Scientific registers and disciplinary diversification: a comparable corpus approach

Elke Teich
Universität des Saarlandes
e.teich@mx.uni-saarland.de

Stefania Degaetano-Ortlieb
Universität des Saarlandes
s.degaetano@mx.uni-saarland.de

Hannah Kermes
Universität des Saarlandes
h.kermes@mx.uni-saarland.de

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
e.lapshinova@mx.uni-saarland.de

Abstract

We present a study on linguistic contrast and commonality in English scientific discourse on the basis of a monolingually comparable corpus. The focus is on selected scientific disciplines at the boundaries to computer science (computational linguistics, bioinformatics, digital construction, microelectronics). The data basis is the English Scientific Text Corpus (SCITEX) which covers a time range of roughly thirty years (1970/80s to early 2000s). In particular, we investigate the disciplinary diversification/relatedness of scientific research articles in terms of register. Our results are relevant for research on multilingually comparable corpora as used in machine translation and related research, since they shed new light on the notion of ‘comparability’.

1 Introduction: Motivation and Goals

In the context of statistical machine translation, comparable corpora are typically bilingual, thematically similar corpora being utilized to extract translation equivalents to enrich translation models. These have proved to be useful, especially for technically specialized texts or for low resource languages where parallel corpora are rare (Chiao and Zweigenbaum (2002); Babych et al. (2007)).

The overarching goal of the paper is to provide evidence that the notion of comparability commonly used in that context is rather coarse and misses important aspects of linguistic variation. We report on a set of experiments in which a monolingually comparable corpus is studied. The corpus contains specialized, technical texts from nine scientific disciplines, related to each other by “interdisciplines” (such as computer science - linguistics - computational linguistics) (cf. Section 2 for details). Our study establishes the linguistic differences and commonalities between the disciplines considered on the basis of the concept of register, i.e., language variation according to situational context. Situational context is conventionally described in terms of field, tenor and mode of discourse (Quirk et al., 1985). It has been shown in numerous corpus-linguistic studies that particular situational settings have specific linguistic correlates at the level of lexico-grammar in the sense of clusters of lexico-grammatical features that occur non-randomly (see notably the work by Biber and colleagues, e.g., Biber (1988, 1993); Biber et al. (1999); Biber (2006, 2012)). Collectively, the linguistic features associated with field, tenor and mode then give rise to registers. More specifically, field of discourse relates to the topic of a discourse and is realized lexico-grammatically in functional verb classes (e.g., activity, communication, etc.) with corresponding arguments (e.g., Actor, Goal, Medium, etc.) and adjunct types (e.g., Time, Place, Manner, etc.). Tenor of discourse relates to the roles and attitudes of the participants in a discourse and is realized lexico-grammatically in mood, modality as well as stance expressions. Mode of discourse relates to the presentational function of language and is realized in Theme-Rheme and Given-New constellations. A register is then characterized by particular distributions of lexico-grammatical features according to a given contextual configuration.

Apart from exhibiting differences in field, tenor and mode, scientific texts are associated with particular discourse “styles” such as technicality, abstractness or informational density, which may again be linguistically realized in different ways and to different degrees across disciplines. Furthermore, in a highly dynamic social domain, such as the scientific one, both registers and discourse styles are relatively versatile and subject to change (cf. Ure (1971, 1982)). This may, for instance,
affect conventional phraseology. Finally, register and stylistic features may be distributed unevenly across document parts, thus giving rise to variation according to document structure. In order to arrive at a comprehensive picture of the linguistic construal of disciplinarity, we thus need to consider the linguistic encodings according to register and the linguistic realization of discursive styles as well as take into account the inherently dynamic nature of scientific discourse.

Relating this back to the notion of comparability, the concept of register may thus provide the basis for a fine-grained description of comparability, as it acknowledges the multi-dimensional nature of linguistic variation.

Our methodology is informed by three sources: corpus linguistics, linguistic theory and data mining. Standard corpus methods are employed for the quantification of instances of linguistic features that are considered to be relevant indicators of variation across scientific disciplines and may be expected to significantly contribute to differences in language use across disciplines. The theoretical basis is provided by Systemic Functional Linguistics (SFL; Halliday (2004)). The reason for choosing SFL to inform analysis is its model of association of contextual variables with lexical-grammatical domains (cf. above on the notion of register).

In contrast to other corpus-based studies on register, our goal is not to uncover dimensions of variation or to discover text classes (as e.g. in Biber et al.’s work). The texts in our corpus are taken from 38 journals from nine disciplines (for details see Section 2) and the text classes are thus extrinsically defined. We can then think of analysis as a task of text classification, where we test whether the extrinsically defined classes have distinctive linguistic correlates and if so, how well the classes are distinguished linguistically and which features contribute most to their distinction. To this end, we employ data mining techniques, in particular automatic text classification (see Section 3 for details). A similar approach to the one developed here, also working on linguistic variation in the scientific domain, has been proposed earlier by Argamon et al. (2008). There is related work in translation studies by Baroni and Bernardini (2006) and Volansky et al. (2011), which uses automatic text classification to describe the specific properties of translations (‘translationese’). The earliest work, to our knowledge, combining SFL with text classification is Whitelaw and Patrick’s work on spam detection (Whitelaw and Patrick, 2004).

2 Corpus

2.1 Corpus Design and Pre-processing

We have built a corpus composed of English scientific research articles — the English Scientific Text Corpus (scitex; cf. Teich and Fankhauser (2010) and Degaetano-Ortlieb et al. (forthcoming)) — that covers nine scientific domains and amounts to approx. 34 million tokens, drawn from 38 sources. SCITEX contains full journal articles from two time periods, the 1970s/early 1980s (SASCITEX) and the early 2000s (DASCITEX). We selected at least two different journals for each discipline in both time slices. As our focus is on se-

Figure 1: Scientific disciplines in the SCITEX corpus

selected scientific domains at the boundaries to computer science and some other discipline, SCITEX has a three-way partition: (1) A-subcorpus: computer science, (2) B-subcorpus: computational linguistics, bioinformatics, digital construction and microelectronics, and (3) C-subcorpus: linguistics, biology, mechanical engineering and electrical engineering, as shown in Figure 1. In the present paper, we are mainly interested in the linguistic evolution of the inter-/transdisciplinary domains represented by the B-subcorpus, as these are the ones that have emerged in the given time frame (1970s/80s to present). We term these domains contact disciplines, since they have come about through contact between two existing dis-
ciplines (here: computer science and another es-
tablished discipline represented in the A and C
subcorpora, which we term seed disciplines). The
main question we are interested in is whether the
seed and contact disciplines have clearly distin-
guishable linguistic correlates in terms of register.

The text sources for SCITEX are full academic
articles in the form of PDF files. These files were
converted to plain text using an existing commer-
cial software including optical character recogni-
tion (OCR).

In further processing we follow the common
practices in corpus linguistics by (a) accounting
for relevant metadata (e.g., author, title, jour-
nal, year of publications) and document structure
(e.g., abstract, conclusion), and (b) using standard
tools for preprocessing (e.g., tokenization, tagging, lemmatization, etc.). For corpus query,
we employ the Corpus Query Processor (CQP)
(CWB; Evert, 2004) which works on the basis of
regular expressions. Utilities of CQP allow for the
extraction of distributional information according
to the annotated metadata and document structure.

3 Methods of Analysis

We carry out a diachronic analysis comparing the
two time slices (1970s/80s vs. 2000s) represented
in the SCITEX corpus, aiming to provide answers
to the following questions:

1. How well are the individual disciplines distin-
guished?
2. How distinct are the contact disciplines from
their seed disciplines?

Thus, analysis involves comparisons along the
temporal and the disciplinary dimensions.

The hypothesis we have about the outcomes of
our analysis is that disciplines will be better dis-
tinguished from one another over time, including
the contact disciplines, reflecting a process of di-
versification within scientific writing over time.

3.1 Feature Selection

In the first step of analysis we need to determine
which features to investigate. These should be fea-
tures that bring out relevant and significant con-
trasts along the dimensions considered (time, dis-
cipline). For the choice of features potentially
distinguishing individual (scientific) registers, we
draw on SFL’s model of register variation in which
the contextual parameters of field, tenor and mode
are associated with particular lexico-grammatical
domains. Since we want to cover all three con-
textual parameters, we choose at least one fea-
ture for each. For field, we analyze functional
verb classes as well as PoS-patterns that are poten-
tially terminology-forming (e.g. noun-noun struc-
tures); for tenor, we analyze modal verbs and for
mode we analyze theme type as well as conjunc-
tive cohesive relations. As another feature, we an-
alyze n-grams on the basis of PoS combinations
(rather than words), since we have seen in a previ-
ous study that they may be involved in processes
of conventionalization (Kermes and Teich, 2012).

Additionally, on an abstract level, scientific
writing is a highly informational production that is
characterized by technicality, information density
and abstractness (cf. Halliday and Martin (1993)).
Among the linguistic features realizing these prop-
erties are a relatively low type-token ratio (techni-
cality), a relatively high lexical density and low
grammatical intricacy (information density) and
the frequent use of nominal categories (nouns, ad-
jectives) (abstractness).

Table 1 displays the features considered in the
analysis together with their associated contextual
variables and/or abstract discourse properties they
 instantiate. Features are extracted from the cor-
pus with CQP. For example, simple queries com-
bine part-of-speech and concrete lemmas (e.g.,
[pos="MD" & lemma="must|should"]; for modal
verbs). More complex queries work with posi-
tional attributes, linguistic annotations and lists
(e.g., <s>\[conj & lemma!=\$modal-adverbs\]... as
part of the extraction of textual Theme, which is
realized in English as the first constituent in the
clause).

3.2 Feature Evaluation

We employ statistical and machine learning meth-
ods to measure (a) how much individual features
contribute to a possible distinction and (b) how
well corpora are distinguished by these features.
We employ classification techniques by using fea-
ture ranking (Information Gain) to determine the
relative discriminatory force of features, and su-
ervised machine learning (decision trees and sup-
port vector machines) to distinguish between the
scientific registers in SCITEX. For these steps we
use the WEKA data mining platform (Witten and
Eibe, 2005).
4 Results and Interpretation

Our analysis addresses the question of how distinctive the subcorpora in SCITEX are comparing the productions of the 1970/80s with those of the early 2000s. Considering the diachronic perspective, we expect to encounter a clearer separation of individual disciplines overall reflecting a process of diversification within scientific writing.

The analysis has two parts: First, we calculate Information Gain of the top twenty features, to see which features are the most discriminatory ones across disciplines. Second, we apply automatic classification, to see how well the subcorpora are distinguished on the basis of these features.

Table 2 shows the twenty most discriminatory features for the 70/80s across all subcorpora. The five highest ranking features are associated with field (NN: IGain 0.39, LEX: IGain 0.36, communication verbs: IGain 0.31) and mode (WL: IGain 0.33, LEX/C: IGain 0.32). In the mid range, we find some tenor features and in the lower range some other field features as well as document structure features.

When we compare these results with the ones for the early 2000s (see Table 3), three main observations can be made. First, features become much more pronounced, the IGain values rising substantially for the top 20 features (1970s/80s are in the range of 0.23 to 0.39, 2000s are in the range of 0.31 to 3.1). This includes the nine features that are identical across SASCITEX and DASCITEX: existence and communication verbs as well as adj-n term pattern for field, obligation modals for tenor, word and sentence length as well as lexical words per clause for mode, bi-grams for conventionalization, and length of main part for document structure, all become more pronounced in DASCITEX (higher IGains) and thus contribute more to the distinction between disciplines. The second observation is that while in SASCITEX only bi-grams ranges among the top 20 features, in DASCITEX we encounter an increase in the contribution of gram-based features to the DASCITEX-internal distinction. This may point to the greater role of conventionalized language in the distinction between disciplines over time. Terminological studies based on n-grams might indicate a thematic comparability of disciplines. Consider one of the key concepts in computer science, ‘algorithm’. The distribution (per million) across the nine disciplines in DASCITEX varies greatly:

Note again that in our analysis, n-grams are based on parts-of-speech, not words.
computer science (3427), microelectronics (1965), bioinformatics (1913), digital construction (1735), computational linguistics (1124), electrical engineering (955), mechanical engineering (129), biology (59) and linguistics (51). When we look at the top frequent token n-grams in which algorithm participates, we find, for example, ‘approximation algorithm’ which is mostly shared between computer science, the contact disciplines and electrical engineering, ‘learning algorithms’ appears practically everywhere, and ‘alignment algorithm’ is almost only mentioned in computational linguistics and bioinformatics (with a few occurrences in computer science and one in biology). The stylistics across the disciplines is also noteworthy: pure stylistic tri-grams, such as the highly frequent ‘in order to’, ‘the number of’, ‘based on the’, ‘as shown in’, etc., are also good discriminators between different disciplines (cf. Kermes and Teich (2012)). Finally, at the levels of contextual and discourse properties, it can be noted that features associated with information density become better discriminators between disciplines in the 2000s having high IGain values, while tenor features step back decreasing in number, tending towards greater uniformity (only one tenor feature (obligation modals) in the top 20 features in the 2000s compared to four in the 70s/80s).

To see how these data are reflected according to disciplines, we perform classification for both corpora.
Table 4: Confusion matrix with decision tree for the 70/80s (SASCITEX)

| A  | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | total | accuracy in % |
|----|----|----|----|----|----|----|----|----|-------|---------------|
| A  | 156| 0  | 1  | 0  | 3  | 7  | 0  | 8  | 37   | 77.23         |
| B1 | 1  | 25 | 11 | 7  | 27 | 3  | 12 | 15 | 125  | 20.80         |
| B2 | 2  | 277| 13 | 4  | 21 | 3  | 24 | 60 | 361  | 43.21         |
| B3 | 8  | 1 | 47 | 13 | 4  | 32 | 46 | 457| 15 | 59.96        |
| B4 | 0  | 14| 8  | 1  | 49 | 26| 15 | 524| 274 | 58.09        |

Table 5: Confusion matrix with SVM for the 70/80s (SASCITEX)

| A  | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | total | accuracy in % |
|----|----|----|----|----|----|----|----|----|-------|---------------|
| A  | 201| 9  | 7  | 1  | 0  | 4  | 2  | 9  | 230  | 87.39         |
| B1 | 0  | 97 | 19 | 1  | 4  | 1  | 0  | 3  | 137  | 70.80         |
| B2 | 5  | 0  | 14 | 6  | 0  | 18 | 6  | 1  | 319  | 84.33         |
| B3 | 3  | 8  | 0  | 8  | 2  | 16 | 30 | 14 | 242  | 69.42         |
| B4 | 2  | 2 | 17 | 0  | 8  | 9  | 1  | 205| 76.10 | 60.67         |

Table 6: Confusion matrix with SVM for the early 2000s (DASCITEX)

| A  | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | total | accuracy in % |
|----|----|----|----|----|----|----|----|----|-------|---------------|
| A  | 135| 6  | 5  | 0  | 9  | 5  | 0  | 111| 81.08 |               |
| B1 | 0  | 7  | 2  | 2  | 3  | 3  | 1  | 351| 95.44 |               |
| B2 | 4  | 1  | 7  | 23| 6  | 0  | 15 | 303| 75.58 |               |
| B3 | 2  | 3 | 42 | 7  | 0  | 4  | 34 | 223| 59.67 |               |

SASCITEX and DASCITEX, first, with decision trees, as they use Information Gain, and second, with support vector machines (SVMs), as they are used for text categorization tasks with many relevant features achieving good results (cf. Joachims (1998)). Classification is performed on all features with 10 fold cross-validation. Table 4 shows the confusion matrix for all subcorpora for the 70/80s and classification accuracy for each subcorpus achieved by decision tree. The overall accuracy is 44.79% only, the correctly classified texts lying on the main diagonal of the matrix.

The confusion matrix produced by SVM is shown in Table 5, with an overall accuracy of 65.07%. Apart from computational linguistics (B1), accuracy goes up by about 10% for digital construction (B3) and linguistics (C1) and about 25-30% for the other subcorpora compared to decision tree. Accuracy with SVM for the contact disciplines (B1-B4) ranges from 20-60% and is much lower than the accuracy achieved for the seed disciplines (A and C1-C4) with around 54-86%. Thus, the contact disciplines are not clearly separated from the seed disciplines. Considering, for instance the triple A-B1-C1, we can see that more texts belong to computational linguistics (B1) are classified into linguistics (C1) than into computational linguistics (27 texts in C1 vs. 26 in B1), i.e., texts in B1 seem to be quite similar to

poras (SASCITEX and DASCITEX).
|       | B1 vs A | B2 vs A | B3 vs A | B4 vs A |
|-------|---------|---------|---------|---------|
| WL    | 0.629   | 0.501   | 0.399   | 0.883   |
| STTR  | 0.509   | LEX     | 0.355   | LEX     |
| LEX   | 0.372   | causal  | 0.334   | n-grams,6 |
| ADJ   | 0.261   | n-grams,6 | 0.306   | STTR    |
| VV    | 0.230   | STTR    | 0.303   | clauses/S |
| n-grams,6 | 0.205   | n-grams,4 | 0.284   | adj-n-n |
| causal | 0.187   | temporal | 0.283   | causal  |
| types | 0.174   | n-grams,5 | 0.282   | NN      |
| adj-c-adj-n | 0.145   | ADJ     | 0.273   | n-grams,4 |
| introduction | 0.129   | causative| 0.197   | ADJ     |
| clauses/S | 0.230   | NN      | 0.269   | LEX/S   |
| ADV   | 0.204   | MD      | 0.264   | main    |
| LEX/C | 0.196   | WH      | 0.198   | n-grams_main |
| NN    | 0.179   | permission | 0.178   | introduction |
| WH/S  | 0.122   | volition | 0.166   | causative |
| LEX   | 0.120   | WL      | 0.147   | exper-theme |
| occurrence | 0.119   | SL      | 0.145   | obligation |
| commun. | 0.112   | WH/S    | 0.137   | n-grams_intro |
| MD    | 0.110   | LEX     | 0.104   | aspectual|
| n-grams_abstr | 0.108   | LEX/C   | 0.098   | LEX/C   |

**A:** Computer Science, B1: Computational Linguistics, B2: Bioinformatics, B3: Digital Construction, B4: Microelectronics, C1: Linguistics, C2: Biology, C3: Mechanical Engineering, C4: Electrical Engineering

Table 7: Feature ranking with IGain for the 70/80s (SASCITEX): Top 20 features contact vs seed disciplines

|       | B1 vs A | B2 vs A | B3 vs A | B4 vs A |
|-------|---------|---------|---------|---------|
| WL    | 0.694   | 0.701   | 0.567   | 0.791   |
| STTR  | 0.631   | main    | 0.680   | causal  |
| SL    | 0.441   | STTR    | 0.678   | STTR    |
| types | 0.402   | n-grams_main | 0.634   | temporal |
| causal | 0.237   | causal  | 0.621   | n-grams,4 |
| n-grams,6 | 0.217   | n-grams,4 | 0.577   | n-grams |
| n-n   | 0.192   | n-grams | 0.552   | n-grams,5 |
| adv   | 0.171   | abstract | 0.537   | n-grams_main |
| adversive | 0.128   | bi-grams | 0.521   | LEX     |
| adj-c-adj-n | 0.125   | introduction | 0.487   | bi-grams |
| occurrence | 0.264   | SL      | 0.366   | WL      |
| adj-adj-n | 0.193   | abstract | 0.518   | VV      |
| ADV   | 0.189   | n-grams_abstr | 0.505   | obligation |
| ADJ   | 0.137   | main    | 0.412   | LEX/C   |
| NN    | 0.128   | introduction | 0.353   | n-grams,5 |
| types | 0.123   | n-grams_main | 0.344   | MD      |
| LEX/C | 0.123   | n-grams_intro | 0.321   | ADJ    |
| main  | 0.118   | WH      | 0.204   | aspectual |
| commun. | 0.107   | MD      | 0.202   | SL      |
| abstract | 0.107   | WH/S    | 0.192   | LEX/S   |

**A:** Computer Science, B1: Computational Linguistics, B2: Bioinformatics, B3: Digital Construction, B4: Microelectronics, C1: Linguistics, C2: Biology, C3: Mechanical Engineering, C4: Electrical Engineering

Table 8: Feature ranking with IGain for the early 2000s (DASCITEX): Top 20 features contact vs seed disciplines

In order to check the separation of disciplines over time, we need to compare classification results across SASCITEX and DASCITEX. We again apply SVM, which returns an overall accuracy of **78.17%**.\(^2\) Comparing the values for the individual subcorpora across SASCITEX and DASCITEX, we can observe that accuracies are now much higher for all subcorpora. Considering the contact disciplines, they have clearly gained distinctiveness in the 2000s in comparison to the 1970/80s, as texts in B1-B4 are classified correctly 69% to 84% of achieving an accuracy of 57.24% only.

\(^2\)Decision tree performed poorly again in comparison
the time (instead of 20-60% in the 1970/80s).

In summary, the classification results match the results obtained by feature ranking, which have shown that the top 20 features increased discriminatory force over time. This is reflected by a higher classification accuracy overall and for the subcorpora. The discriminatory force of features in the 1970s/80s instead, was not strong enough to clearly separate disciplines.

To see whether there are any particular features involved in the differentiation of the contact disciplines in particular vis-à-vis computer science on the one hand and the other seed disciplines on the other hand, we inspect the confusion matrices as well as the IGains of each B vs. A and each B vs. the respective C, both for SASCITEX and DASCITEX. In the comparison to computer science (A), we can see that the confusion matrices produced with SVM (cf. Table 5 and 6) show few texts that are misclassified from the contact disciplines (Bs) into computer science (A) for both time slices. Thus, the features employed distinguish Bs from A quite well. Considering the IGain values (see Table 7 and 8 for the top 10 features), besides computational linguistics (B1; relatively low classification accuracy of 20% in the 70/80s), the contact disciplines have the following features in common: word length (WL), STTR, causal verbs in the top 10 as well as four-grams, lexical words (LEX) and temporal conjunctions in the top 20 features. Except lexical words (LEX), all features have a higher IGain in the 2000s. In the comparison to the other seed disciplines (Cs), the confusion matrices show more misclassifications of Bs into Cs. Considering the IGain values there are no tendencies uniformly applying to the contact disciplines (Bs). They rather show individual tendencies for each pair (B1 vs. C1, B2 vs. C2, B3 vs. C3, B4 vs. C4). Features that contribute to a better classification diachronically lie in the following parameters: (a) field (occurrence, term-patterns, ADV) for computational linguistics (B1), (b) document structure (abstract, main, intro), information density (SL) and conventionalization (n-grams_abstract) for bioinformatics (B2), (c) information density (WL) and technicality (VV) for digital construction (B3) and microelectronics (B4).

5 Summary and Conclusions

We have looked at disciplinary linguistic diversification in English scientific writing in terms of register, discourse styles and document structure. The results of our analysis provide evidence of major motifs of development in scientific writing over time, showing dynamicity over a time span of only thirty years. Diversification over time is clearly borne out for the contact disciplines but is also true for most of the other disciplines.

Considering the contact disciplines we have seen that (1) they can be distinguished quite well from computer science with the same features being involved in better classification results, (2) they show individual feature constellations in their distinction from their seed disciplines. Moreover, n-grams have gained discriminatory force over time and are ranked relatively high among our features in the 2000s subcorpus. As they are also relevant in terms of terminology, they give an insight in the relatedness of disciplines.

In terms of methods, we have combined state-of-the-art corpus processing with techniques of data analysis as developed in data mining. As such techniques become more accessible to linguistic, literary and cultural analysis, the repertoire of methods for such analysis will be greatly enhanced in that sounder empirical evidence can be sought in text-based socio-cultural and historical studies at large (cf. Jockers (2013)). The crucial factor in employing such methods is the motivation of the features to be used in analysis. Here, we have deliberately not relied on word-based features but instead mainly employed lexico-grammatical patterns. While bags-of-words are strong discriminators between texts/text classes, they can only tell us something about lexical variation (e.g., as an indicator of text topic). However, when register or style rather than topicality are in the focus (such as e.g. the linguistic construal of technical, dense or abstract discourse or the expression of field, tenor or mode relations), it will not be sufficient to study lexical word distributions (cf. Cohen et al. (2010); Teich and Fankhauser (2010) for some other studies). Instead, one needs to identify lexico-grammatical patterns that are potential indicators of the more abstract discursive and contextual properties that are in focus.

The insight to be gained from our study for multilingually comparable corpora is that more elaborate definitions of ‘comparability’ might be re-

---

3There are only two exceptions: C1 (linguistics) goes slightly down (around 2.5%), C4 (electrical engineering) goes down by over 20% to 50.67% accuracy, i.e., it is not really distinguishable any more.
quired. Our approach offers such a definition of comparability by being firmly based on an established model of linguistic variation, which has also been widely applied in multilingual contexts, such as for example, automatic text generation (see e.g., Matthiessen and Bateman (1991); Bateman (1997); Kruijff et al. (2000)). The parameters of variation we employ (register: field, tenor, mode; discourse styles; time) provide a fine-grained grid of features involved in linguistic variation, which can be applied to other languages as well. For example, we can extract and analyze field features, such as term patterns (as produced for German by Weller et al. (2011)), tenor features, such as modal verbs, as well as the other features investigated using the same tools applied here (part-of-speech tagger, CQP, R-scripts and WEKA modules) with only little adaptations (e.g., tag sets, query formulation). Overall, we would expect that applying the concept of register to the problem of comparability will enable finer-tuned comparable corpora and thus contribute to their fuller potential for multilingual language technology.

Acknowledgments

We wish to thank the anonymous reviewers for their helpful comments. We are especially grateful to Peter Fankhauser for critically assessing our data and to Noam Ordan for valuable suggestions regarding the structure of the paper.

References

Shlomo Argamon, Jeff Dodick, and Paul Chase. Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics*, 75(2):203–238, 2008.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. Translating from under-resourced languages: Comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen, Denmark, 2007.

Marco Baroni and Silvia Bernardini. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, 2006.

John A. Bateman. Enabling technology for multilingual natural language generation: The KPML development environment. *Journal of Natural Language Engineering*, 3(1):15–55, 1997.

Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

Douglas Biber. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5-6):331–345, 1993.

Douglas Biber. *University Language: A Corpus-based Study of Spoken And Written Registers*, volume 23 of *Studies in Corpus Linguistics*. John Benjamins Publishing, Amsterdam/Philadelphia, 2006.

Douglas Biber. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1):9–37, 2012.

Douglas Biber, Stig Johansson, and Geoffrey Leech. *Longman Grammar of Spoken and Written English*. Longman, Harlow, 1999.

Yun-Chuang Chiao and Pierre Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international Conference on Computational Linguistics (COLING)*, Vol. 2, pages 1–5, Taipei, Taiwan, 2002.

Kevin Bretonnel Cohen, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11(1):492, 2010.

CWB. The IMS Open Corpus Workbench, 2010. http://www.cwb.sourceforge.net.

Stefania Degaetano-Ortlieb, Kermes Hannah, Ekaterina Lapshinova-Koltunski, and Teich Elke. SciTex a diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP), Vol. 3. Narr, Tübingen, forthcoming.

Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, IMS, University of Stuttgart, 2004.
M.A.K. Halliday. *An Introduction to Functional Grammar*. Arnold, London, 2004.

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

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.

Matthew L. Jockers. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Hannah Kermes and Elke Teich. Formulaic expressions in scientific texts: Corpus design, extraction and exploration. *Lexicographica*, 28 (1):99–120, 2012.

Geert-Jan Kruijff, Elke Teich, John Bateman, Ivana Kruijff-Korbayová, Hana Skoumalová, Serge Sharoff, Lena Sokolova, Tony Hartley, Kamenka Staykova, and Jiří Hana. Multilinguality in a text generation system for three Slavic languages. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Vol. 1, pages 474–480, Saarbrücken, Germany, 2000.

Christian M.I.M. Matthiessen and John A. Bateman. *Text generation and systemic-functional linguistics: Experiences from English and Japanese*. Communication in Artificial Intelligence Series. Pinter, 1991.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, London, 1985.

Elke Teich and Peter Funkhauser. Exploring a corpus of scientific texts using data mining. In S. Gries, S. Wulff, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*, pages 233–247. Rodopi, Amsterdam and New York, 2010.

Jean Ure. Lexical density and register differentiation. In G. E. Perren and J. L. M. Trim, editors, *Applications of Linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, pages 443–452. Cambridge University Press, 1971.

Jean Ure. Introduction: Approaches to the study of register range. *International Journal of the Sociology of Language*, 35:5–23, 1982.

Vered Volansky, Noam Ordan, and Shuly Wintner. More human or more translated? Original texts vs. human and machine translations. In *Proceedings of the 11th Bar-Ilan Symposium on the Foundations of AI with Israeli Seminar on Computational Linguistics (ISCOL)*, Ramat Gan, Israel, 2011.

Marion Weller, Helena Blancafort, Anita Gojun, and Ulrich Heid. 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, 2011.

Casey Whitelaw and Jon Patrick. Selecting systemic features for text classification. In Ash Asudeh, Cécile Paris, and Stephen Wan, editors, *Proceedings of the Australasian Language Technology Workshop*, pages 93–100, Sydney, Australia, 2004.

Ian H. Witten and Frank Eibe. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, Morgan Kaufmann Publishers, Amsterdam, Boston, second edition, 2005.