



Characterizing genre/register: The case of scientific text

Elke Teich

Institut für Sprach- und Literaturwissenschaft
Technische Universität Darmstadt

Peter Fankhauser

Forschungszentrum L3S
Leibniz Universität Hannover
Germany

Background



- Project **Linguistic profiles of interdisciplinary registers**
(*Linguistische Profile interdisziplinärer Register*, DFG)
- Context
 - **interdisciplinary** research areas – linguistic reflexes?
 - registers at the boundaries of **computer science** and some other scientific discipline (e.g., bioinformatics, computational linguistics); language: English
- Goals
 - insights on **emerging registers**: describe how registers develop in contact situations
 - **scientific writing**: test observations about genre/register from *Systemic Functional Linguistics* (SFL; Halliday & Martin 1993)
 - develop **methodology for register comparison** using state-of-the-art corpus methods and techniques

Overview of this talk

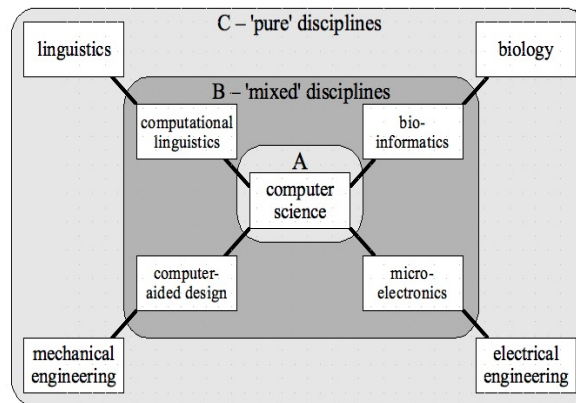


- (1) Darmstadt Scientific Text Corpus: Corpus design
- (2) Corpus processing
- (3) Analysis: Features of scientific texts
 - How coherent is the corpus?
 - How distinct are the subcorpora?
- (4) Summary and envoi

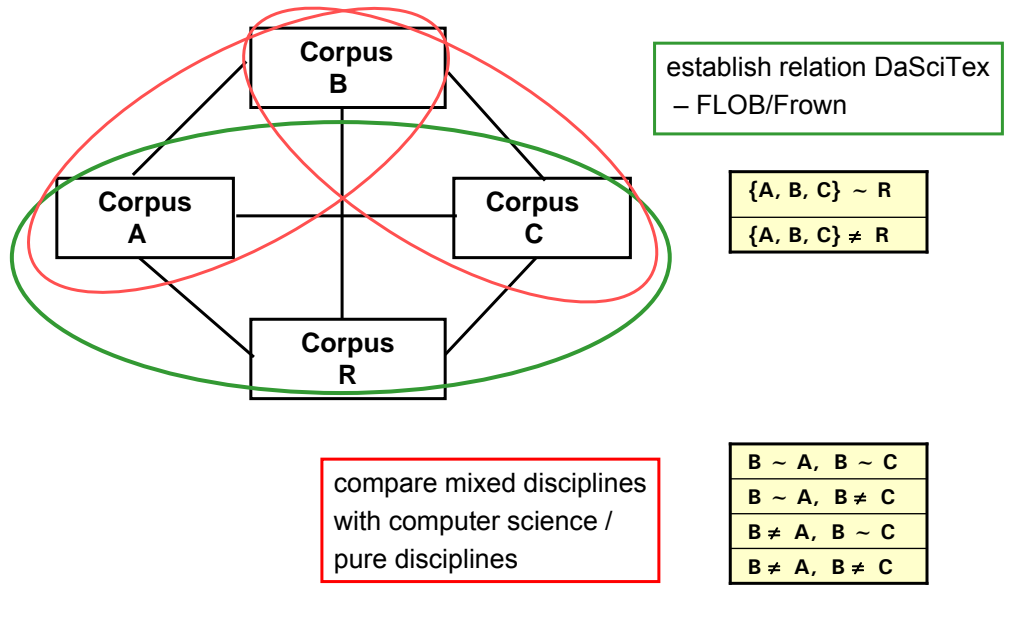
(1) Corpus design




- Registers at the boundaries to computer science: commonalities and differences
- **Darmstadt Scientific Text Corpus (DaSciTex):**
 - Corpus **A: computer science**
 - Corpus **B: “mixed” disciplines** - computational linguistics (B1), bioinformatics (B2), computer aided design (B3), microelectronics (B4)
 - Corpus **C: “pure“ disciplines** - linguistics (C1), biology (C2), mechanical engineering (C3), electronics (C4)



- Sources: full journal articles
- Time: - 2007
- Sampling: 3-4 journals per discipline
- Size: 100.000 (min) – 2 Mio (max) words per subcorpus (discipline); \sum ca. 20 Mio
- Reference corpora (R):
 - Synchronic: FLOB/Frown (RS2)
 - Diachronic: LOB/Brown (RD1), Helsinki corpus (RD2)



(2) Corpus processing



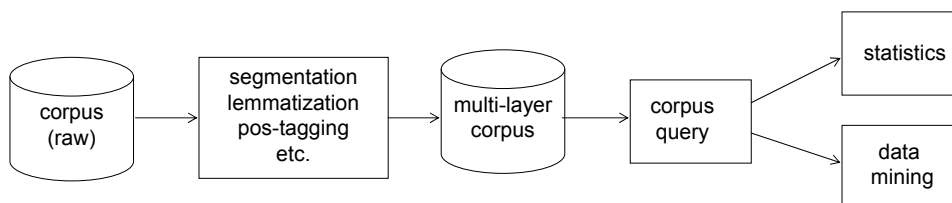
TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Pre-processing: format transformation; encoding
- Linguistic processing
 - segmentation: sentence splitting, tokenization
 - annotation: part-of-speech tagging (Tree tagger; Schmid 1994)
 - processing pipelines: AnnoLab (Eckart 2006, Eckart & Teich 2007)
- Data analysis: data/text mining (Weka; Witten & Eibe 2005)

(1) pre-processing



(2) linguistic processing and data analysis



(3) Analysis: features of scientific text

- How coherent is the corpus?
- How distinct are the subcorpora?
 - all subcorpora
 - triple of corpora
- Method: data mining
- Tool: Weka
- Steps:
 - extract features (= corpus query plus vectorization)
 - evaluate features (select, classify/cluster)
 - inspect result (classification/clustering accuracy, confusion matrix, texts)

Question 1: How coherent is the corpus?

- Scientific texts are generically characterized by
 - abstractness → high ratio of nouns (low ratio of verbs)
 - technicality → low type-token ratio (TTR)
 - information density → high lexical density (LD)
- Comparison with a generically/registerially mixed corpus (here: FLOB)
- Hypotheses DaSciTex vs. FLOB: more nouns, lower TTR, higher LD
- Approach
 - Extract features (parts-of-speech, TTR, LD)
 - Evaluate features (select by information gain; classification/clustering)
 - Inspect results

Question 2: How distinct are the subcorpora?

- **2.a** Experiential metafunction: field (lexis)
 - distinctive lexical words
- Hypothesis: Subcorpora can be distinguished lexically
- Approach
 - Extract features (nouns, verbs)
 - Evaluate features (dito)
 - Inspect results

Question 2: How distinct are the subcorpora?

- **2.b** Interpersonal metafunction:
How do the authors construe themselves?
→ usage of 1st person pronouns
triple: A=CompSci; B1=CompLing; C1=Ling
- Approach
 - Extract features (1st person pl bigrams: we + lexical verb)
 - Evaluate features
 - Inspect results
- Examples:
We have proved ...
We evaluate...
We argue...

Answer 1: How coherent is the corpus?

- Hypotheses DaSciTex vs. FLOB: more nouns, lower TTR, higher LD
- Most discriminative features are (ranked)
 - 1 (low) TTR (91% of texts are correctly classified just on this basis!)
 - 2 (high) noun ratio
 - 3 low adverb ratio (indirectly measures verb ratio)
 - 4 high LD
- Misclassification arises between FLOB GOVERNMENT and SCIENTIFIC

FLOB vs. DaSciTex

classification: 89.7%

clustering: 83.8%

C\P	DaSciTex	FLOB
DaSciTex	163	23
FLOB	38	367

C\P	DaSciTex	FLOB
DaSciTex	179	7
FLOB	89	316

FLOB – H (Government) – J (Scientific) vs. DaSciTex

classification: 96.5%

clustering: 92.1%

C\P	DaSciTex	FLOB
DaSciTex	178	8
FLOB	9	288

C\P	DaSciTex	FLOB
DaSciTex	179	7
FLOB	31	266

Answer 2.a: How distinct are the subcorpora? (experiential: lexis)

- Classification by nouns (no. of nouns: 7204/500)
- Accuracy: 96%
- Confusion Matrix

A vs B: 15

B vs C: 27

Eng: 31

Other: 3

Sum: 76 of 1834

C/P	A	B1	B2	B3	B4	C1	C2	C3	C4
A	217	0	2	2	1	0	0	0	5
B1	3	76	0	0	0	10	0	0	0
B2	3	1	275	0	0	0	6	0	0
B3	3	0	0	215	0	0	0	2	4
B4	1	0	0	1	204	0	0	0	0
C1	0	4	0	0	0	95	0	0	1
C2	0	0	5	0	0	1	236	0	0
C3	0	0	0	0	0	0	0	246	6
C4	4	1	0	4	0	1	0	5	203

no. of texts: A: 227; B1: 89; B2:284; B3: 224; B4: 206; C1: 100; C2: 242; C3: 252; C4: 218

- Classification by verbs (no. of verbs: 3620/250)
- Accuracy: 87%
- Confusion Matrix

A vs B: 28

B vs C: 59

Eng: 139

Other: 3

Sum: 239 of 1834

C/P	A	B1	B2	B3	B4	C1	C2	C3	C4
A	200	3	1	9	2	0	1	1	10
B1	5	65	5	1	1	11	0	0	1
B2	2	7	266	2	1	1	3	1	2
B3	5	4	0	180	1	0	0	17	17
B4	1	2	0	2	200	0	0	0	1
C1	0	9	0	1	0	90	0	0	0
C2	0	0	6	1	0	2	229	4	0
C3	1	0	1	13	1	1	2	222	11
C4	13	0	1	32	6	0	0	14	152

no. of texts: A: 227; B1: 89; B2:284; B3: 224; B4: 206; C1: 100; C2: 242; C3: 252; C4: 218

- Subcorpora are well distinguished by nouns (surprise?)
- Subcorpora are rather well distinguished by verbs (surprise!)
- Broader areas are well separated:
 - Engineering
 - Humanities (linguistics)
 - Science (biology)
- Misclassifications arising (by frequency):
 - 1 Among CompSci (A) and engineering (C3, C4)
 - 2 Between original and mixed disciplines (B vs. C)
 - 3 Between CompSci and mixed disciplines (A vs. B)
 - interesting?

Answer 2.b: How distinct are the subcopora? (interpersonal: construal of self)

A=CompSci; B1=CompLing; C1=Ling

What we do in:

- CompSci (A): *prove, show, obtain* (“formal”)
- CompLing (B1): *examine, implement, use* (“experimental”)
- Ling (C1): *propose, suggest, argue* (“semiotic”), *feel, see* (“mental”)

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

- Most misclassifications occur for CompLing (B1: 19+47+21+32), very few (3+3) occur between CompSci (A) and Ling (C1)
→ May this be an indicator of CompLing being “in between” CompSci and Ling?

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

- CompLing is more often misclassified as Ling (47) than as CompSci (19), and Ling is more often misclassified as CompLing (32) than CompSci is misclassified as CompLing (21)
→ Could this mean CompLing is more similar to Ling?

(4) Summary and envoi

- Summary
 - scientific registers in contact (focus: disciplines mixing with computer science): linguistic reflexes, emerging registers
 - innovation:
 - combine functional linguistic framework (here: SFL) with data mining
 - extend data mining to work on annotated text
- Future work
 - more linguistic analyses:
 - find interesting register/genre features
 - compare subcorpora for register/genre features
 - data mining for linguistic analysis *and* linguistic analysis for NLP applications (automatic text classification, question-answering etc)

Thanks to the Team

Sabine Bartsch
Richard Eckart
Monica Holtz
Anke Schulz
Lara Schwarz