Exploring a corpus of scientific texts using data mining

Elke Teich

Technische Universität Darmstadt, Germany

Peter Fankhauser

L3S, Leibniz Universität Hannover, Germany

Abstract

We report on a project investigating the linguistic properties of English scientific texts on the basis of a corpus of journal articles from nine academic disciplines. The goal of the project is to gain insights on registers emerging at the boundaries of computer science and some other discipline (e.g., bioinformatics, computational linguistics, computational engineering). The questions we focus on in this paper are (a) how characteristic is the corpus of the meta-register it represents, and (b) how different/similar are the subcorpora in terms of the more specific registers they instantiate. We analyze the corpus using several data mining techniques, including feature ranking, clustering and classification, to see how the subcorpora group in terms of selected linguistic features. The results show that our corpus is well distinguished in terms of the meta-register of scientific writing; also, we find interesting distinctive features for the subcorpora as indicators of register diversification. Apart from presenting the results of our analyses, we will also reflect upon and assess the use of data mining for the tasks of corpus exploration and analysis.

1. Introduction

The broader context in which the present paper is placed is corpus comparison. Corpus comparison is involved in many areas of corpus linguistics, ranging from the comparative analysis of registers/genres, varieties and languages (including translations), both from a synchronic and from a diachronic perspective (Biber, 1988, 1995; Mair, 2006, 2009; Teich, 2003). With this comes a concern for methodologies of corpus comparison, laying out the principal work flows in corpus compilation and corpus processing (annotation, query). In this context, questions of data analysis have recently received some attention, addressing the important issue of appropriate statistical measures for interpreting quantitative data. Examples are Kilgariff (2001) who discusses a range of statistical techniques and their applicability to lexically-based corpus comparison, or Gries (2006) who presents a method of measuring variability within and between corpora.

The primary concern of the present paper is a methodological one. The concrete background of our work is a research project on the specifics of language use in

interdisciplinary scientific contexts, with a focus on scientific registers at the boundaries of computer science (such as bioinformatics, computational linguistics or computational engineering). The research questions we are interested in are of the following kind: What are the linguistic effects of a scientific discipline coming into contact or merging with computer science? To what extent are the linguistic conventions of the original discipline retained? Are there any tendencies to adopt the language of computer science? Or are there new registers developing?

The data we work on is the *Darmstadt Scientific Text Corpus* (DaSciTex), which contains full English scientific journal articles compiled from 23 sources covering nine scientific disciplines. The corpus comes in two versions, a large one comprising around 19 million words, and a small one comprising around one million words (Holtz and Teich, in preparation).[i] The corpus includes texts from the broader areas of humanities, science and engineering and has a three-way partition (see also Figure 1 for a diagrammatic representation):

A computer science
B 'mixed' disciplines:
  B1: computational linguistics
  B2: bioinformatics
  B3: computer aided design/construction in mechanical engineering
  B4: microelectronics/VLSI
C 'pure' disciplines
  C1: linguistics
  C2: biology
  C3: mechanical engineering
  C4: electrical engineering

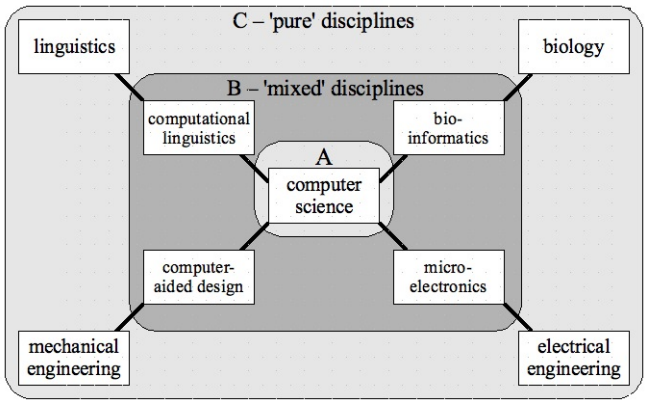

Figure 1: Design of DaSciTex

At a methodological level, the project is concerned with developing a methodology for corpus comparison with a special view to fine-grained linguistic variation: What are the most distinctive features between the corpora under investigation and how can we obtain these features? To approach this question, we explore a set of techniques from an area known as data mining (see Chakrabarti (2003) and Witten and Eibe (2005) for comprehensive introductions as well as Manning and Schütze (1999:Chapter 16) and Sebastiani (2002) for overviews). To our knowledge, except for clustering (e.g., Biber (1993)), data mining methods have hitherto hardly been explored in the context of corpus comparison.

We have carried out several analyses using data mining which address the following two types of questions:

(1) How well is the corpus distinguished in terms of features characterizing the meta-register of scientific writing?
(2) How different/similar are the subcorpora in terms of features characterizing the individual registers?

To explore the first question, we have compared the DaSciTex corpus with the registerially-mixed FLOB corpus.[ii] If DaSciTex represents scientific writing, then the texts contained should exhibit some typical properties of this meta-register, such as (relative) abstractness, technicality and informational density (cf. Halliday (1985); Halliday & Martin (1993)), which are not exhibited by a registerially-mixed corpus such as FLOB. To explore the second question, we have compared the subcorpora of DaSciTex in order to determine the relative position of the mixed disciplines vis à vis their corresponding pure disciplines and computer science. Inspecting the texts in the corpus, we observed that one potential source of difference lies in the roles and attitudes participants adopt in the discourse situation (cf. Section 2 on parameters of register variation). One indicator of this is the self-construal of the authors in terms of the types of activities engaged in. For some examples see (1)-(3) below, (1) instantiating two material processes with *we* as Actor, (2) instantiating a mental process with *we* as Senser and (3) a verbal process with *we* as Sayer.

(1) *We analyze* and *compare* different queue policies…
(2) *We believe* that competitive analysis gives important insights…
(3) *We argue* that novel instances of verb adjective sequences are based on analogies to previous experiences ...

Analyses of this kind, testing for a variety of potential register features, will bring out the differences and commonalities between the registers under investigation.

The remainder of the paper is organized as follows. We start with an introduction to the underlying linguistic-theoretical framework we work with, Systemic

Functional Linguistics (SFL; Halliday, 2004), which has at its core a model of register variation (Section 2). This is followed by the presentation of the analyses we have carried out using selected techniques of data mining (Section 3). Section 4 concludes the paper with a summary and discussion.

2. Theory and model of register variation: Systemic Functional Linguistics

In descriptive linguistics, the notion of register refers to linguistic variation according to *use in context* (cf. Quirk et al. (1985); Biber et al. (1999)). Register variation is a well researched topic that typically requires working in a corpus-based fashion. In order to account for register differences in a corpus of texts, one needs a sound model of register variation that allows addressing the research questions involved. One such model is Systemic Functional Linguistics (SFL; Halliday, 2004). The notion of register is at the core of the language model put forward by SFL (Halliday et al., 1964; Halliday, 1985; Halliday and Hasan, 1989; Matthiessen, 1993). SFL considers language a multi-dimensional resource for making meaning. The two dimensions of interest here are *stratification* and *instantiation* (see Figure 2 below). According to stratification, the linguistic system is organized along the levels of lexico-grammar, semantics and context, where lexico-grammar is taken to realize semantics and semantics is taken to realize context. Instantiation refers to the relation between the linguistic system and a text (i.e., an instance). Each instance is characterized by the selection of particular linguistic (semantic, lexico-grammatical) features according to a context of situation. This situated language use results in *registers* or *text types*. For example, different sets of linguistic features will be chosen by speakers involved in a casual conversation compared to a highly-structured and planned discourse, such as a written academic paper.

In SFL-based analysis of texts, an account of the contextual configuration forms an integral part. The instrument provided for accounting for a contextual configuration is given by three parameters that are said to characterize the level of context. These are *field*, *tenor* and *mode* (cf. Quirk et al. (1985) who suggest a similar classification into field, attitude and medium of discourse). Field of discourse is concerned with subject matter and the goal orientation of a text (e.g., expository, narrative, instructional). At the level of lexico-grammar, field is reflected in configurations of processes and the participants therein, such as Actor, Goal, Medium, and accompanying Circumstantials of Time, Place, Manner etc. (see again examples (1)-(3) in Section 1 above). Tenor of discourse is concerned with the roles and attitudes of the participants in a discourse. Linguistic reflexes can be found in choices of mood and modality as well as appraisal. Mode of discourse is concerned with the role language itself plays in the discourse, e.g., whether it is substantive or ancillary, whether the channel of communication is visual and/or auditive, whether the medium is written or

spoken. Linguistic reflexes of this parameter are primarily textual (e.g., thematic structure, information structuring, informational density vs. grammatical intricacy). A register is thus constituted by particular settings of these three parameters – the contextual configuration – together with the sets of linguistic features *typically* chosen according to that contextual configuration (cf. Halliday and Hasan (1989)).

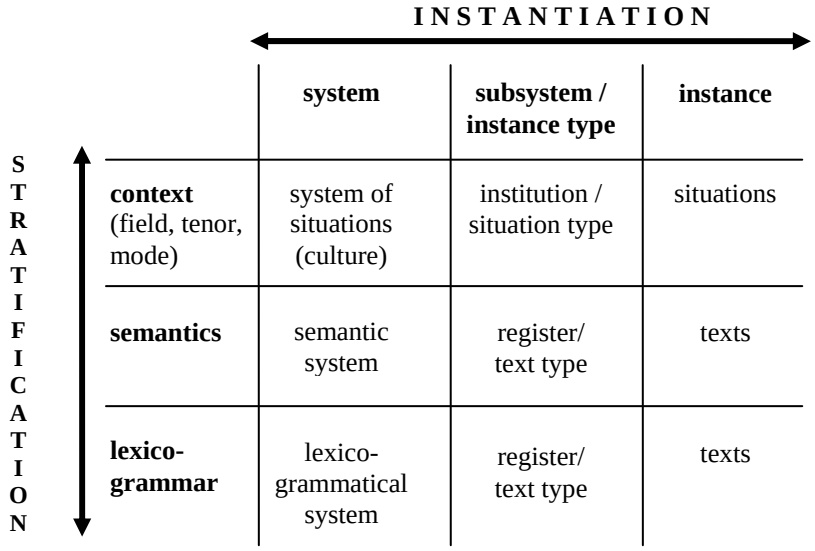

Figure 2: Register, stratification and instantiation

An analysis of the register(s) of a text or set of texts crucially involves statements about the distribution of features, i.e., it is a quantitative account (cf. Halliday (2005)). Also, since a text may exhibit some of the features typical of a register but not others (or, in other words, for a text to belong to a given register is a matter of degree), it is desirable to be able to bring out this kind of fuzziness in the analysis. Finally, since there is typically more than one feature involved in register differentiation, multivariate techniques qualify better than univariate ones (cf. again Biber (1993)). A set of methods that offer most of the desirable functionality is provided by an area known as data mining. The following section describes the analyses we have carried out on our corpus addressing the questions posed in Section 1, employing data mining techniques such as feature ranking, clustering and classification.

3. Data mining for corpus comparison: experimental setup and results

To approach the questions formulated in Section 1, we have carried out a number of analyses on the basis of selected, potentially discriminatory features using the WEKA data mining platform (Witten and Eibe, 2005). In this section we describe the two set-ups for these analyses and their results. The first set-up addresses the first question by comparing DaSciTex with FLOB (Section 3.1); the second addresses the second question, comparing the subcorpora in DaSciTex (Section 3.2).

3.1 Analysis (1): Comparing DaSciTex with FLOB

The data sets we work with in these analyses are the small version of DaSciTex (around one million tokens, 186 texts) and FLOB (also containing around one million tokens, 405 texts). Both corpora have been tokenized and part-of-speech tagged with the Tree Tagger (Schmid, 1994).[iii] This first set of analyses aims at comparing the DaSciTex corpus as an instance of scientific writing with the FLOB corpus which contains instances of various different registers (including a partition of scientific writing).

We have investigated the following candidate features as potential indicators of scientific writing, focusing on the properties of abstractness, technicality and informational density:

- the relative number of nouns (NN), lexical verbs (VV), and adverbs (ADV) as possible indicators of abstract language,
- the standardized type-token ratio (STTR) as a potential indicator of technical language,
- the average number of lexical words per clause (LEX/C), i.e., lexical density, as a measure for the informational density of a text.

Table 1 gives the overall averages for FLOB and DaSciTex for these features evaluated for their discriminatory force by the technique of Information Gain (IGain) and, for comparison, by the T-Test. Both Information Gain and T-Test measure how well a feature distinguishes between classes.[iv] For a set of features these measures also provide a ranking: the features in the table appear ranked in the order of top (highest discriminatory force: STTR) to bottom (lowest discriminatory force: VV) according to Information Gain. Note that this ranking matches the ranking by the T-Test exactly.

Table 1: Results for selected features comparing DaSciTex and FLOB

	FLOB	DaSciTex	IGain	T-Test
STTR	43.6	34.0	0.48	23.8
ADV	0.056	0.034	0.33	21.2
NN	0.27	0.33	0.33	-18.3
LEX/C	6.16	8.39	0.26	-15.6
VV	0.114	0.097	0.12	10.3

The best discriminator is the standardized type token ratio (STTR), with texts in the DaSciTex corpus having a significantly lower type token ratio than the texts in FLOB.[v] The T-Statistics is 23.8, which is well above the critical value of 1.9 for 0.95 confidence. Accordingly, also the Information Gain of STTR is fairly high with 0.48. The next two features are the relative numbers of adverbs and nouns. DaSciTex has a larger number of nouns and a smaller number of adverbs than FLOB, a difference that is again significant, as shown by both the T-Statistics and Information Gain. The average number of lexical words per clause is ranked as the fourth feature. Again, DaSciTex has a larger number of lexical words per clause than FLOB. Interestingly, the relative number of lexical verbs, shown here as the last feature, is a less strong discriminator than the number of adverbs , but still the number of verbs is significantly smaller in DaSciTex than in FLOB.

We have also investigated a number of other candidate features, including the number of words per sentence (T-Statistics: 3.4) and the number of clauses per sentence (T-Statistics: -7.6) as possible alternatives for LEX/C, and the ratio of lexical words vs. function words. While all these features have a T-Statistics above the critical value, they are far less discriminatory than the five features given above.

Information Gain and T-Statistics measure how well an individual feature distinguishes between classes. In order to better understand how these features *together* distinguish DaSciTex and FLOB, we have used them to train a classifier based on a linear support vector machine, and clustered them based on K-Means. See Table 2 below for the results. The top four features achieve a classification accuracy of 90% (with ten-fold cross validation), and a clustering accuracy of 84%.[vi] While the clustering accuracy is naturally lower than the classification accuracy, it shows that the features separate the two corpora into two clusters, with one cluster mainly containing texts from DaSciTex and the other one mainly from FLOB. Clearly, one cannot expect a 100% accuracy, as FLOB is composed of several registers including official reports (H - Government) and scientific writing (J - Learned), which are expected to have similar characteristics as the texts in DaSciTex with respect to the investigated features. Indeed, most of the texts from H and J are grouped into the DaSciTex cluster. When removing these from FLOB, leaving 297 texts, the classification accuracy goes up to 97%, and

the clustering accuracy goes up to 92%, as shown in Table 2.[vii] Here, the most discriminatory feature alone – STTR – achieves a classification accuracy of 91%.

Table 2: Results for classification and clustering comparing (a) DaSciTex and FLOB, and (b) DaSciTex and FLOB minus H, J (FLOB')

	classification	**clustering**
DaSciTex vs. FLOB	90%	84%
DaSciTex vs. FLOB'	97%	92%

Not considering H and J will then also increase the results for Information Gain and T-Test, as shown in Table 3 below.

Table 3: Results for selected features comparing DaSciTex and FLOB'

	FLOB'	**DaSciTex**	**IGain**	**T-Test**
STTR	45.3	34.0	0.75	29.5
ADV	0.060	0.034	0.50	23.8
NN	0.27	0.33	0.41	-19.0
LEX/C	5.76	8.39	0.38	-18.4
VV	0.12	0.097	0.19	12.2

Thus we can conclude that type token ratio, number of nouns and adverbs, and number of lexical words per clause distinguish DaSciTex from FLOB quite well. However, within the subcorpora of DaSciTex, these features are not distinctive. For example, when contrasting computer science and the mixed disciplines (A+B1+B2+B3+B4) with all pure disciplines (C1+C2+C3+C4), the Information Gain of all features but STTR is 0, with STTR still a very low 0.09. This indicates that DaSciTex is not only well distinguished from FLOB, but also rather coherent in itself with respect to these features.

3.2 Analysis (2): Comparing subcorpora in DaSciTex

The data set we work with in these analyses is the full set of texts in DaSciTex (1843 texts with about 19 million tokens). The texts are tokenized and part-of-speech tagged with the Tree Tagger (Schmid, 1994). This second type of analysis aims at comparing the subcorpora in DaSciTex to see how well the registers are discriminated.

As pointed out above, shallow features such as a low type-token ratio clearly characterize the meta-register of scientific writing, but they cannot distinguish between individual disciplines. Of course one can expect that disciplines are well

distinguished by their subject specific vocabulary represented mainly by nouns and to a lesser extent by verbs. To analyze this, we have selected the 500 most distinctive nouns in terms of their Information Gain and transformed the texts to their term vectors representing the frequencies of these 500 nouns. This representation we have used to train and test a classifier that classifies texts into the nine disciplines.[viii] The achieved classification accuracy is 96%, and the misclassifications, which indicate overlaps between the subcorpora, exhibit an interesting pattern. See the confusion matrix in Table 4, where each row gives the predicted classes for an actual class. The main diagonal (in bold) gives the number of correctly classified texts. 28 misclassifications (1.5%) occur between C4 and other engineering disciplines (A, B3, C3) (shaded in dark grey). 27 misclassifications (1.5%) occur between a mixed discipline (B1 through B4) and its corresponding pure discipline (C1 through C4) (two secondary diagonals, shaded in light grey). 15 misclassifications (0.8%) occur between computer science (A) and one of the mixed disciplines (shaded in grey), and only 6 otherwise. Thus we can observe that the engineering disciplines have the largest overlap, and the mixed disciplines have a larger overlap with their corresponding pure disciplines than with computer science, but overall, the overlap is fairly small.

Table 4: Confusion matrix for classification by nouns

C/P	**A**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**	**C4**	**Sum**
A	**217**	**0**	**2**	**2**	**1**	0	0	0	**5**	227
B1	**3**	**76**	0	0	0	**10**	0	0	0	89
B2	**3**	1	**275**	0	0	0	**6**	0	0	285
B3	**3**	0	0	**215**	0	0	0	**2**	**4**	224
B4	**1**	0	0	1	**204**	0	0	0	**0**	206
C1	0	**4**	0	0	0	**95**	0	0	1	100
C2	0	0	**5**	0	0	1	**236**	0	0	242
C3	0	0	0	**0**	0	0	0	**246**	**6**	252
C4	**4**	1	0	**4**	**0**	1	0	**5**	**203**	218

A: computer science
B1: computational linguistics; **B2**: bioinformatics; **B3**: computer aided design; **B4**: microelectronics
C1: linguistics; **C2**: biology; **C3**: mechanical engineering; **C4**: electrical engineering

Classification with the top 250 lexical verbs still achieves a fairly high accuracy of 87% and confirms the general pattern of overlap (see Table 5). We get 97 misclassifications (5.3%) among the engineering disciplines (C4 vs. A, B3, C3), 66 misclassifications (3.6%) between a mixed discipline and its corresponding pure discipline, 28 misclassifications (1.5%) between computer science and mixed disciplines, and 48 misclassifications (2.6%) otherwise.

Table 5: Confusion matrix for classification by verbs

C/P	**A**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**	**C4**	**Sum**
A	**200**	**3**	**1**	**9**	**2**	0	1	1	**10**	227
B1	**5**	**65**	5	1	1	**11**	0	0	1	89
B2	**2**	7	**266**	2	1	1	**3**	1	2	285
B3	**5**	4	0	**180**	1	0	0	**17**	**17**	224
B4	**1**	2	0	2	**200**	0	0	0	**1**	206
C1	0	**9**	0	1	0	**90**	0	0	0	100
C2	0	0	**6**	1	0	2	**229**	4	0	242
C3	1	0	1	**13**	1	1	2	**222**	**11**	252
C4	**13**	0	1	**32**	**6**	0	0	**14**	**152**	218

A: computer science
B1: computational linguistics; **B2**: bioinformatics; **B3**: computer aided design; **B4**: microelectronics
C1: linguistics; **C2**: biology; **C3**: mechanical engineering; **C4**: electrical engineering

In order to analyze further how the usage of verbs differs across disciplines, we investigate the colligation patterns of verbs with the pronoun *we* in Subject position. As illustrated in Section 1 this is interesting from the point of view of tenor of discourse (self-construal of the authors). These patterns can be extracted fairly accurately from the part-of-speech tagged corpus by means of a regular expression that selects all lexical verbs following *we*, possibly interleaved with adverbs and auxiliary verbs. Moreover, we split the individual texts into chunks of 30 subsequent occurrences of *we* + verb to balance the different text lengths in DaSciTex. Again, we only take the top 250 verbs into account.

Table 6 gives the confusion matrix for the triple of A (computer science), B1 (computational linguistics) and C1 (linguistics), uniformly sampled, such that each register contributes the same number of instances (234). The achieved classification accuracy is 81% (87% for the full set, which contains more instances for A). As is to be expected, this is smaller than the accuracy achieved by classification with all verbs. Again the largest number of misclassifications (79 = 11.3%) occurs between B1 and C1, followed by 40 misclassifications (5.7%) between A and B1. Only 6 instances from A are misclassified as C1 or vice versa.

Table 6: Confusion matrix for classification by *we* + verb

C/P	**A**	**B1**	**C1**
A	**210**	**21**	3
B1	**19**	**168**	**47**
C1	3	**32**	**199**

A: computer science, **B1**: computational linguistics, **C1**: linguistics

Because we use a linear support vector machine for classification, which assigns a negative or positive weight for each feature, we can also determine the most typical verbs for each subcorpus.

Table 7: The nine most typical *we* + verb for each pair of subcorpora

A vs. B1	A vs. C1	B1 vs. A	B1 vs. C1	C1 vs. A	C1 vs. B1
show	define	train	describe	argued	turn
prove	use	adopt	collect	argue	speculate
present	show	describe	examine	turn	feel
choose	present	induce	simplified	don	coded
save	denote	examine	use	read	assume
obtain	save	constrain	separated	examine	met
touch	evaluate	combined	evaluated	feel	read
get	describe	downloaded	given	suggesting	find
proved	obtain	separated	define	Saw	presenting

A: computer science, **B1**: computational linguistics, **C1**: linguistics

Table 7 shows the nine most typical verbs for each pair of subcorpora. Verbs typical of A in contrast to B1 and C1 are shaded in light grey, verbs typical of B1 compared to A and C1 are shaded in grey, and verbs typical of C1 compared to A and B1 are shaded in dark grey. Also from this perspective, B1 is clearly positioned between A and C1; *define*, *use*, *evaluate*, *describe* are typical of A and B1 in contrast to C1, and *examine* is typical of B1 and C1 in contrast to A. At a more abstract level, we can say that the types of activities authors typically engage in in computer science texts (A) are of a formal nature (*prove*, *define*) whereas the authors in computational linguistics texts (B1) act experimentally *(collect*, *examine*), and in linguistics (C1) they act verbally (*argue*) as well as cognitively (*see*, *feel*).

4. Summary and discussion

In this paper, we have explored selected data mining techniques for the purpose of analyzing register discrimination. The overarching question we are interested in is whether in a situation of register contact between scientific disciplines something like ‘interdisciplinary language’ is emerging. Addressing this question involves register comparison. At a more technical level, this is an exercise in corpus comparison. In order to carry out corpus comparisons, we have set up an infrastructure for corpus processing which includes some standard tools, such as sentence splitting, tokenization, part-of-speech tagging etc. (see Teich (2009) for

a typical processing pipeline). This provides the basis for corpus query and data analysis employing methods of data mining.

The crucial issue in this endeavor is how to discover those linguistic features that are good indicators of register differences, provided they exist. We have explored a set of features potentially distinguishing between the meta-register of scientific writing and other registers, on the one hand, and between individual registers of scientific writing, on the other hand. In the first type of analysis, we have compared the DaSciTex corpus with the registerially-mixed FLOB corpus with regard to the properties of abstractness, technicality and informational density using the following features: part-of-speech distribution, type-token ratio and number of lexical items per clause. To determine the discriminatory force of these features, we have employed three data mining methods: feature ranking, clustering (an unsupervised machine learning method) and classification (a supervised machine learning method). The results are consistent on all three methods providing evidence that DaSciTex is well distinguished from FLOB according to the selected features. In the second type of analysis, we have compared the subcorpora in DaSciTex with regard to selected lexico-grammatical features (nouns, verbs, *we*+verb) using classification. Here, the results are also conclusive: the subcorpora are distinguished well lexically. While this would be more or less expected because nouns and verbs are the main carriers of subject matter (i.e., they realize field of discourse), what is interesting to observe are the misclassifications arising between the subcorpora. We have suggested that these misclassifications might be systematic, indicating a similarity between some subcorpora but not others. Here, the consistent pattern for all analyzed feature sets is that the mixed disciplines (B corpora), the pure disciplines (C corpora) and computer science (A corpus) are well distinguished from one another, while at the same time the mixed disciplines are more similar to their corresponding pure disciplines than to computer science, and the engineering disciplines exhibit the largest similarity. This tendency is corroborated by a number of other studies in which we looked at the grammatical preferences of selected parts-of-speech (e.g., noun/verb colligations; Holtz and Teich, in preparation). These analyses were conducted using more traditional descriptive statistical techniques as well as similarity measures between corpora. The results point in the same direction: triples of A-B-C corpora are clearly distinct with regard to the features investigated according to various measures (including e.g., chi-square, cosine distance, classification) *and* the B corpora (mixed disciplines) are “closer” to the C corpora (pure disciplines) than to the A corpus (computer science). At a more abstract level, this means that register contact results in mixed registers which, however, cannot deny their points of origin. Finally, investigating *we*+verb, we have found some colligations specific to the mixed disciplines. Provided that more evidence can be found of such distinctive lexico-grammatical patterns, one could conclude that interdisciplinary registers negotiate between integration and identification: They integrate the linguistic conventions of two different

disciplines, while at the same time attempting to create their own identity as a unique discipline.

To assess the proposed methodology, three types of comments are in place. First, concerning the application of data mining, while simple descriptive statistics, such as a statistical test on a feature distribution, fulfills a similar purpose, there is a two-fold added value in using data mining methods. In addition to ranking features by their individual discriminatory power, we can explore their *collective* contribution to register discrimination (multivariate analysis). Also, having available information about misclassifications in the form of the confusion matrix, we can investigate the context of typical features/terms in correctly classified and in misclassified texts, analyzing differences and commonalities between registers at class level as well as at instance level. From the perspective of data mining, register analysis is an interesting application, because we are not primarily concerned with finding the most discriminatory feature set for optimal classification, but rather with analyzing patterns of misclassifications for understanding register diversification. Second, from the viewpoint of register analysis, the features we have investigated here are obviously rather shallow, operating with strings and parts-of-speech. As part of the methodology, these have to be related to contextual configurations in a more principled way (consider again Biber's work on interpreting feature bundles in terms of more abstract dimensions of register variation). While words exhibit a strong discriminatory force between registers (i.e., an analysis on the basis of words provides a good classification result), they do not offer much interpretation space other than in terms of *field* variation ('a text t1 *is about* x and a text t2 *is about* y') (cf. again Section 2). Lexis-in-context (e.g., word/part-of-speech bi-grams), on the other hand, may not be such a strong discriminator (i.e., there will be more misclassifications), but it offers more interesting interpretation directions ('a text t1 *construes* x as y', 'a text t2 *construes* x as z'). This is critical to get a grasp of other parameters of register diversification, such as writer-reader role relations (*tenor* variation) or textual organization (*mode* variation).[ix] Third, from the perspective of Systemic Functional Linguistics, we have proposed a possible operationalization for modeling register variation which allows interpreting feature distributions in terms of their contributions to register diversification. Here, a crucial aspect is the opportunity of representing the inherent fuzziness of registers: Any one text purportedly belonging to a particular register may be more or less exhibiting the linguistic properties typically ascribed to that register. This can be read off the confusion matrix, which thus turns out to be a convenient instrument for human inspection of analysis results.

In our future work we plan to carry out more analyses on the basis of aggregated linguistic information (such as part-of-speech n-grams) in order to explore other parameters of variation. For example, in the area of mode of discourse, colligations of nouns at the level of the nominal group could be a source of interesting differences between registers. Also, we are currently annotating a part

of the corpus for process types and Theme-Rheme structure (Schwarz et al., 2008). We plan to use these annotations as a basis for training a classifier to annotate larger amounts of data from the DaSciTex corpus in order to be able to analyze differences and commonalities between registers at higher levels of linguistic organization.

Acknowledgements

We are grateful to *Deutsche Forschungsgemeinschaft* (DFG) who support this work as grant TE 198/1-1 *Linguistische Profile interdisziplinärer Register* (Linguistic Profiles of Interdisciplinary Registers). Many thanks also go to Richard Eckart and Monica Holtz for their collaboration in corpus compilation and basic processing of the corpus, as well as the anonymous reviewer for making helpful suggestions for improving our paper.

References

Biber, D., S. Johansson and G. Leech (1999), *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D. (1988), *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993), 'The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings', *Computers and the Humanities*, 26: 331-345.

Biber, D. (1995), *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Chakrabarti, S. (2003), *Mining the web. Discovering knowledge from hypertext data*. Amsterdam, Boston: Elsevier, Morgan Kaufmann Publishers.

Gries, S. Th. (2006), 'Exploring variability within and between corpora: some methodological considerations', *Corpora*, 1(2):109-151.

Halliday, M.A.K., A. McIntosh and P. Strevens (1964), *The linguistic sciences and language teaching*. London: Longman.

Halliday, M. A. K. and R. Hasan (1989), *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K. and J. R. Martin (1993), *Writing Science. Literacy and discursive power*. London: The Falmer Press.

Halliday, M. A. K. (1985), *Spoken and Written Language*. Victoria: Deakin University Press.

Halliday, M. A. K. (2004), *Introduction to Functional Grammar*. Third edition. London: Arnold.

Halliday, M. A. K. (2005), 'Towards probabilistic interpretations', in: J. Webster (ed.) *Computational and quantitative studies*, volume 6 in the collected works of M. A. K. Halliday, London and New York: Continuum. Reprint of Halliday M. A. K. (1991), 'Towards probabilistic interpretations', in:

E. Ventola (ed.) *Functional and systemic linguistics: approaches and uses*, Berlin and New York: Mouton de Gruyter.

Holtz, M. and E. Teich (in preparation), 'Scientific registers in contact: An exploration of the lexico-grammatical properties of interdisciplinary discourses'.

Kilgarriff, A. (2001), 'Comparing Corpora', *International Journal of Corpus Linguistics*, 6(1): 1-37.

Mair, C. (2006), *Twentieth-Century English. History, variation and standardization*. Cambridge: Cambridge University Press.

Mair, C. (2009), 'Corpora and the Study of Recent Change', in: A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics*, HSK – Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science, Berlin: Mouton de Gruyter, 1109-1125.

Manning, C. and H. Schütze (1999), *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT Press.

Matthiessen, C. M. I. M. (1993), 'Register in the round, or diversity in a unified theory of register', in: M. Ghadessy (ed.) *Register analysis. Theory and Practice*, London: Pinter, 221-292.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman.

Schmid, H. (1994), 'Probabilistic Part-of-Speech Tagging Using Decision Trees', in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, September 1994.

Schwarz, L., S. Bartsch, R. Eckart and E. Teich (2008), 'Exploring automatic theme identification: a rule-based approach', in: A. Storrer, A. Geyken, A. Siebert and K.-M. Würzner (eds.) *Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing (KONVENS 2008)*, Berlin and New York: Mouton de Gruyter., 15-26.

Sebastiani, R. (2002), 'Machine learning in automated text categorization', *ACM Computing Surveys*, 34(1): 1-47.

Teich, E. (2003), *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin and New York: Mouton de Gruyter.

Teich, E. (2009), 'Linguistic computing', in: M.A.K. Halliday and J. A. Webster (eds.) *Companion to Systemic Functional Linguistics*. London: Equinox, 113-127.

Witten, I. H. and F. Eibe (2005), *Data mining. Practical machine learning tools and techniques*. Second edition. Amsterdam, Boston: Elsevier, Morgan Kaufmann Publishers.

Notes

[i] The corpus was compiled from pdf files which were automatically transformed into plain text. The resulting data is not completely clean (e.g., erroneous splitting/contraction of tokens). For some types of investigations, one can live with this quality of data, but for others it is crucial to have them absolutely clean. We thus decided to manually clean a one million token extract from the corpus which is referred to here as the "small version".

[ii] FLOB was chosen for two reasons: First, it was readily available for us to carry out our own processing (pos-tagging, parsing, manual annotation); second, it is comparable in size to the small version of DaSciTex.

[iii] The reported accuracy of the TreeTagger is 96%.

[iv] Information Gain measures the reduction of uncertainty about a class *C* (e.g., FLOB vs. DaSciTex) when knowing an attribute *A* (e.g. STTR); more formally, it is defined by $H(C) - H(C|A)$, with $H(C)$ the entropy of *C*, and $H(C|A)$ the conditional entropy of *C* given *A*. The T-Test tests whether the observed means of two (normally distributed) populations differ significantly. We employed Welch's T-Test, assuming unequal variances. Note that unlike Information Gain, T-Test takes into account sample size.

[v] To avoid the effect of different text lengths on type-token ratio, we have employed the standardized TTR.

[vi] A linear support vector machine is a linear classifier that separates two classes with a maximum margin between the two classes. We have used the standard SVM implementation shipped with Weka. Other classifiers such as naïve Bayes and decision tree learners achieve a similar accuracy. The clustering accuracy is lower, because clustering operates unsupervised, whereas classification operates supervised.

[vii] Gries (2006) provides a more systematic account on analyzing the variability of a single feature in a hierarchical corpus organized into registers and subregisters.

[viii] In more detail, the term frequencies are measured by TF/IDF (term frequency by inverse document frequency). Classification is performed by means of ten-fold cross validation, i.e., the sample is systematically split ten times into 90% training data, and 10% testing data, and accuracies are averaged over the ten runs. Both are fairly standard procedures in the field of text classification.

[ix] A similar position is adopted by other approaches which acknowledge the importance of investigating the interplay of lexis and grammar (colligation and related notions) in linguistic analysis (e.g., Pattern Grammar or Construction Grammar).