Register profiling of scientific texts:

Experiences in linguistic description and corpus-based methods

Sabine Bartsch
Technische Universität Darmstadt
Department of Linguistics & Literary Studies

Elke Teich
Universität des Saarlandes
Englische Sprach- und Übersetzungswissenschaft

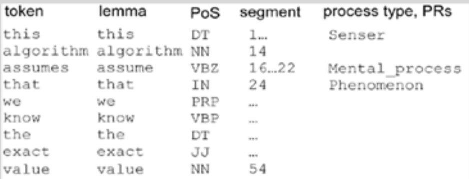
Collaborators: Stefanie Degaetano, Richard Eckart de Castilho, Peter Fankhauser, Mônica Holtz

Prerequisites for register studies

Registers are characterized by typical clusters of features which have a greater-than-random tendency to occur (Halliday & Martin 1993, p. 54).

Register analysis is inherently quantitative; because frequency of occurrence is relative, register analysis must be comparative. Quantitative studies are typically corpus-based; appropriate corpus designs are a necessary prerequisite for comparative studies.

The corpus which has served as the basis for a variety of studies of English scientific registers over the last few years is the Darmstadt Scientific Text Corpus (DaSciTex), a multilayer annotated corpus of scientific research articles.



token	lemma	PoS	segment	process type, PRs
this	this	DT	1…	Senser
algorithm	algorithm	NN	14	
assumes	assume	VBZ	16…22	Mental_process
that	that	IN	24	Phenomenon
we	we	PRP	…	
know	know	VBP	…	
the	the	DT	…	
exact	exact	JJ	…	
value	value	NN	54	

Figure 1: Annotation example

The Corpus

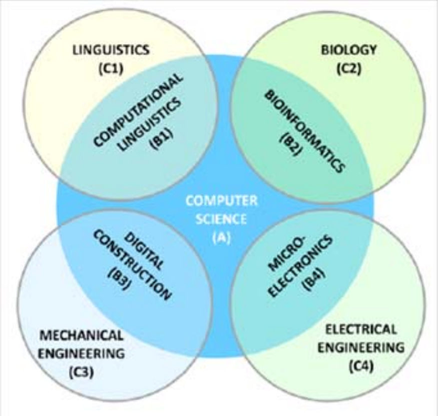

LINGUISTICS (C1)
BIOLOGY (C2)
COMPUTATIONAL LINGUISTICS (B1)
BIOINFORMATICS (B2)
COMPUTER SCIENCE (A)
DIGITAL CONSTRUCTION (B3)
MICRO-ELECTRONICS (B4)
MECHANICAL ENGINEERING (C3)
ELECTRICAL ENGINEERING (C4)

Figure 2: Corpus design of DaSciTex

Corpus profile:

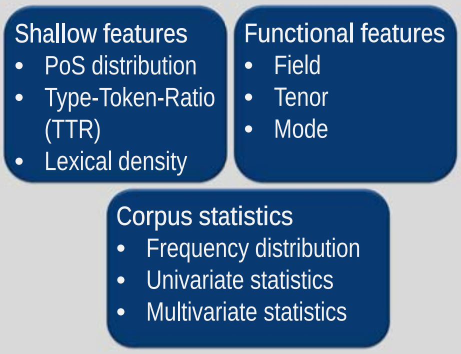
- Sources: full journal articles from 2007
- Sampling: 3 – 4 journals per discipline
- Size: ca. 17 mio. tokens; 1 mio. tokens cleaned and hand annotated.

Features of register variation

The studies presented are concerned with

(a) the distinctive linguistic properties of scientific writing (compared to less specialized texts) (cf. Teich & Fankhauser (2010)) and

(b) the distinctive linguistic properties of individual disciplines (cf. Teich & Holtz (2009), Degaetano & Teich (2011), Bartsch et al. (submitted)).



Shallow features
- PoS distribution
- Type-Token-Ratio (TTR)
- Lexical density

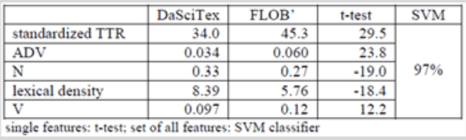
Functional features
- Field
- Tenor
- Mode

Corpus statistics
- Frequency distribution
- Univariate statistics
- Multivariate statistics

Figure 3: Features of register variation

Distinctive properties of scientific writing

Sets of features that are characteristic of science writing are identified and statistically evaluated against standard corpora.



	DaSciTex	FLOB'	t-test	SVM
standardized TTR	34.0	45.3	29.5	97%
ADV	0.034	0.060	23.8	
N	0.33	0.27	-19.0	
lexical density	8.39	5.76	-18.4	
V	0.097	0.12	12.2	

single features: t-test; set of all features: SVM classifier

Table 1: Comparison FLOB vs. DaSciTex

Distinctive features ranked:

- Type-Token Ratio (TTR) → relatively low (91 % distinctiveness)
- Noun ratio → relatively high
- Adverb ratio → relatively low
- Lexical density → relatively high

Methods:

- Univariate statistics: t-test
- Multivariate statistics: Classification (linear Support Vector Machines (SVM))

Distinctive properties of individual disciplines

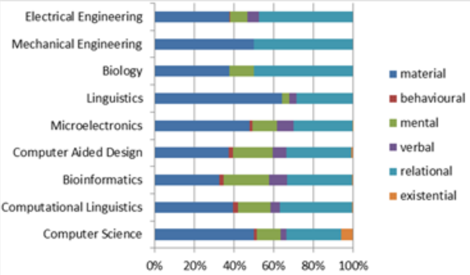


Diagram 1: Process type distribution
'algorithm' + VERB

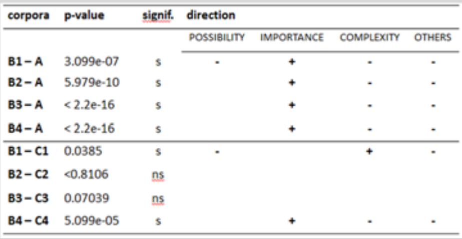
Examples:

(1) *… a centralized polynomial algorithm that works in the spirit of LIS.* [COMPSCI: mat.]

(2) *…our algorithm assumes that no ambiguity arises.* [COMPLING: mental]

These results suggest that the B subcorpora differ significantly from the A corpus; while, if we compare B to C, results are signifcant for B4, but not for any of the others (B1-3).

Distinctive properties of individual disciplines



corpora	p-value	signif.	direction			
			POSSIBILITY	IMPORTANCE	COMPLEXITY	OTHERS
B1 – A	3.099e-07	s	-	+	-	-
B2 – A	5.979e-10	s		+	-	-
B3 – A	< 2.2e-16	s		+	-	-
B4 – A	< 2.2e-16	s		+	-	-
B1 – C1	0.0385	s	-		+	-
B2 – C2	<0.8106	ns				
B3 – C3	0.07039	ns				
B4 – C4	5.099e-05	s		+	-	-

Table 2: Distribution of stance expressions

Examples from Computer Science:

(1) *It is often easiest to pick them at random*.

(2) *It is impossible to eliminate packets*.

(3) *It is also important to consider losses*.

Results:

- B subcorpora make more use of the IMPORTANCE-group than computer science (A)
- bioinformatics (B2) and DigiConst (B3) similar to their pure disciplines
- distinctive difference microelectronics (B4) (differs in the same way from A and C4)
- less pronounced difference of compling (B1)

Deutsche Forschungsgemeinschaft DFG
Linguistische Profile interdisziplinärer Register (2006-2009) & Register im Kontakt (2011-2014)