Stefania Degaetano-Ortlieb / Hannah Kermes / Ekaterina Lapshinova-Koltunski / Elke Teich

SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers

Abstract
In this paper, we report on a project[1] investigating the diachronic development of scientific registers that have emerged in the last forty years or so through interdisciplinary contact of selected scientific disciplines with computer science (e.g., computational linguistics or bioinformatics). Our main goal is to gain a better understanding of the principles of register formation in highly specialized scientific domains in this kind of context. For this purpose, we have built a diachronic corpus, the English Scientific Text Corpus (SciTex). Our theoretical framework is Systemic Functional Linguistics (Halliday 2004) and register/genre theory (Halliday and Hasan 1989; Biber 1988, 1995; Martin 1992). Methodologically, we adopt a variationist approach, looking at lexico-grammatical differences and commonalities between registers under the perspective of recent language change (cf. Mair 2006).

1. Introduction

The investigation of scientific texts is a very active research area. Studies are carried out on small text samples or on corpora, ranging from the analysis of single registers (Halliday 1988, O'Halloran 2005) to studies with a wider focus on scientific or academic language as such (see, e.g., Halliday and Martin 1993, Ventola 1996, Biber 2006, Hyland 2007). However, an issue that has received little attention so far is the diachronic evolution of scientific registers (see, e.g., Halliday 1988, Banks 2008). In the scientific domain, to pursue new knowledge and technological innovation, the boundaries of established scientific disciplines are transcended and new, interdisciplinary research fields emerge (e.g., in more recent time, bioinformatics, mechatronics or biomechanics). Linguistically, we encounter here a situation of *register contact*, where a newly emerging scientific field draws on the linguistic conventions of two or more established scientific disciplines and possibly develops a new register.

The overarching goal of our research is to develop a model of register formation in specialized scientific domains, tracing the major motifs governing the

[1] Project 'Registers in contact', funded by Deutsche Forschungsgemeinschaft (DFG) under grant TE-198/2.

development of a new scientific discipline – *diversification* and s*tandardization* – in linguistic terms. This involves addressing the following questions:

(1) What are the linguistic features involved in the process of register formation in the scientific domain? Registers are linguistically manifested by particular distributions of lexico-grammatical patterns that are relatively stable in time. The diagnostic for a new register developing would thus be the observation of redistributions of such patterns. Thus, in analysis, a contrastive approach comparing different scientific registers over time is required.

(2) Which contextual settings do the linguistic features involved in register formation realize? Register variation is situation-dependent variation. The canonical view is that situations can be characterized by the parameters of field, tenor and mode of discourse (cf. Halliday 1985, Quirk et al. 1985, Halliday and Hasan 1989). Field denotes the social action participants are engaged in within a situation (e.g., processes and participants), tenor concerns the relationship between participants (e.g., roles and attitudes of participants), and mode is about the symbolic organization of information (information flow, foregrounding and backgrounding of information etc). These situational parameters are encoded by particular linguistic subsystems (field: lexis/colligation, tenor: mood and modality, mode: theme-rheme, given-new). It is thus part of the analytical task to interpret the observed linguistic features (and their distributions) in terms of their contextual settings.

The main goal of the paper is to present the particular corpus design (Section 2) and the principal methodology we adopt to pursue our research goals (Section 3). In order to illustrate our approach, we provide selected examples of analysis carried out using the corpus (Section 4). We conclude with a summary and envoi (Section 4).

2. SciTex: corpus design and processing

To investigate register formation in the scientific domain, we focus on the situation of interdisciplinary contact between computer science and selected other disciplines. We have built a corpus comprising texts from computer science (A-subcorpus), four interdisciplinary fields (B-subcorpus: computational linguistics, bioinformatics, computer-aided design, microelectronics), and their respective disciplines of origin (C-subcorpus: linguistics, biology, mechanical engineering and electrical engineering) (cf. Teich and Holtz 2009, Teich and Fankhauser 2010) – the SciTex corpus (see Figure 1).

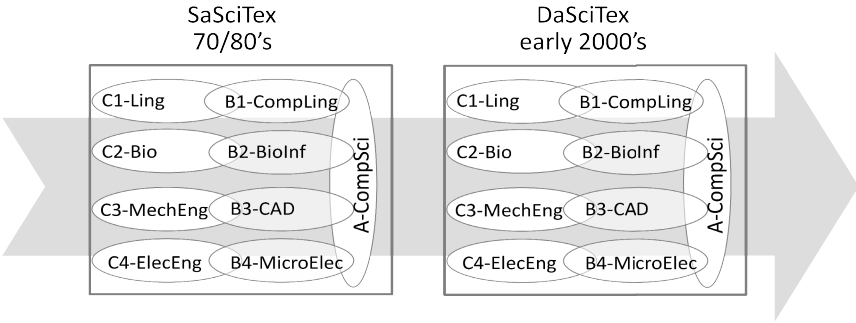

Figure 1: Disciplines represented in SciTex

Diachronically, SciTex covers two time periods, the 1970's/early 1980's (SaSciTex) and the 2000's (DaSciTex).[2] The sources for the corpus are full journal articles; for each discipline at least two different journals were selected. The sources were collected in the form of pdf files and converted to plain text format. Altogether, SciTex contains around 35 million tokens (i.e., approx. 17.5 million per time slice). A smaller cross-sectional subcorpus of around two million tokens, i.e., one million per time slice, was created that was cleaned from erroneous data produced by the OCR conversion. Additionally, formulas, mostly from computer science, and examples, as in linguistics, were annotated with particular tags to be able to exclude them from linguistic searches. This extensive procedure was important in order to obtain high quality text data at least for a portion of the corpus that can be employed for detailed analyses that may require further (manual) annotation.
Furthermore, a dedicated processing pipeline (cf. Kermes 2011) was implemented for (1) conversion of the corpus from text files to xml files while maintaining information about document structure (e.g., paragraphs, sections etc), (2) tokenization, lemmatization and part-of-speech tagging of the corpus files using the TreeTagger (Schmid 1994), (3) transformation of the corpus files into a verticalized text format for segmentation, and (4) encoding of the corpus for query by the Corpus Query Processor (CQP; Evert 2005).

3. Methodology: Complementary comparisons

The chosen corpus design allows two kinds of perspectives that are necessary for our purposes: the temporal and the disciplinary. From the temporal perspective, we can carry out both *synchronic* (within a time slice) and *dia-*

[2] Compare the Brown corpus family with the Brown (AmE) and LOB (BrE) corpora from the early 1960's and the Frown (AmE) and FLOB corpora (BrE) from the early 1990's for a similar design (cf. Kucera and Francis 1967, Hundt, Sand and Siemund 1999, Hundt, Sand and Skandera 1999).

chronic (across time slices) comparisons. From the disciplinary perspective, we can compare the different disciplines in terms of *register*. To address our overarching research question, i.e., register formation in the scientific domain, we obviously need to combine the two perspectives. Doing so, we may want to consider just one triple of A-B-C corpora, zooming in on one particular interdisciplinary field (e.g., computational linguistics) compared to its discipline of origin (e.g., linguistics) and computer science in order to detect a trend in that particular interdisciplinary field; or we may look at the interdisciplinary fields as a whole (all B corpora) in the search of a general trend. Furthermore, we may want to compare SciTex as a whole to a registerially mixed corpus, such as the BNC or the Brown family (cf. Teich & Fankhauser 2010). This may be of interest for comparing diachronic trends in the language as whole (cf. e.g., Mair 2006) to the development of scientific language.

In order to detect diachronic trends, we need to determine features that are potentially relevant for the formation of new registers and that ultimately bring about significant feature redistributions. The theoretical framework of Systemic Functional Linguistics provides a map of lexico-grammatical domains to look into for such features. Lexico-grammatical features typically associated with the contextual variables of field, tenor and mode are:

- field: experiential lexis, collocation/colligation, predicate-argument structure
- tenor: mood, modality, expressions of stance
- mode: theme-rheme, given-new

On the basis of the annotated corpus as described in Section 2, we can then proceed to extract instances of these features. The extraction tool we employ is CQP, which allows us to detect feature instances by means of regular expressions, offering several functionalities for extraction (e.g., context expansion) and sorting purposes (e.g., counting, grouping of results). This flexibility is very useful when working with linguistic features at various cut-off points of the grammar-lexis cline. The obtained feature frequencies are then evaluated in terms of their discriminatory effects across registers and time slices, using univariate methods (e.g., chi-square test) on single features as well as multivariate methods (e.g., principal component analysis, correspondence analysis etc) on sets of features.

In the following section, we show two examples of analysis using two features associated with field and tenor, respectively, and employing univariate evaluation techniques.

4 Sample analyses

4.1 Discourse field: Lexis (most frequent words/keywords)

In the development of a scientific discipline, the creation of a distinctive vocabulary, esp. terminology, is a key issue. To get a first impression, we extract the most frequent nouns from the subcorpora of SciTex for both time slices. The most frequent nouns provide a first indication of the topics in a discipline. Table 1 illustrates two triples (A-B1-C1, A-B2-C2) with the five most frequent nouns.

discipline	70/80's	early 2000's
A-CompSci	set	algorithm
	time	time
	function	problem
	proof	graph
	algorithm	set
B1-CompLing	word	word
	sentence	translation
	rule	sentence
	structure	system
	system	example
C1-Ling	rule	language
	sentence	verb
	form	case
	case	example
	verb	word
B2-BioInf	system	gene
	computer	protein
	time	method
	program	sequence
	value	model
C2-Bio	DNA	gene
	fragment	sequence
	site	protein
	gene	cell
	plasmid	DNA

Table 1: Five most frequent nouns for two triples in both time slices

It can be seen from the table that diachronically, the most frequent nouns have changed to different degrees for each discipline. For example, in the triple A-B1-C1, all three disciplines have slightly changed their five most frequent nouns. However, when we look at the interdisciplinary field in this triple, computational linguistics (B1), there is apparently no diachronic change regarding its relation to linguistics (C1) and computer science (A): both in the 1970's/80's and in the early 2000's it leans more towards linguistics than to computer science. Looking at the triple A-B2-C2, a different development is indicated. In the 1970's/80's the interdisciplinary field of bioinformatics (B2) leans more towards computer science (A) in the nouns

used most frequently (e.g., *computer*, *time*, *program*), while in the early 2000's, there is a larger overlap with biology (C2) (e.g., *gene*, *sequence*, *protein*).

To further explore these tendencies, we calculate the keyness of the most frequent nouns for each subcorpus, again comparing triples of subcopora. Keyness is calculated by means of the log likelihood statistics. The higher the log likelihood value, the more significant is the difference between corpora. A log likelihood of 3.8 or higher indicates a significant difference between two corpora (p-value < 0.05). Positive and negative values indicate which corpus makes more or less use of a given word.

discipline		**nouns**	**keyness in comparison to**	
			C1-Ling	A-CompSci
B1-CompLing	70/80's	**word**	+ 475.50	+ 1767.27
		sentence	- 3.37	+ 2834.26
		rule	- 2328.87	+ 850.19
		structure	+ 0.90	+ 993.88
		system	+ 639.33	+ 242.12
	early 2000's	**word**	+ 1811.54	+ 8416.12
		translation	+ 5602.33	+ 7785.86
		sentence	+ 1265.74	+ 7921.52
		system	+ 2147.56	+ 2576.71
		example	+ 83.20	+ 2230.67
			C2-Bio	A-CompSci
B2-BioInf	70/80's	system	+ 1711.89	+ 1581.33
		computer	+ 3176.64	+ 3703.04
		time	+ 1138.78	+ 51.52
		program	+ 2931.57	+ 585.25
		value	+ 1106.19	+ 960.09
	early 2000's	gene	- 48.20	+ 15216.64
		protein	- 192.48	+ 8429.29
		method	+ 3032.11	+ 5185.33
		sequence	- 305.71	+ 2303.13
		value	+ 2486.16	+ 1623.78

Table 2: Log likelihood for B1-CompLing and B2-BioInf

Table 2 shows the log likelihood values for the comparison of the two triples A-B1-C1 and A-B2-C2 for both time slices.[3] From these values we can observe that in the 1970's/80's computational linguistics (B1) was more similar to linguistics (C1) than to computer science (A) – except for the word *rule*, all log likelihood values are smaller for B1 vs. C1 than for B1 vs. A. In the early 2000's the differences to both C1 and A become greater. Regarding bioinformatics (B2), the differences to computer science (A) increase over

[3] Nouns that are common to two time slices are highlighted by bold face; negative values indicate a less frequent use relative to an interdisciplinary field

time, too (larger log likelihood values for B2 vs. A), while the difference to biology (C2) decreases (lower log likelihood values for B2 vs. C2 for all words in B2 except *value*).

4.2 Discourse tenor: Modal verbs

One of the linguistic features relevant for discourse tenor is modality. Here, we discuss an analysis of modal verbs[4]. The overall trend we detect from the quantitative results is that in the early 2000's, the amount of modal verbs used is rather stable across disciplines with 500-1 000 occurrences per million words in each discipline. This is in contrast to the 1970's/80's which exhibit a relatively high variability across disciplines, some using quite a lot of modal verbs (e.g., in linguistics: around 3 000 modal verbs per million words) and others rather few (e.g., in computer-aided design: under 500 modal verbs per million words). Overall, we obtain similar results to Mair (2006) who reports a decrease of the modals *shall*, *ought to*, *need to* as well as *must* and *may* in the Brown corpus family (see Mair 2006: 327). However, in contrast to the relative stability of the use of *can* as reported by Mair (2006) in English generally, we observe a relative increase of *can* (approx. 10 to 20 %) in our corpus.

To see how the interdisciplinary fields have developed in the given time period, we have investigated all A-B-C triples across the two time slices. For this purpose, we apply the following meaning groups as used by Biber et al. (1999: 485):

- permission/possibility/ability: *can*, *cannot*, *could*, *may*, *might*;
- obligation/necessity: *must*, *have to*, *need to*, *ought to*, *should*;
- volition/prediction: *will*, *would*, *shall*.

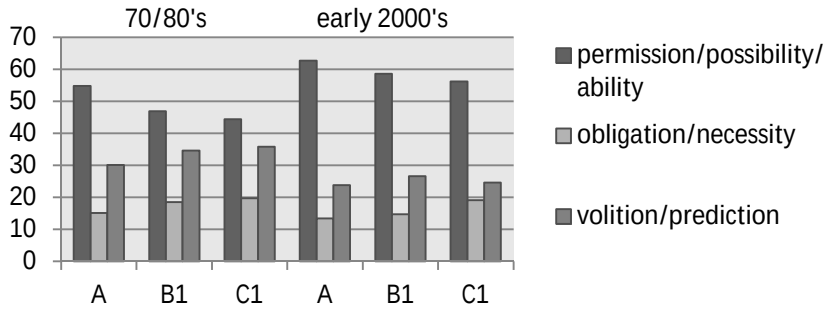

Figure 2: Distribution of modal meanings across the A-B1-C1 triples

[4] For other features in the tenor parameter, e.g., evaluative patterns, modal adverbs, see Teich and Degaetano 2011 and Degaetano 2011.

Figure 2 shows the A-B1-C1 triple in both time slices in percentages (100% = all modal verbs used). In the 1970/80's computational linguistics (B1) seems to be more similar to linguistics (C1), but different from computer science (A), comparing the percentages of the modal meanings (less use of obligation/necessity and volition/prediction in A compared to B1 and C1). In the early 2000's, the picture changes: computational linguistics (B1) seems more similar to computer science (A), while differing from linguistics (C1) in the use of the obligation/necessity group. The diachronic tendency of computational linguistics (B1) being first similar to linguistics (C1) and later on moving towards computer science (A) is confirmed by calculating the p-values using the chi-square test (see Table 3): B1 shows higher significant differences to A than to C1 in the 1970/80's, but lower significant differences to A than to C1 in the early 2000's.

Time slice	**discipline**			**p-value**
1970/80's	B1-CompLing	-	A-CompSci	< 2.2e-16
	B1-CompLing	-	C1-Ling	0.0004012
early 2000's	B1-CompLing	-	A-CompSci	8.37e-14
	B1-CompLing	-	C1-Ling	< 2.2e-16

Table 3: p-values for diachronic comparison of the A-B1-C1 triple

Diachronic changes have also been observed for the other interdisciplinary fields: bioinformatics (B2) moved from being similar to computer science (A) to differing from both biology (C2) and computer science (A); computer-aided design (B3) differs from both mechanical engineering (C3) and computer science (A), thus creating its own variation; microelectronics (B4), instead, remains similar to its discipline of origin (C4: electrical engineering) and differs from computer science (A) in both time slices.

5. Summary and Conclusions

In this paper, we have introduced a project on the diachronic development of highly specialized scientific registers having emerged by register contact. In our investigation, we focus on the situation of interdisciplinary contact of selected disciplines with computer science. The prerequisite for our research is an appropriate corpus. We have introduced the SciTex corpus which is compiled from research articles from nine scientific disciplines (Section 2). The SciTex corpus enables us to investigate register contact from both the synchronic and the diachronic angle. The framework of Systemic Functional Linguistics and register/genre theory provide the linguistic and contextual categories relevant for the analysis of register variation (see Section 3). We have then shown two examples of analysis of the corpus, one using a field-related feature, the other using a tenor-related feature (Section 4). The analy-

ses have focused on diachronic trends in the four interdisciplinary fields contained in SciTex.

So far, we obtain indications of both of the principal motifs of scientific evolution, diversification and standardization. However, the picture is not uniform across the four interdisciplinary fields investigated: some change from being more similar to their discipline of origin to being more similar to computer science (e.g., modal meanings in computational linguistics), others change from being more similar to computer science to greater similarity to their discipline of origin (e.g., lexis/keywords in bioinformatics) and some seem to create their own patterns of variation (e.g., modal meanings in computer-aided design). Moreover, the tendencies may differ for different contextual parameters, e.g., differences in field, but similarities in tenor. Obviously, we need to study more features in order to cover the full spectrum of potential variation. Also, with a larger feature set we will be able to use other, more powerful methods of feature evaluation, such as automatic clustering or classification, which will allow us a more comprehensive and differentiated interpretation. In the analysis of vocabulary, we will explore more advanced methods such as topic models (cf. Blei to appear) which promise to get a tighter grip on diachronic topic shifts.

References

Banks, David (2008): The Development of Scientific Writing, Linguistic features and historical context. London: Equinox.

Biber, Douglas (2006): University language: A corpus-based study of spoken and written registers. Amsterdam: Benjamins.

Biber, Douglas (1995): Dimensions of Register Variation. A Cross-linguistic Comparison. Cambridge: Cambridge University Press.

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

Biber, Douglas/Johansson, Stig/Leech, Geoffrey (1999): Longman Grammar of Spoken and Written English. London: Longman.

Blei, David (to appear): Introduction to probabilistic topic models. Communications of the ACM.

Degaetano, Stefania (2011): Evaluative options and their choice - modal adjuncts vs. evaluative patterns in academic writing. Paper presented at the International Evaluation Conference (IntEval). Madrid: 6-8 Oct. 2011.

Evert, Stefan (2005): The CQP Query Language Tutorial. IMS, Universität Stuttgart.

Halliday, Michael A.K. (2004): Introduction to Functional Grammar. 3rd edition (with Christian M.I.M. Matthiessen). London: Edward Arnold.

Halliday, Michael A.K. (1988): On the language of physical science. In: Ghadessy, Mohsen (ed.). Registers of Written English: Situational Factors and Linguistic Features. London: Pinter, 162–177.

Halliday, Michael A.K. (1985). Spoken and written language. Victoria: Deakin University Press.

Halliday, Michael A.K./Martin, James R. (1993): Writing science: Literary and discursive power. London and Washington D.C: The Falmer Press.

Halliday, Michael A.K./Hasan, Ruqaiya (1989): Language, Context and Text: a social semiotic perspective. London: Oxford University Press.

Hundt, Marianne/Sand, Andrea/Siemund, Rainer (1999): Manual of Information to accompany The Freiburg – LOB Corpus of British English ('FLOB'). Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg. http://khnt.aksis.uib.no/icame/manuals/flob/INDEX.HTM.

Hundt, Marianne/Sand, Andrea/Skandera, Paul (1999): Manual of Information to accompany The Freiburg – Brown Corpus of American English ('Frown'). Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg. http://khnt.aksis.uib.no/icame/manuals/frown/INDEX.HTM.

Hyland, Ken (2007): Disciplinary Discourses. Social Interactions in Academic Writing. Ann Arbor: The University of Michigan Press.

Mair, Christian (2006): Twentieth-Century English: History, Variation and Standardization. Cambridge: Cambridge University Press.

Martin, James R. (1992): English Text: System and Structure. Amsterdam: John Benjamins.

Kermes, Hannah (2011): Automatic corpus creation. Manual. Institute of Applied Linguistics, Translation and Interpreting. Saarbrücken: Universität des Saarlandes.

O'Halloran, Kay L. (2005): Mathematical Discourse: Language, Symbolism and Visual Images. London: Continuum.

Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1985). A Comprehensive Grammar of the English Language. London: Longman.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 44-49.

Teich, Elke/Degaetano, Stefania (2011): The lexico-grammar of stance: an exploratory analysis of scientific texts. In: Dipper, Stefanie/Zinsmeister Heike (eds.). Beyond semantics: Corpus-based investigations of pragmatic and discourse phenomena. Bochumer Linguistische Arbeitsberichte 3, 57-66.

Teich, Elke/Fankhauser, Peter (2010): Exploring a corpus of scientific texts using data mining. In: Gries, Stefan Th./Wulff, Stefanie/Davies, Mark (eds.). Corpus-linguistic applications. Current studies, new directions. Amsterdam and New York: Rodopi, 233-247.

Teich, Elke/Holtz Mônica (2009): Scientific registers in contact: A methodology and some findings. In: International Journal of Corpus Linguistics 14(4): 524-548.

Ventola, Eija (1996): Packing and unpacking of information in academic texts. In: Ventola, Eija/Mauranen Anna (eds.): Academic Writing. Intercultural and Textual Issues. Amsterdam/Philadelphia: John Benjamins, 153-194.