments. *Senioren-Ratgeber*, *TV-Gesund* and *Diabetes-Ratgeber* are magazines for lay persons interested in general aspects of health and medicine or in particular in diabetes. The French medical on-line journals (*Revue française de rhumatologie* and *Médecine/sciences*) are written for specialists and for students. In addition, we selected some Ph.D. theses and scientific articles available from Scientext. For popular science, we used some web sites maintained by health insurances for their members, Web sites presenting some rare diseases (www.orphanet.fr), Web sites built by patient associations (www.forum-santé.fr).

For the computer science domain, it is not easy to find scientific articles in German or French, as most scientific publications (e.g., conference proceedings) are in English. We finally decided to use doctoral dissertations, which exist in substantial numbers in non-English languages. We gathered French theses from the CNRS Open archive (http://hal.cnrs.fr) and German theses from the university library document repositories of several universities. Also, there is a considerably broader spectrum within the popular science category of computer science that there is for medicine. It comprises both sources aimed at beginners and complete laypersons (http://www.01net.com and *L'internaute-High Tech* for French, *ComputerBild* for German) and those aimed at amateurs and using a more technical style (*PCWorld* for French, *c't* for German).

4. Extraction and annotation of the corpora

As Tables 1 and 2 show, almost none of the texts are in plain text format; the two predominant formats are PDF and HTML. These formats represent complementary challenges. This section describes how the corpora were collected, annotated with metadata, and analysed syntactically.

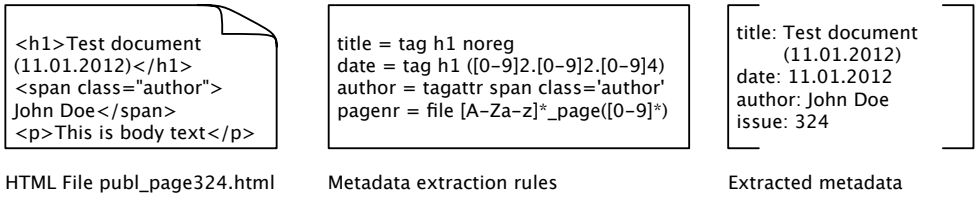

Figure 1: Example of metadata extraction from HTML file

	Type	Publication/Source	Words	Format
FR	SCI	articles, PhD theses	565 k	PDF
	POP	*L'Internaute, PC World*	386 k	HTML
DE	SCI	PhD theses	565 k	PDF
	POP	*ComputerBild*	1500 k	PDF
	POP	*c't*	2900 k	HTML

Table 2: Sources of computer science texts for French and German, scientific (SCI) vs. popular science (POP), text type, size in words, and original format

4.1. Extraction of text from HTML pages

The extraction of text from web pages is a well-researched problem in the context of creating web corpora (Baroni and Kilgarriff, 2006). The first step is always the collection of pages from the web (crawling). In a second step, each page is typically subject to three processing mechanisms: boilerplate stripping (the removal of generic material on pages, like link lists or scripts), function word filtering (the removal of pages that consist only of keyword lists), and porn filtering.

We have developed a flexible tool for the download of HTML archives of online publications. The first step is similar to the creation of web corpora: we employ the free web crawler WebHTTrack[1]. WebHTTrack can download individual files, but if the user specifies the file structure in which the articles are stored on the server, all articles can be downloaded.

Regarding the second step, we make different assumptions, though. Since the user of our tool manually selects which pages to download, we can do without function word or porn filtering. As for boilerplate removal, i.e., the identification of the actual page content, Baroni and Kilgarriff (2006) measure the tag density, assuming that the page content has a low tag density compared to link lists or other types of boilerplate text. However, we found that depending on the house style of publications, page contents frequently contain tags to indicate text structure (`<p>`) or text formatting (`<b>`). The alternative strategy that we pursue is to let the user specify one tag that contains the page content. This is typically (`<p>`), and in all cases that we considered, this tag was constant across all pages from one publication. Additionally, we discard lines that occur on 50% or more of all pages from one server (e.g., "Click here for more information").

We also add a third step, namely the extraction of document metadata which is highly relevant for building a rich corpus. Fortunately, many HTML pages explicitly mark article information like author, date of publication, article title, and file URL. The way in which these meta data are represented is again consistent within one publication but differs between publications. For this reason, our tool allows the user to specify which parts of the HTML page or its URL contain the individual metadata information, using regular expressions. This is illustrated on an example in Figure 1. The middle column shows the extraction "rules". The first and second rule extract the title and the date, respectively. They are examples of "tag" rules. They return the content of particular HTML tags, in this case `h1`. The *date* rule additionally specifies through a regular expression which part of the tag content should be returned. The match is marked, as usual, with round brackets. The title rule does not give a regular expression ("noreg"). In this case, the complete tag is returned. The *author* rule is a "tagattr" rule which allows the user to specify not only a tag but also an attribute of the tag. Finally, the *pagenr* rule is not applied to the HTML file but to its name ("file") and again defines a regular expression for the filename string.

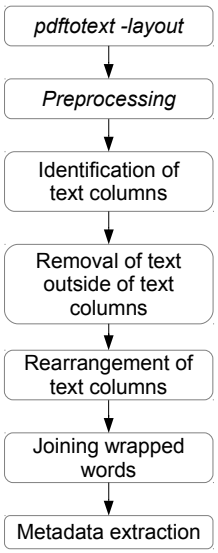

Figure 2: PDF postprocessing pipeline

4.2. Extraction from PDF files

We developed a second tool to extract text from PDF files. A number of such extractors exist already (e.g. the Unix tool *pdftotext* which is part of the xpdf suite[2]), but most tools only work satisfactorily if the PDF document contains only one text column. In documents with several columns, text from these columns is often extracted in an incorrect order. We therefore run *pdftotext* in layout-preserving mode (-

[1] `http://www.httrack.com`

[2] `http://www.foolabs.com/xpdf`

Genre	GF	# Gold	Recall	Precision	F_1
SCI	SBJ	97	76%	74%	74%
	OBJ	22	73%	64%	67%
	PRD	9	89%	57%	70%
	ALL	128	76%	69%	72%
POP	SBJ	90	69%	74%	71%
	OBJ	57	75%	74%	74%
	PRD	4	100%	29%	44%
	ALL	151	72%	70%	71%

Table 3: Domain-specific German parser evaluation on medicine data: argument-specific (SBJ, OBJ, PRD) and totals

Genre	GF	Base	Recall	Precision	F_1
SCI	SBJ	84	95 %	93 %	94 %
	OBJ	166	61 %	84 %	71 %
	PRD	4	100 %	66 %	80 %
	ALL	254	73 %	87 %	79 %
POP	SBJ	105	94 %	97 %	95 %
	OBJ	138	55 %	71 %	62 %
	PRD	10	62 %	88 %	72 %
	ALL	253	69 %	62 %	65 %

Table 4: Domain-specific evaluation of French parser on medicine data: argument-specific (SBJ, OBJ, PRD) and totals

layout) and postprocess its output with completely self-developed heuristics. Figure 2 shows that this process involves several successive steps, at the end of which we recover single-column text.

The first step ("preprocessing") removes undesirable characters like e.g. ligatures or control characters. Next, the typographical structure of the page is analyzed and a set of text columns is identified by identifying the positions of uninterrupted text sequences on the page to determine column boundaries. After we have determined the boundaries, we delete all elements that do not correspond to any of the text columns. This step gets rid of tables (typically recognizable by a large number of isolated numbers), captions (which typically span more than one column), and material in headers/footers (which are also outside the columns). Then the columns are rearranged. Yet many words are still wrapped around line breaks. We join as many as possible without removing genuine hyphens. To do so, we create a word list of the entire document and join wrapped words only if their combination without hyphen occurs in the word list. An evaluation on a small set of documents showed that this simple heuristic improves the quality of the subsequent POS tagging step by an average of 7%.

Finally, the metadata is extracted from each page of the input document. This is, however, considerably more difficult than for HTML files, since the output of the extraction is unstructured text in which metadata information is very hard to detect. Under the assumption that it is better not to extract metadata when it is unreliable, we found that we could only retrieve the issue number and the page number automatically with an acceptable precision.

4.3. Linguistic annotation and quality

We parsed the corpora for both languages with the MATE dependency parser (Bohnet, 2010), using parsing models based on the German TIGER treebank(Brants et al., 2002) and the French treebank (Abeillé et al., 2003), respectively. The parser provides output in the CONLL format, a widely used simple column-based format[3]. We also conducted small manual evaluations of the parser output for both languages with a focus on subject and object relations, to check if the quality of potential (morpho)-syntactic features for text classification might be degraded by parsing problems.

[3] http://ufal.mff.cuni.cz/conll2009-st/task-description.html

German. We gained an impression of the parse quality on German medical text by drawing two random samples of 100 sentences from the SCI and POP genres each. We manually annotated the (surface) subjects (SBJ), direct objects (OBJ), and predicatives (PRD) of main clause predicates, ignoring embedded clauses completely. These three categories are arguably most relevant for genre-based feature extraction (see Section 5. for details).

The results of comparing manual and parser categories (exact match) are shown in Table 3. Not surprisingly, the performance of the parser was below the reported numbers on offical benchmarks: not only were the texts that we analysed fairly dissimilar to the newspaper text that the parser was trained on, but they also contained various artifacts (see below). Contrary to our expectations, however, both text types were approximately equally difficult to parse. The errors were substantially different, though: In scientific text, the two main error types were attachment errors, and subjects misclassified as objects due to morphological ambiguity (note the large number of passive sentences which can be inferred from low number of objects). Both of these error types require advances in parsing technology for future improvements. Only in third place do we find a preprocessing-related error class, namely parsing errors caused by the inclusion of headings in text body sentences. In contrast, for the popular science corpora such preprocessing-related problems account for a clear majority of all errors. Examples are colon-separated semi-sentences and enumerations where the ordinal is erroneously included in the sentence text. Such errors can can presumably be alleviated by better sentence splitting and handling of document structure.

French. For French, we found better results for subject identification. We take randomly 100 phrases from each genre. To obtain a reference corpus, we manually analyse the parser output and we correct the border identification and the dependency errors. We annotate missing subjects or objects. Parsing results were comparable with German data, as displayed in Table 4.3., with slightly better results for scientific texts. We note that the number of border identification errors is greater than the number of wrong dependency labels. Subject and object borders show some errors due to attachment problems (appositions, conjunctions). Objects are hit harder for French corpora, in particular for popular science. Subject and object labeling errors are less frequent than in German, due to the more fixed constituent order. At

the same time, many errors arise from auxiliary-participle constructions, where the parser wrongly attaches the subject to the auxiliary and the object to the main verb. Some complex verbal phrases are not uniformly processed : the object marked by the same preposition modifying the same verb is labelled sometimes as direct object and sometimes as prepositional object (in constructions as *permet d'accéder* 'allows to access' and *permet de connecter* 'allows to connect'). Other errors identified in the scientific texts were generated by long enumerative sequences and by explanations (between round brackets). Preprocessing (titles included in the main document, POS errors) also form a large source of errors. Finally, when some adverbs are used between the verb and the adjective, the adjective is frequently mislabeled as modifier.

Availability. The parts of the corpora which have been freely available from the web can be downloaded.[4].

5. Analysis of text type differences

This section presents our analysis of medical and of computer science corpora. We study several properties proposed in the literature as being representative for scientific and popular science. Several properties characterize scientific texts, such as the author's point of view (Hyland, 2009), argumentation patterns, and a preference for relational adjectives (Daille, 1999). Popular science texts are characterized by rhetorical questions (Hyland, 2009) and by definition and reformulation patterns. For French, term density (Kocourek, 1991) is considered to be a specific feature for scientific texts, while explanations (Jacobi et al., 1988) are typical for popular science. In addition, we study simple features such as word and sentence length, content word frequency (verbs, nouns, adverbs, adjectives), but also features proposed in (Biber and Conrad, 2009), namely passive constructions and the syntactic structure of subjects or objects.

We study the two corpora to identify a range of linguistic expressions realizing these textual properties. We use a concordancer and both raw texts and parsed texts (including POS tags, lemma information and dependency relations). For each feature, we compute its relative frequency as its absolute frequency in the corpus divided by the size of the corpus. In this way, the size of the corpus does not come into play.

First, we study *simple statistical* features as the average word length or sentence length. Also, we compare the frequence of content words in the two corpora: nouns, adjectives, verbs, adverbs. This comparison points out that some of these features (adverb's frequencies, long words or phrases) where specific to scientific texts. Relational adjectives might be identified with POS filtering and some specific endings (in French : *-al, -ique, -ien*, in German *-iv, -al*).

Second, we identify *linguistic* features by means of morphosyntactic patterns. For example, the author's point of view can be indicated in the text by 1st person pronouns, by argumentative sequences, or by modal expressions. Argumentative sequences are characterized by the use of cognitive verbs (*penser 'to think', réfléchir 'to consider'*), the use of communication verbs (*présenter, exposer 'to present'*) or by the use of impersonal expressions (*il est possible/probable 'it is possible/probable'*). As shown in Table 6, 1st and 2nd person pronouns (singular or plural) are very frequent in popular science texts with respect to science texts.

Definition patterns are identified by some lexico-syntactic patterns (in French *X est défini comme Y 'X is defined as Y'* , *X est nommé Y 'X is called Y'*). We use these patterns and the concordancer to compare the frequency of these elements in popular vs scientific texts. In French, this property seems to be specific to popular science rather than scientific texts. In German, several definition patterns are used: *X ist eine Art Y 'X is a kind of Y", X bedeutet Y 'X means Y', heisst 'is called', nennt man 'one calls'*. All of them are preferred for popular science.

Explanations, which are typical for popular science texts are enclosed into brackets (),[] or by specific markers (*autrement dit, c'est-à-dire 'in other words'*). In French, reformulation markers are very frequent for popular science in both domains, as well as passive constructions (see Table 5). However, caution is necessary in the computer science domain, since brackets might introduce a code fragment and these elements are found both in popular and scientific texts.

Impersonal pronouns are substantially more frequent in scientific texts. For French, these occurences must be identified in context (followed by a verb like 'be' and a specific impersonal adjective *il est nécessaire* 'It is necessary').

The complexity of subjects and objects is an important issue, as presented in Table 5. Complex terms (noun phrases modified by several PPs or by relative clauses) are frequent in scientific texts rather than popular texts and occur in subject or object positions.

Domain term identification fails to give interesting results. Indeed, we compared the first 50 domain terms extracted from the scientific corpora and from the popular science corpora, using the TTC multilingual term extractor (Weller et al., 2011). This tool extracts terms from monolingual corpus by applying some syntactic patterns and by comparing the domain-specific corpus with a general language corpus. In our case, domain-specific terms were found both in scientific and popular texts, so they have limited influence as text feature.

6. Selected Features and Contrastive Study

Following the observations reported in the previous section, we categorize the features for genre classification into the following groups:

- statistical features (word length, sentence length – can be computed language-independently from text without linguistic analysis);
- lexical features (nouns, adverbs, adjectives, verbs – domain-specific);
- morphosyntactic features (simple and complex noun phrases, simple subject, simple object, complex subject, complex object – computed from dependency analyses of sentences).

[4] http://www.ims.uni-stuttgart.de/~kisselmx/classyn.html

	Definition	Reformulating markers	1st person pronoun	passive	complex subject
MED SCI	814	214	2316	3069	41536
MED POP	2212	132	274	2546	10451
COM SCI	1704	113	2049	657	15789
COM POP	11070	530	20	167	4324

Table 5: Comparison of some features on French corpora : MED - medicine; COM- computer science; SCI - scientific tetxs; POP - popular science. Numbers are relative frequencies in parts per milion.

Source	Type	?	!	(,)	[,]	1. Pers Sg	2. Pers Sg	1. Pers Pl	2. Pers Pl
Arzneimittelbrief	SCI	3830	180	23770	60	190	0	970	1050
Ärzteblatt	SCI	1160	350	12530	240	1460	30	1740	2450
Diabetes-Ratgeber	POP	3200	1410	6720	0	1820	190	1100	2920
Senioren-Ratgeber	POP	6870	960	2570	10	7310	240	2960	4170

Table 6: Comparison of some features on German medicine corpora : SCI - scientific tetxs; POP - popular science. Numbers are relative frequencies in parts per milion.

In our analysis, we found that scientific writing and popular science can be distinguished on the basis of morphosyntactic properties, as we expected, but also just by statistical features, in particular average sentence and word length: scientific writing contains longer sentences and more specialized terminology (i.e. longer words). In German scientific writing, hyphenated compounds abound. The same holds for measuring units, indications of percentages, indications in parentheses or brackets.

In terms of morphosyntactic properties, French scientific writing is characterized by the author's politeness 1st person plural form "nous", which does not occur in popular science articles. Interestingly, this feature is not discriminative in German, where the popular science corpus contains many reports by patients, where the 1st person plural is used. The two languages coincide, however, in that they use more 1st and 2nd person singular forms in popular science, which is again a result of personal reports not found in scientific text. The syntactic complexity of subject and object noun phrases is also a distinctive criterion: popular science is characterized by a less compact style and thus has more simple NPs (`Det (Adj) N` in German, `Det N (Adj)` in French) and a smaller number of complex NPs (with pre- or post-modification) than scientific writing. Other indicators of scientific writing include impersonal constructions (*il est important de...* 'it is important'), definitional contexts and the use of less adverbs than in popular science.

On the basis of this analysis, we have developed a module to extract features from parsed corpora. The module counts the relative frequency of each feature in the document. The module applies the patterns presented below to extract all the features and can be used subsequently for genre classification.

7. Conclusion and Further Work

We present the ressources (corpora and tools) built for a genre-based classification systems. We build comparable corpora, from two various domains (computer science and medicine), composed of scientific and popular science texts. We present the tools developed to gather corpora from the Web. To classify documents, we parse the documents with a statistical dependency parser available for French and German. We analyse the differences between popular science and scientific text types and we compare the specific properties across languages. The results of this analysis will be used to select relevant features for classification purposes.

Acknowledgments. This work was partly supported by the DAAD PROCOPE project and the French Ministy of Foreign Affairs. We would like to thank Bernd Bohnet for his support with the parser, and Hailian Jiang for her work on parser evaluation for German.

8. References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. *Treebanks*, pages 165–187.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Douglas Biber and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Sabine Brants, Sabine Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Thierry Charnois, Anne Doucet, and Yann Mathet. 2008. Trois approches du GREYC pour la classification de textes. In *Proceedings of TALN'08*, Avignon, France.

Beatrice Daille. 1999. Identification des adjectifs relationnels en corpus. In *Proceedings of the TALN Conference*, Cargèse, France.

Clément de Groc. 2011. Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France.

Mark A. K. Halliday and Ruqaiya Hasan. 1985. *Language context and text: Aspects of language in a socialsemiotic perspective*. Oxford University Press.

Ken Hyland. 2009. *Academic Discourse*. London: Continuum.

Daniel Jacobi, Bernard Schiele, and Jean Marie Albertini. 1988. *Vulgariser la science*. Seyssel, Éditions Champ Vallon (Milieux).

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of COLING 94*, pages 1071–1075.

Rostislav Kocourek. 1991. *La langue française de la technique et de la science*. Oscar Brandstetter Verlag, Wiesbaden.

Céline Poudat, Guillaume Cleuziou, and Viviane Clavier. 2006. Catégorisation de textes en domaines et genres: complémentarité des indexations lexicale et morphosyntaxique. *Document numérique*, 9:61–76.

Marina Santini. 2007. Characterizing genres of web pages: Genre hybridism and individualization. In *Hawaii International Conference on Systems Science*, page 71, Waikoloa, HI, USA.

Fabrizio Sebastiani. 2005. Automatic classification of text. In Keith Brown, editor, *The Encyclopedia of Language and Linguistics*, volume 14. Elsevier Science Publishers, Amsterdam, NL, second edition.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26:461–485.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Agnès Tutin. 2010. Evaluative adjectives in academic writing in the humanities and social sciences. In R. Lores-Saz, P. Mur-Duenas, and E. Lafuente-Millan, editors, *Constructing Interpersonality: Multiple perspectives on written academic genres*. Cambridge Scholars Publishing.

Marion Weller, Helena Blancafort, Anita Gojun, and Ulrich Heid. 2011. Terminology extraction and term variation patterns: a study of French and German data. In *Proceedings of the GSCL: German Society for Computational Linguistics and Language Technology*, Hamburg, Germany.