Data Mining with Shallow vs. Linguistic Features to Study Diversification of Scientific Registers

Stefania Degaetano-Ortlieb*, Peter Fankhauser†, Hannah Kermes*, Ekaterina Lapshinova-Koltunski*, Noam Ordan*, Elke Teich*

*Universität des Saarlandes; †Institut für Deutsche Sprache (IDS)
Universität Campus A2.2, 66123 Saarbrücken, Germany; R 5, 6-13, 68161 Mannheim, Germany
s.degaetano, h.kermes, e.lapshinova, n.ordan, e.teich@mx.uni-saarland.de; fankhauser@ids-mannheim.de

Abstract

We present a methodology to analyze the linguistic evolution of scientific registers with data mining techniques, comparing the insights gained from shallow vs. linguistic features. 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) (Degaetano-Ortlieb et al., 2013; Teich and Fankhauser, 2010). In particular, we investigate the diversification of scientific registers over time. Our theoretical basis is Systemic Functional Linguistics (SFL) and its specific incarnation of register theory (Halliday and Hasan, 1985). In terms of methods, we combine corpus-based methods of feature extraction and data mining techniques.
Keywords: data mining, text classification, register

1. Introduction

Our central research interest is in the evolution of lexico-grammatical patterns in the scientific domain, asking whether individual scientific disciplines develop their own, distinctive linguistic characteristics over time, and if so, what these distinctive characteristics are.

Most obviously, disciplines can be well distinguished by domain specific vocabulary. Thus, a bag-of-words approach as used in text categorization tasks and stylometric studies (e.g., Joachims, 1998; Koppel et al., 2002; Rybicki, 2006; Argamon et al., 2008; Fox et al., 2012), will clearly indicate a distinction between disciplines. Yet, these kinds of approaches do not account for the full potential of language variation according to situational context, which may provide more insights into the evolution of lexico-grammatical patterns. In this paper, we investigate what additional information we can gain from approaches relying on linguistic features rather than shallow (bag-of-words) features.

To investigate the above question, we employ the notion of register, i.e., language variation according to situational context described in terms of *field*, *tenor* and *mode* of discourse (Halliday and Hasan, 1985; Quirk et al., 1985). Numerous corpus-linguistic studies have shown that particular situational settings have linguistic correlates at the level of lexico-grammar in the sense of clusters of lexico-grammatical features that occur non-randomly (see e.g., Biber, 1988; Biber, 1993; Biber, 2006; Biber, 2012). In addition, we consider *time* as another relevant contextual factor in register analysis, as language use continuously adapts to changing social contexts (cf. Ure, 1971; Ure, 1982).

Our methodology is informed by three sources: empirical linguistics (in particular corpus linguistics), linguistic theory and data mining. There is related work in translation studies by e.g., Baroni and Bernardini (2006) or Lembersky et al. (2012). 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. Data

As data basis we use the English Scientific Texts Corpus (SCITEX; cf. Teich and Fankhauser, 2010; Degaetano-Ortlieb et al., 2013) built from full English journal articles, which covers nine scientific domains amounting to around 34 million tokens, drawn from 38 sources.

Our focus lies on selected scientific domains at the boundaries to computer science ('contact' disciplines) and some other 'seed' discipline. This is captured in SCITEX by 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 (see Figure 1). SCITEX comprises two time slices, the 70/80s (SASCITEX) and the early 2000s (DASCITEX), covering a thirty year time span similarly to the Brown corpus family (Kučera and Francis, 1967; Hundt et al., 1999). The corpus has been tokenized, lemmatized and part-of-speech (PoS) tagged using TreeTagger (Schmid, 1994). Additionally, each document has been enriched with meta-information (such as author(s), title, scientific journal, academic discipline, and year of publication) and document structure (e.g., section types, section titles, paragraphs and sentence boundaries). SCITEX is encoded in the Corpus Query Processor (CQP) format (Evert, 2005) and can be queried with CQP by using regular expressions in combination with positional (e.g., PoS) and structural attributes (e.g., sentence, sections).

3. Methods of analysis

We carry out three types of analysis, one based on shallow features, the other on linguistic features, and the third one combines both feature sets, comparing the two time slices

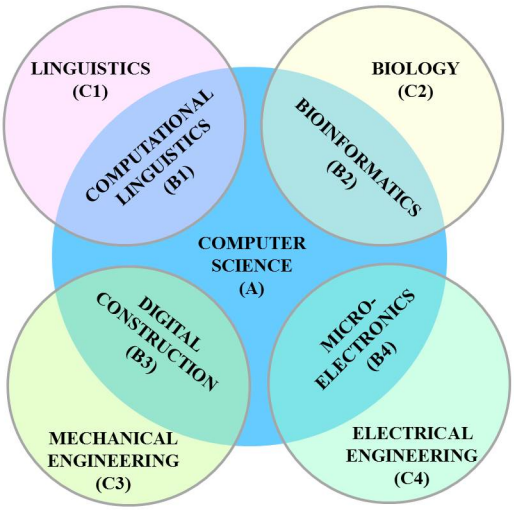

Figure 1: Scientific disciplines in the SCITEX corpus

(1970s/80s vs. 2000s) represented in the SCITEX corpus. We aim to provide answers to the following questions:

1. How do contact disciplines and seed disciplines differ?
2. How do contact disciplines evolve over time?

In all three analysis, we need to compare the B subcorpus (contact disciplines) with the A and C subcorpora (seed disciplines).

Feature selection For the first analysis, we use shallow features that can distinguish between individual registers. As disciplines can be well distinguished by domain specific vocabulary, we choose to use the 500 most distinctive nouns calculated by Information Gain.
For the second analysis, 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 contextual parameters, we choose at least one feature for each (see Table 1). Additionally, we analyze PoS n-grams as well as features for technicality, information density and abstractness which are linguistic features associated with scientific writing (Halliday and Martin, 1993) (also shown in Table 1).
In the third analysis, both feature sets are combined.

Feature evaluation We employ statistical and machine learning methods to measure (a) how well corpora are distinguished by the selected features and (b) how much individual features contribute to the distinction. We employ classification techniques by using feature ranking (Information Gain and SVM weights) to determine the relative discriminatory force of features, and supervised machine learning (support vector machines) to distinguish between the scientific registers in SCITEX on the basis of shallow and linguistic features. For this, we use the WEKA data mining platform (Witten and Frank, 2005).

4. Analyses

Analysis 1: Shallow features In the first analysis, we look at how distinctive the subcorpora of SCITEX are diachronically by shallow features. Here, we consider the 500 most distinctive nouns calculated by Information Gain to classify the texts into the nine disciplines by training and testing a support vector machine classifier.
For the classification of texts of the 70/80s, we achieve a classification accuracy of **91.08%**. This rises slightly up to **91.55%** for the 2000s. Figures 2 and 3 show the confusion matrices for both time slices, respectively. Each row gives the predicted classes for an actual class. The number of correctly classified texts is shown on the main diagonal (in bold). Misclassifications indicate overlaps between the subcorpora. For both time slices, there are three kinds of interesting misclassifications (shown in three shades of gray): (1) electrical engineering (C4) overlaps with other engineering disciplines (A, B1-B4, C3; gray), (2) there are misclassifications between contact disciplines (Bs) and corresponding seed disciplines (A and Cs; dark gray), (3) while the seed disciplines (Cs) are relatively well distinguished from each other, the contact disciplines (Bs) show overlaps (light gray).
Diachronically, misclassifications between contact and seed disciplines slightly rise (1.8%) for both A and Cs (see Table 2). Misclassifications of electrical engineering (C4) into other engineering disciplines decrease as well as misclassifications of contact disciplines into other contact disciplines (see again Table 2).

overlaps	**70/80s**	**2000s**
engineering	3.61	1.81
contact into seed	2.58	4.34
contact into A	0.61	1.65
contact into Cs	1.97	2.69
contact into contact	2.52	1.27

Table 2: Diachronic comparison of misclassifications in % by SVM with shallow features

Thus, contact disciplines become more distinguishable among each other, evolving as disciplines in their own right, but show greater overlap with seed disciplines over time.

Analysis 2: Linguistic features In the second analysis, we compare the subcorpora in SCITEX by linguistic features (see Section 3) to see whether we can gain more insights about the evolution of lexico-grammatical patterns. As in Analysis 1, we perform classification for both time slices with support vector machines (SVMs).
For the 70/80s, we obtain a classification accuracy of **65.07%**, which rises to **77.24%** in the 2000s. As expected this is lower than for the shallow features (see Analysis 1). The confusion matrices (see Figures 4 and 5) show similar tendencies to Analysis 1 (see Table 3 showing misclassifications among engineering disciplines, contact and seed disciplines as well as among contact disciplines). However, the amount of overlap drops for all comparisons over time, except for misclassifications between the contact disciplines and computer science (A). Thus, the contact disciplines are less distinguished diachronically from computer

contextual parameter/ abstract property	feature category	feature subcategory
FIELD	term patterns	NN-of-NN, N-N, ADJ-N
	verb classes	activity (e.g., *make, show*) aspectual (e.g., *start, end*) causative (e.g., *let, allow*) communication (e.g., *note, describe*) existence (e.g., *exist, remain*) mental (e.g., *see, know*) occurrence (e.g., *change, grow*)
TENOR	modality	obligation/necessity (e.g., *must*) permission/possibility/ability (e.g., *can*) volition/prediction (e.g., *will*)
MODE	theme	experiential theme (e.g, *The algorithm...*) interpersonal theme (e.g., *Interestingly...*) textual theme (e.g., *But...*)
	conjunctive cohesive relations	additive (e.g., *and, furthermore*) adversative (e.g., *nonetheless, however*) causal (e.g., *thus, for this reason*) temporal (e.g., *then, at this point*)
TECHNICALITY	type-token ratio lexical vs. function words	STTR no. of lexical PoS categories
INFORMATION DENSITY	lexical density grammatical intricacy	lexical items per clause/sentence clauses per sentence wh-words per sentence sentence length
ABSTRACTNESS	PoS distribution	no. of nominal vs. verbal categories
CONVENTIONALIZATION	n-grams on PoS basis length of sections	2-to-6-grams overall/per section tokens per section

Table 1: Linguistic features used in analysis

	A	B1	B2	B3	B4	C1	C2	C3	C4
A-CompSci	**189**	2	1	6	0	0	0	2	2
B1-CompLing	3	**98**	4	2	0	14	0	0	4
B2-BioInf	2	4	**418**	20	0	0	0	2	11
B3-DigCon	5	3	29	**283**	1	0	0	8	32
B4-MicroElec	0	0	3	12	**246**	0	0	3	8
C1-Ling	1	5	1	0	0	**211**	1	0	0
C2-Bio	0	1	4	1	1	0	**528**	0	0
C3-MechEng	0	0	6	15	1	0	0	**366**	0
C4-ElecEng	4	2	5	32	7	0	0	5	**469**

Figure 2: Confusion matrix (shallow features) with SVM for the 70/80s (SASCITEX)

science (A) not only by nouns shown in Analysis 1, but also by linguistic features.

To learn more about which linguistic features contribute to a better distinction between the contact disciplines and the seed disciplines, we look at their SVM weights. We group the distinctive features for each discipline according to their contextual parameter (field, tenor, mode) or abstract property (e.g., technicality) and calculate the sum of their SVM weights. This is done for each pair of seed vs. contact discipline (A-CompSci vs. B1-CompLing, A-CompSci vs. B2-BioInf, etc.). We then inspect which feature categories contribute most to the distinction.

overlaps	70/80s	2000s
engineering	9.72	5.23
contact into seed	5.97	4.59
contact into A	0.58	1.03
contact into Cs	5.39	3.55
contact into contact	7.10	3.04

Table 3: Diachronic comparison of misclassifications in % by SVM with linguistic features

	A	B1	B2	B3	B4	C1	C2	C3	C4
A-CompSci	**205**	3	2	10	5	0	0	0	5
B1-CompLing	0	**125**	1	0	0	11	0	0	0
B2-BioInf	5	5	**291**	3	1	0	9	3	2
B3-DigCon	8	2	1	**208**	7	1	0	5	10
B4-MicroElec	2	1	0	6	**188**	0	0	0	5
C1-Ling	0	9	0	1	0	**101**	0	0	0
C2-Bio	0	0	15	1	0	1	**334**	0	0
C3-MechEng	0	0	1	2	1	0	1	**290**	8
C4-ElecEng	7	0	1	13	1	0	0	4	**197**

Figure 3: Confusion matrix (shallow features) with SVM for the 2000s (DASCITEX)

For the diachronic comparison between computer science and the contact disciplines, consider Figures 6 and 7. In these bar charts we visualize the sum of SVM weights of distinctive features for computer science (A) on the left-hand side and for the contact disciplines (Bs) on the right-hand side (Bs) for each feature category. The length of the bar indicates the amount of contribution of features for computer science (left) and for the contact disciplines overall (right). The colors indicate which pair the sum of the weights belongs to (e.g., blue for A-CompSci-vs-B1 and B1-CompSci-vs-A). In the 70/80s, computer science (A) is mostly distinguished by conventionalization and mode features. The contact disciplines (Bs), instead, make more use of field, abstractness, and information density features. Thus, computer science uses a more conventionalized language in comparison to the contact disciplines in the 70/80s.[1] Comparing this to the 2000s (Figure 7), we see that conventionalization is less distinctive for computer science (A), but has gained discriminatory force for most of the contact disciplines (B1, B2 and B4). Additionally, while information density features have gained discriminatory force, abstractness and field features show less discriminatory force. This shows some parallels to Analysis 1, where nouns, which also belong to the contextual parameter of field, show a diminished discriminatory force between computer science and the contact disciplines over time.

In the comparison to the other seed disciplines (Cs) (see Figures 8 and 9), there are no tendencies uniformly applying to the contact disciplines (Bs). They rather show individual tendencies for each pair (B1 vs. C1, etc.). Figure 8 shows that in the 70/80s computational linguistics (B1) is for the most part distinguished from linguistics (C1) by field and abstractness features, bioinformatics (B2) from biology (C2) by field, tenor, mode, abstractness, and information density features, digital construction (B3) from mechanical engineering (C3) by mode, tenor, and field features, and microelectronics (B4) from electrical engineering (C4) by abstractness, conventionalization, and information density features. In the 2000s, the contact disciplines are much less distinguished by field features (see Figure 9). Additionally, computational linguistics (B1) has gained discriminatory force in conventionalization features, similarly to bioinformatics (B2). Digital construction (B3) has gained discriminatory force in information density, mode, and field features, and microelectronics (B4) remains discriminated by conventionalization, abstractness, and information density features.

In summary, contact disciplines seem to become more conventionalized and more distinct from each other, but have greater overlaps with the seed disciplines, especially with computer science in terms of field.

Analysis 3: Shallow and linguistic features In the third analysis, we combine shallow and linguistic features to see whether the classification improves.

register	shallow	shallow + ling.	difference
A-CompSci	93.56	91.58	-1.98
B1-CompLing	78.40	68.00	-10.40
B2-BioInf	91.47	89.50	-1.97
B3-DigCon	78.39	75.90	-2.49
B4-MicroElec	90.44	89.34	-1.10
C1-Ling	96.35	92.24	-4.11
C2-Bio	98.69	97.76	-0.93
C3-MechEng	94.33	93.56	-0.77
C4-ElecEng	89.50	92.56	3.06

Table 4: Shallow features vs. shallow and ling. features for the 70/80s

The classification for the 70/80s achieves an overall accuracy of **89.82%** which is much higher than the accuracy of the classification on the linguistic features on their own (compare 65.07% from Analysis 2), but lower than the accuracy of the classification on the shallow features (compare 91.08% from Analysis 1). Thus, the linguistic features do not improve classification for this time period. This is due to the fact that the disciplines in the 70/80s, especially the contact disciplines, are not clearly distinct from one another in terms of linguistic features (see Analysis 2). In the 2000s, classification achieves an overall accuracy of **92.92%** which is higher than the linguistic and shallow features taken on their own (compare 91.55% for the shal-

[1]Note that only the conventionalization feature actually indicates a relative degree of use (high or low). For the other features we only show the level of contribution to a given distinction but not whether that feature is relatively highly or rarely used.

	A	B1	B2	B3	B4	C1	C2	C3	C4
A-CompSci	**156**	0	3	4	0	1	1	0	37
B1-CompLing	1	**26**	23	11	7	27	3	12	15
B2-BioInf	2	2	**274**	47	13	4	32	37	46
B3-DigCon	8	1	72	**156**	21	3	16	24	60
B4-MicroElec	0	1	14	8	**158**	1	49	26	15
C1-Ling	2	11	12	0	0	**183**	0	5	6
C2-Bio	2	0	28	4	12	0	**463**	9	17
C3-MechEng	3	4	53	18	22	2	40	**213**	33
C4-ElecEng	30	2	41	25	12	1	24	12	**377**

Figure 4: Confusion matrix (linguistic features) with SVM for the 70/80s (SASCITEX)

	A	B1	B2	B3	B4	C1	C2	C3	C4
A-CompSci	**200**	0	0	9	7	1	0	2	11
B1-CompLing	4	**96**	5	18	1	9	1	0	3
B2-BioInf	5	0	**265**	14	6	0	18	10	1
B3-DigCon	4	2	10	**169**	6	0	4	33	14
B4-MicroElec	3	2	14	16	**151**	0	7	9	0
C1-Ling	1	9	5	4	0	**92**	0	0	0
C2-Bio	0	0	6	1	2	2	**336**	3	1
C3-MechEng	4	1	9	28	8	0	12	**223**	18
C4-ElecEng	18	3	3	41	7	0	4	43	**104**

Figure 5: Confusion matrix (linguistic features) with SVM for the 2000s (DASCITEX)

register	**shallow**	**shallow + ling.**	**difference**
A-CompSci	89.13	93.04	3.91
B1-CompLing	91.24	89.05	-2.19
B2-BioInf	91.22	95.30	4.08
B3-DigCon	85.95	89.67	3.72
B4-MicroElec	93.07	91.09	-1.98
C1-Ling	90.00	90.09	0.09
C2-Bio	95.16	97.15	1.99
C3-MechEng	95.71	96.04	0.33
C4-ElecEng	88.34	87.44	-0.90

Table 5: Shallow features vs. shallow and ling. features for the 2000s

70/80s		**2000s**	
A-CompSci vs B3-DigCon			
N_fig	1.28	**ttr**	**2.01**
N_manuscript	1.25	**word length**	**1.87**
adj-n-n	**1.13**	**sentence length**	**1.17**
word length	0.92	N_method	**0.93**
n-n	**0.90**	N_solution	0.90
B3-DigCon vs C3-MechEng			
N_program	1.85	**sentence length**	**1.56**
N_computer	1.56	**word length**	**1.12**
N_manuscript	1.31	N_datum	0.96
N_device	1.13	N_function	0.95
N_simulation	1.04	N_algorithm	**0.94**

Table 6: Top 5 most distinctive features for digital construction (B3) across time

low features and 77.24% for the linguistic features). Thus, while the combination of shallow and linguistic features does not have a positive impact on the classification accuracy in the 70/80s, it does provide a better classification in the 2000s. This is also reflected across registers. Consider Tables 4 and 5 which show the accuracies for each classification (shallow vs. shallow + ling.) by register. While in the 70/80s the accuracies for each register are lower for almost all registers in the combined feature classification (except for C4-ElecEng), the accuracies in the 2000s increase for most registers (except for B1-CompLing, B4-MicroElec and C4-ElecEng) with the combined feature classification. Focusing on the disciplines with improved classification accuracy combining shallow features with linguistic features (A-CompSci, B2-BioInf, B3-DigCon, C1-Ling, C2-Bio, C3-MechEng), we can observe that in the 70/80s 0 to 3 features of the top 10 distinctive features are linguistic features, while in the 2000s 2 to 5 out of the top 10 are linguistic features (see Table 6 for an example). Thus, in

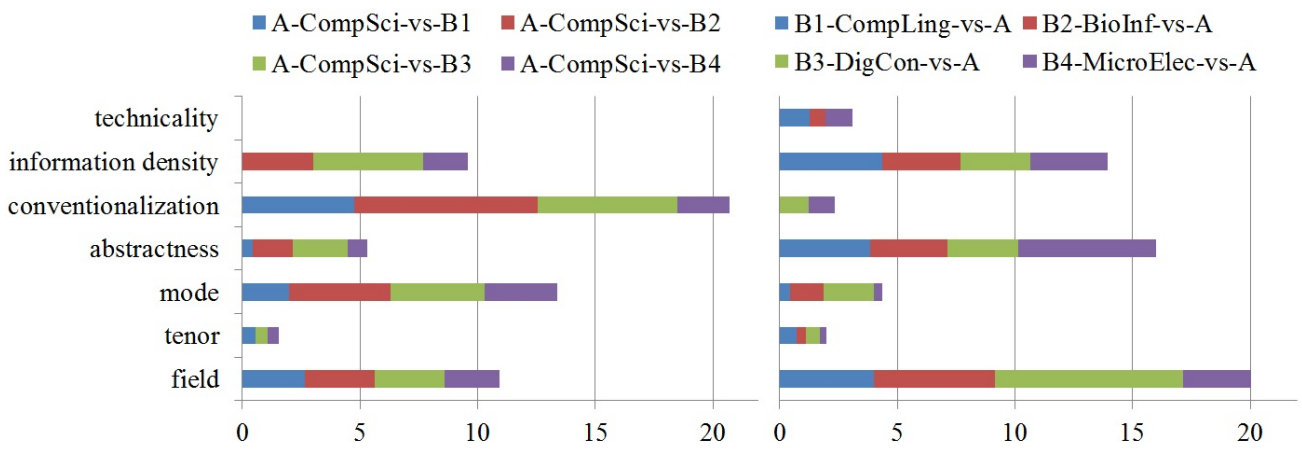

Figure 6: SVM weights for A vs. Bs 70/80s

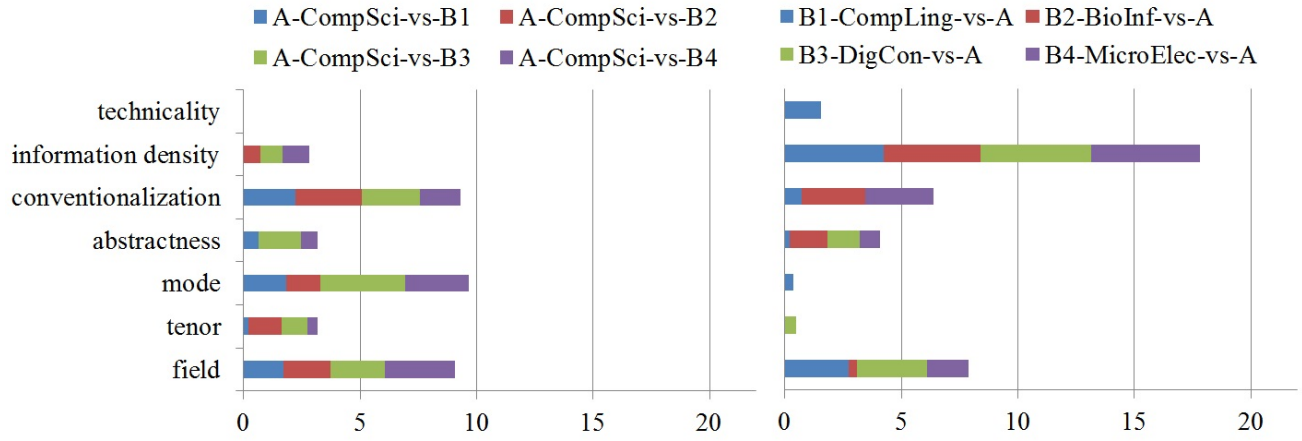

Figure 7: SVM weights for A vs. Bs 2000s

the 2000s linguistic features contribute to a better classification.

5. Conclusions

We have shown how lexico-grammatical patterns in the scientific domain change over time and how individual scientific disciplines emerged by register contact develop their own distinctive linguistic characteristics or adopt them from their seed disciplines.

In terms of methods, we have performed a bag-of-words classification (500 most distinctive nouns as shallow features) and a classification based on linguistic features. While the results on shallow features can only give insights on vocabulary differences between the disciplines (topicality), the linguistic features reveal more fine-grained linguistic differences. In general, we can say that the linguistic distinctness of registers increases. However, while the contact disciplines become more distinct from one another, they do not 'forget' about their seed disciplines as more topics and possibly methods from the seed disciplines are incorporated.

Additionally, we have shown that classification is improved for most disciplines in the 2000s by considering linguistic features on top of shallow features.

6. References

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

Marco Baroni and Silvia Bernardini. 2006. 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.

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

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

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

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

Stefania Degaetano-Ortlieb, Hannah Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich. 2013. SciTex -

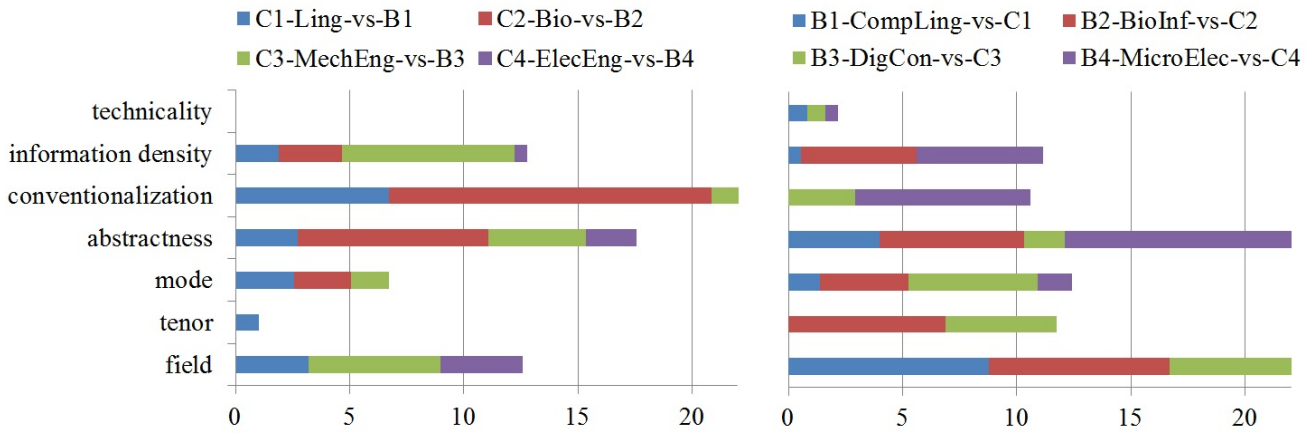

Figure 8: SVM weights for Cs vs. Bs 70/80s

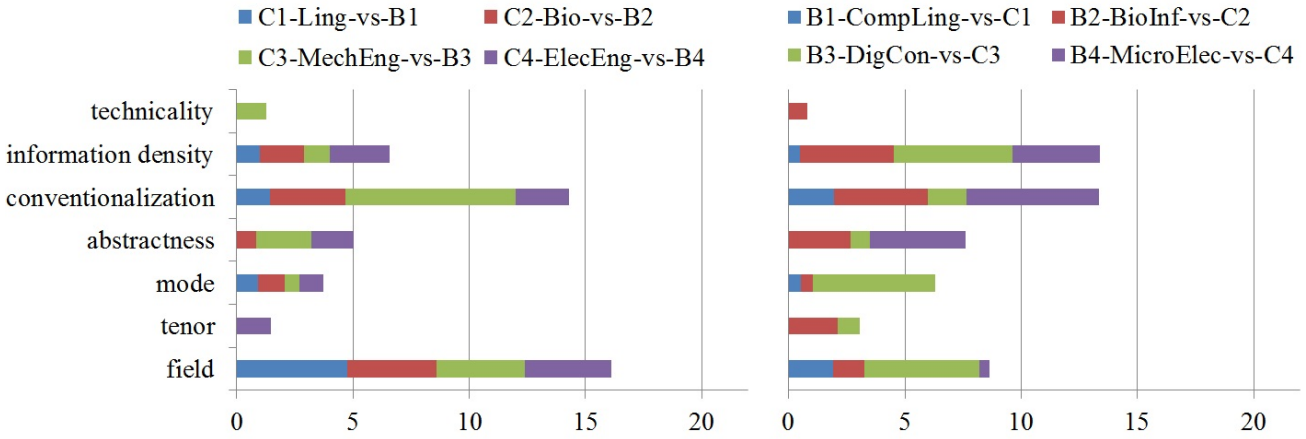

Figure 9: SVM weights for Cs vs. Bs 2000s

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*, volume 3 of *Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*, pages 93–104. Narr.

Stefan Evert, 2005. *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

Neal Fox, Omran Ehmoda, and Eugene Charniak. 2012. Statistical Stylometrics and the Marlowe - Shakespeare Authorship Debate. In *Proceedings of the Georgetown University Roundtable on Language and Linguistics (GURT)*, Washington, D.C, USA.

M.A.K. Halliday and Ruqaiya Hasan. 1985. *Language, Context and Text: A Social Semiotic Perspective*. Language and Learning Series. Deakin University Press, Geelong, Victoria.

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

Marianne Hundt, Andrea Sand, and Rainer Siemund, 1999. *Manual of Information to Accompany The Freiburg LOB Corpus of British English (FLOB)*. Freiburg: Department of English, Albert-Ludwigs-Universität Freiburg.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Machine Learning: ECML-98*, pages 137–142.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

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

Jan Rybicki. 2006. Burrowing into translation: Character idiolects in Henryk Sienkiewicz's trilogy and its two English translations. *Literary and Linguistic Computing*, 21(1):91–103.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Elke Teich and Peter Fankhauser. 2010. Exploring a cor-

pus 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.

Jean Ure. 1971. 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.

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

Casey Whitelaw and Jon Patrick. 2004. 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.

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