Information Density and Scientific Literacy in English

Stefania Degaetano-Ortlieb, Hannah Kermes,
Ashraf Khamis, Noam Ordan, Elke Teich

IDeaL
SFB 1102

Aim

Investigate

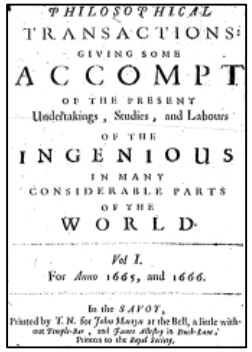
- in general: formation of discourse types/registers
- specific: diachronic development of written scientific English

Motifs

- Specialization
  (greater encoding density)
- Conventionalization
  (greater linguistic uniformity)

Balance in Information Density (ID)

Example: Abstract


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

- basic ling structure:
  complex NPs,
  simple clause structure

- basic linguistic structure:
  complex NPs,
  simple clause structure
- local effects:
  omission of determiners
- global effects:
  high TTR, high lexical density

The Excesses of the Sines of the Refraction of several sorts of Rays above their common Sine of Incidence when the Refractions are made out of divers denser Mediums immediately into one and the same rarer Medium, suppose of Air, are to one another in a given Proportion.

We report the discovery of a novel downstream target of BCR-ABL signalling, PRL-3 (PTP4A3), an oncogenic tyrosine phosphatase. Analysis of CML cancer cell lines and CML patient samples reveals the upregulation of PRL-3.

Assumptions and Hypotheses

- Higher encoding density over time (cf. Halliday,1989: On the language of physical science)
  - scientific texts will exhibit higher ID *relative* to other productions / "general language" over time → *specialization*
  - *within* scientific productions, ID will decrease over time → *conventionalization*
- Correlation between variation in ling encoding and ID
  - Ling features marking specialization and conventionalization serve optimizing ID in scientific writing
    - longer, expanded ling forms → less predictable, more informative
    - shorter, reduced ling forms → more predictable, less informative

Research Design and Methodology

Corpus-based approach: Find out linguistic features indicative of ID

LM approach: Measure ID

- synchronically by comparison of Abstracts vs. Research articles
  Corpus: SciTex, size: ~34M words, 9 academic disciplines, divisions annotation (two time periods 70/80s and 2000s)

- diachronically by comparison of historical stages of text productions (17th century - present)
  Corpus: RSC, size: ~23M words, time periods: 1665-1870

Research Design and Methodology

Corpus-based approach:

- Feature selection, extraction, evaluation
- Classification with SVM

Synchronic analysis

- Abstracts vs. Research articles

Diachronic analysis

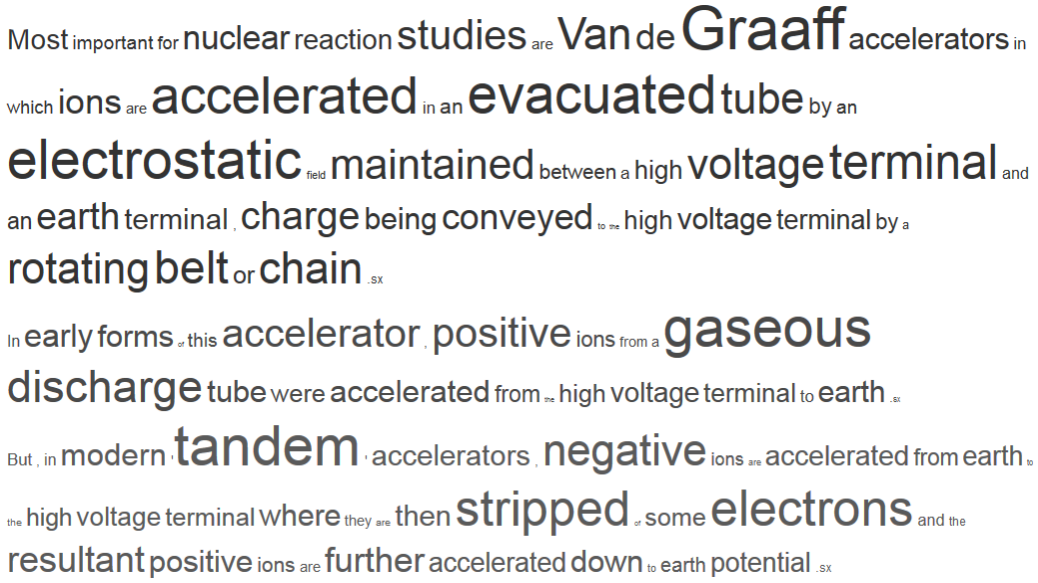
- Historical stages

higher encoding density	lower encoding density
high TTR	low TTR
high TTR on lexical words	low TTR on lexical words
high lexical density	low lexical density
complex NPs	simple NPs
many nominalizations	few nominalizations
few formulaic expressions	many formulaic expressions
simple sentence structure ('x is y')	more varied sentence structures
many nominal compounds	few nominal compounds
few determiners (omissions)	many determiners
few relativizers (omissions)	many relativizers
few complementizers (omissions)	many complementizers

Research Design and Methodology

LM-based approach:

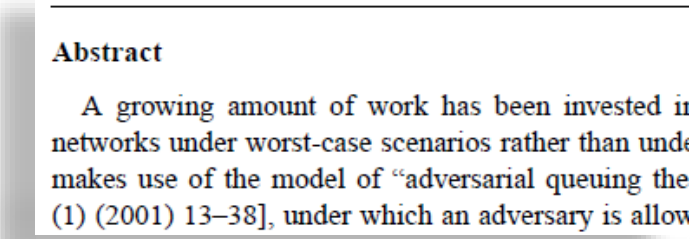
- ID in single texts
  - Cross-entropy based on Genzel and Charniak (2002))
  - Calculate entropy at each token position
  - Explore idea of entropy rate constancy
  - Sliding window of 4 tokens
  - Simple model of memory decay



(by Peter Fankhauser, IDS Mannheim)

Research Design and Methodology

- Relative ID across production types and disciplines
  (by Kullback-Leibler divergence)
  - synchronic: abstracts > research articles; diachronic: t2 > t1

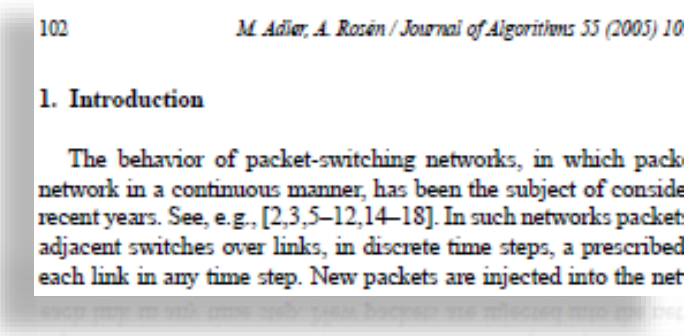


Abstract

A growing amount of work has been invested i
networks under worst-case scenarios rather than unde
makes use of the model of "adversarial queuing the
(1) (2001) 13–38], under which an adversary is allow

Corpus abs

$P(unit \mid \text{Abs})$



102 M. Adler, A. Rosén / Journal of Algorithms 55 (2005) 10

1. Introduction

The behavior of packet-switching networks, in which pack
network in a continuous manner, has been the subject of conside
recent years. See, e.g., [2,3,5–12,14–18]. In such networks packet
adjacent switches over links, in discrete time steps, a prescribed
each link in any time step. New packets are injected into the net

$P(unit \mid \text{RA})$

Corpus RA

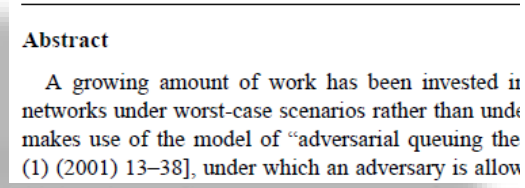
$$H(\text{Abs}) = -\sum_i P(unit_i \mid \text{Abs}) \log_2 P(unit_i \mid \text{Abs})$$

⟶ compare

$$H(\text{RA}) = -\sum_i P(unit_i \mid \text{RA}) \log_2 P(unit_i \mid \text{RA})$$

Research Design and Methodology

- Relative ID across production types and disciplines (by Kullback-Leibler divergence) to identify linguistic patterns that might have changed/ are different
  - Use LM trained on one type on a different type
  - Evaluate largest change in log likelihood



Abstract

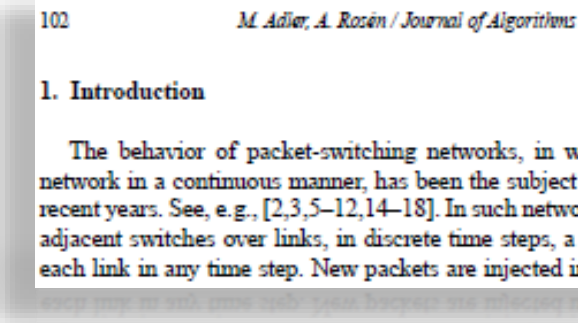
A growing amount of work has been invested ir
networks under worst-case scenarios rather than unde
makes use of the model of "adversarial queuing the
(1) (2001) 13–38], under which an adversary is allow

Corpus Abs

$P(unit \mid \text{Abs})$

apply

Corpus RA



102 M. Adler, A. Rosén / Journal of Algorithms

1. Introduction

The behavior of packet-switching networks, in w
network in a continuous manner, has been the subject
recent years. See, e.g., [2,3,5–12,14–18]. In such netwo
adjacent switches over links, in discrete time steps, a
each link in any time step. New packets are injected i

$$D(\text{RA} \,||\, \text{Abs}) = \sum_{w} P(w \mid \text{RA}) \log_2 \frac{P(w \mid \text{RA})}{P(w \mid \text{Abs})}$$

Analyses

Synchronic analyses on Abstracts vs. Research articles

1. Classification with SVM
2. Entropy rates on single texts
3. LMs

Analysis 1 – SVM Classification Abstracts vs. RAs

Research design

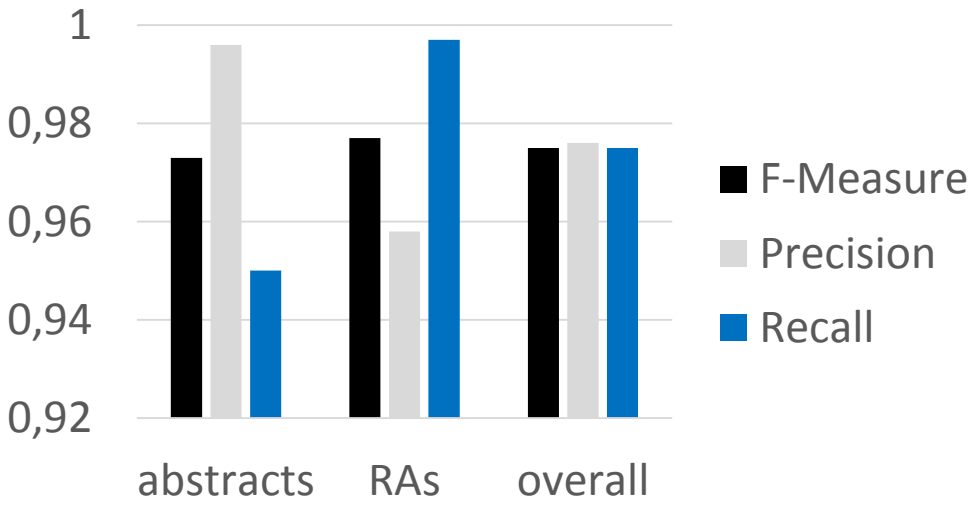
- By possible features involved in reduction and densification
- With Weka (SMO), 10-folds cross-validation, normalized data

type	feature (example)	type	feature (example)
Densification	Sttr Lex. word (nn, adj, adv, vv) / sent.	**Modality**	Modal verbs (*can, would*) modal meanings (obligation)
Complex NPs	term patterns (adj-n, n-of-n)	**Theme**	Experiential (NPs) Interpersonal (*Interestingly*) Textual (*But, Therefore*) Conj. types at sentence beg. (adversative, additive)
Simple clauses	X-be-Y		
Reduction	Personal pronouns (*we*, *our*) Definite/indefinite article (*the*, *a*) Relativizer (*which*, *that*) Affixes (*non-*, *-like*) Hyphen-words (*degree-3-vertex*)	**Expansion**	Conjunctions (*as, since*), Prepositions (*at, by*)

Analysis 1 – SVM Classification Abstracts vs. RAs

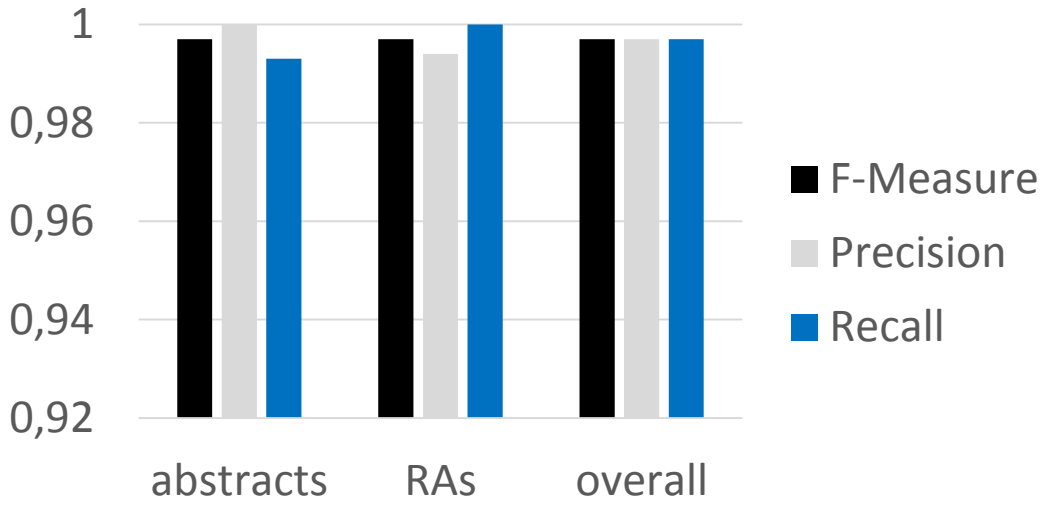
70/80s (SASCITEX)

	abstracts	RAs
abstracts	2488	130
RAs	10	2992



2000s (DASCITEX)

	abstracts	RAs
abstracts	1983	13
RAs	0	2108



Analysis 1 – SVM Classification Abstracts vs. RAs

70/80s (SASCITEX)

Typical for abstracts

type	feature	svm-weight
theme	experiential-sb	-5.03
densification	sttr	-4.66
	lex/s	-3.82
determiners	dt-indef-sb	-4.02
	dt-def-sb	-2.92

Typical for RAs

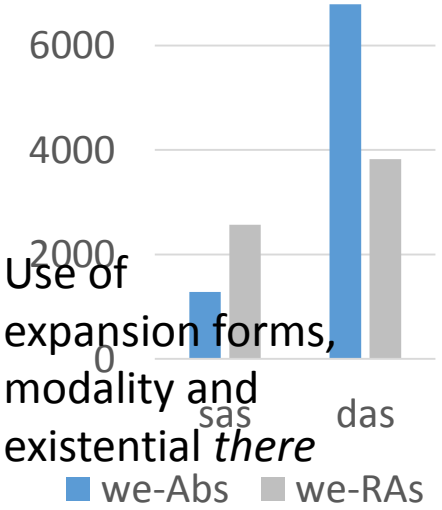
type	feature	svm-weight
expansion	conjunctions	2.73
simple clause	x-be-y	1.43
determiners	dt-def	1.35
modality	obligation	1.28
	volition	1.08

2000s (DASCITEX)

type	feature	svm-weight
densification	sttr	-5.66
	lex/s	-0.88
simple clause	x-be-y	-2.15
determiners	dt-def-sb	-1.06
reduction	pers-pronoun	-0.86

type	feature	svm-weight
expansion	conjuctions	0.97
	noun-relativizer	0.53
modality	modals	0.90
	obligation	0.57
theme	ex-there	0.41

- Densification quite typical
- Simple clause usage
- Reduction to *we* (~60% > than in RAs)

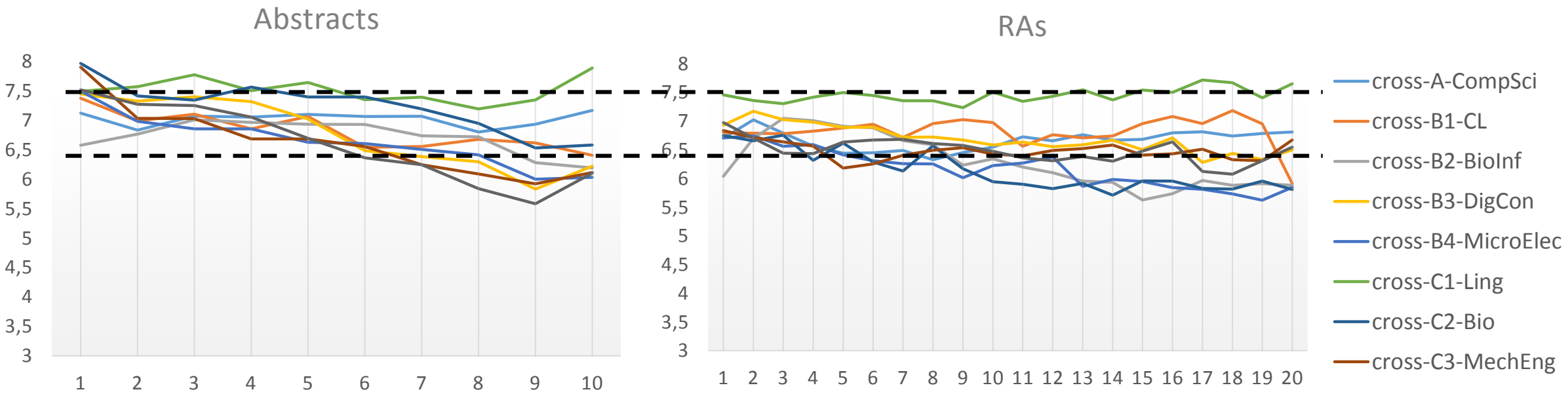


- Use of expansion forms, modality and existential *there*

Analysis 2 – Cross-entropy Abstracts vs RAs

Research design

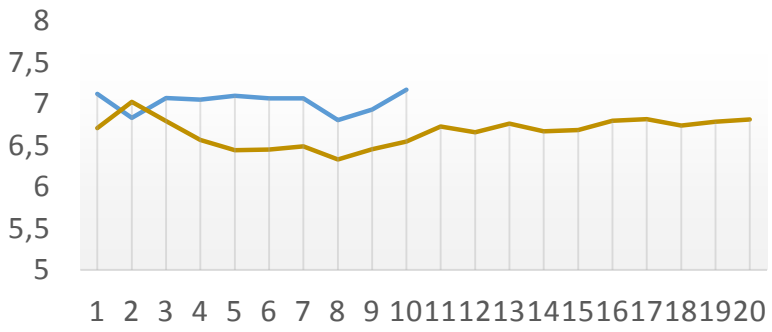
- Based on Genzel and Charniak (2002)
- Procedure designed by **Peter Fankhauser**
  - Cross-entropy calculated on full articles from 2000s (DASCITEX)
  - Cross-entropy rates observed separately for Abstracts and RAs
    - Headlines not considered (low cross-entropy rates)



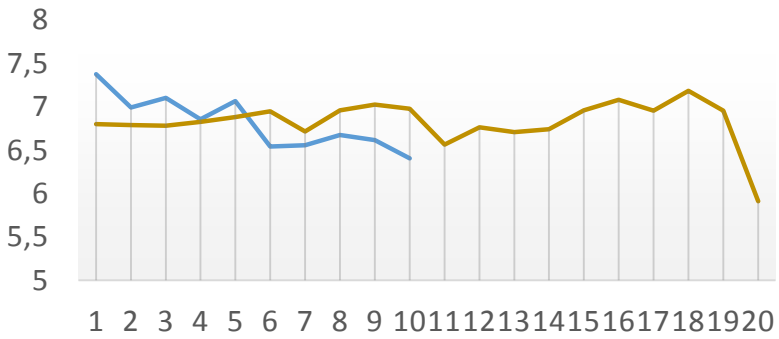
Analysis 2 – Cross-entropy Abstracts vs RAs

- Focus on 5 disciplines from 2000s (DASCITEX)

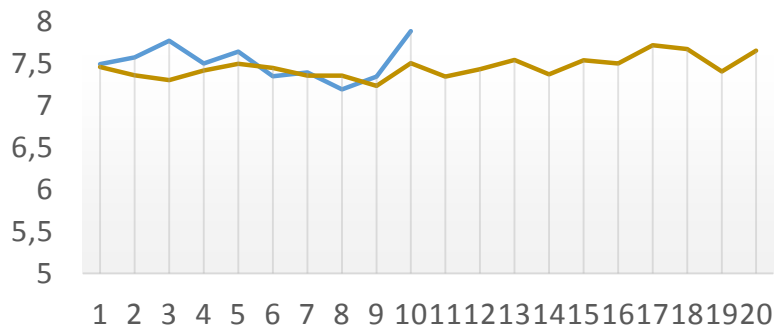
Computer science



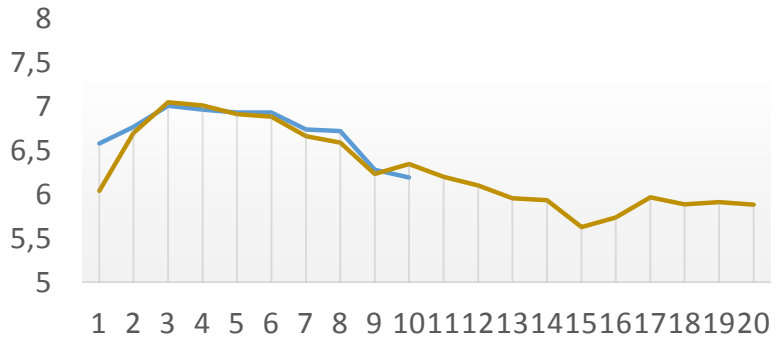
Computational Linguistics



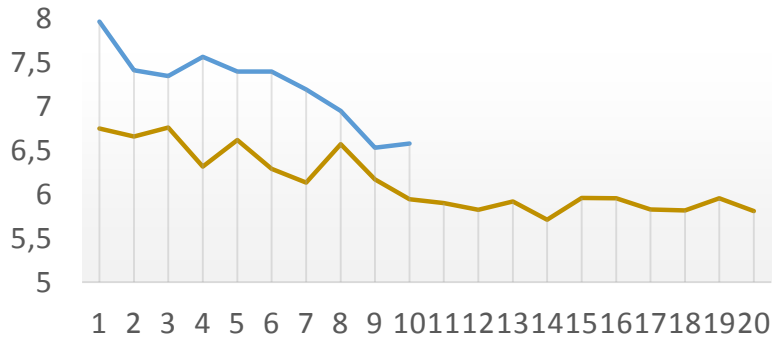
Linguistics



Bioinformatics



Biology



Analysis 2 – Word Cross-entropy Rates Abstracts-CompSci

ABS-sent1	cross_sent									
Younas2006	15.221	Composite	web	services	provide	promising	prospects	for	conducting	cross-organizational
Tan2006a	11.805	TCP	over	bandwidth	asymmetric	networks	such	as	Cable	TV
Liao2006	10.65	Authentication	ensures	that	system's	resources	are	not	obtained	fraudulently
Genest2006	10.257	Message	sequence	charts	(	MSC	)	and	High-level	MSC
Pemantle2005a	10.23	Gosper's	algorithm	is	a	cornerstone	of	automated	summation	of
Buhler2005	10.188	Large-scale	comparison	of	genomic	DNA	is	of	fundamental	importance
Cole2006	10.099	We	study	economic	incentives	for	influencing	selfish	behavior	in
Badger2004	10.089	Many	fundamental	questions	in	evolution	remain	unresolved	despite	the
Koch2005	9.902	Graph-based	specification	formalisms	for	access	control	(	AC	)
Lee2005a	9.757	Asynchronous	cellular	automata	(	ACA	)	are	cellular	automata

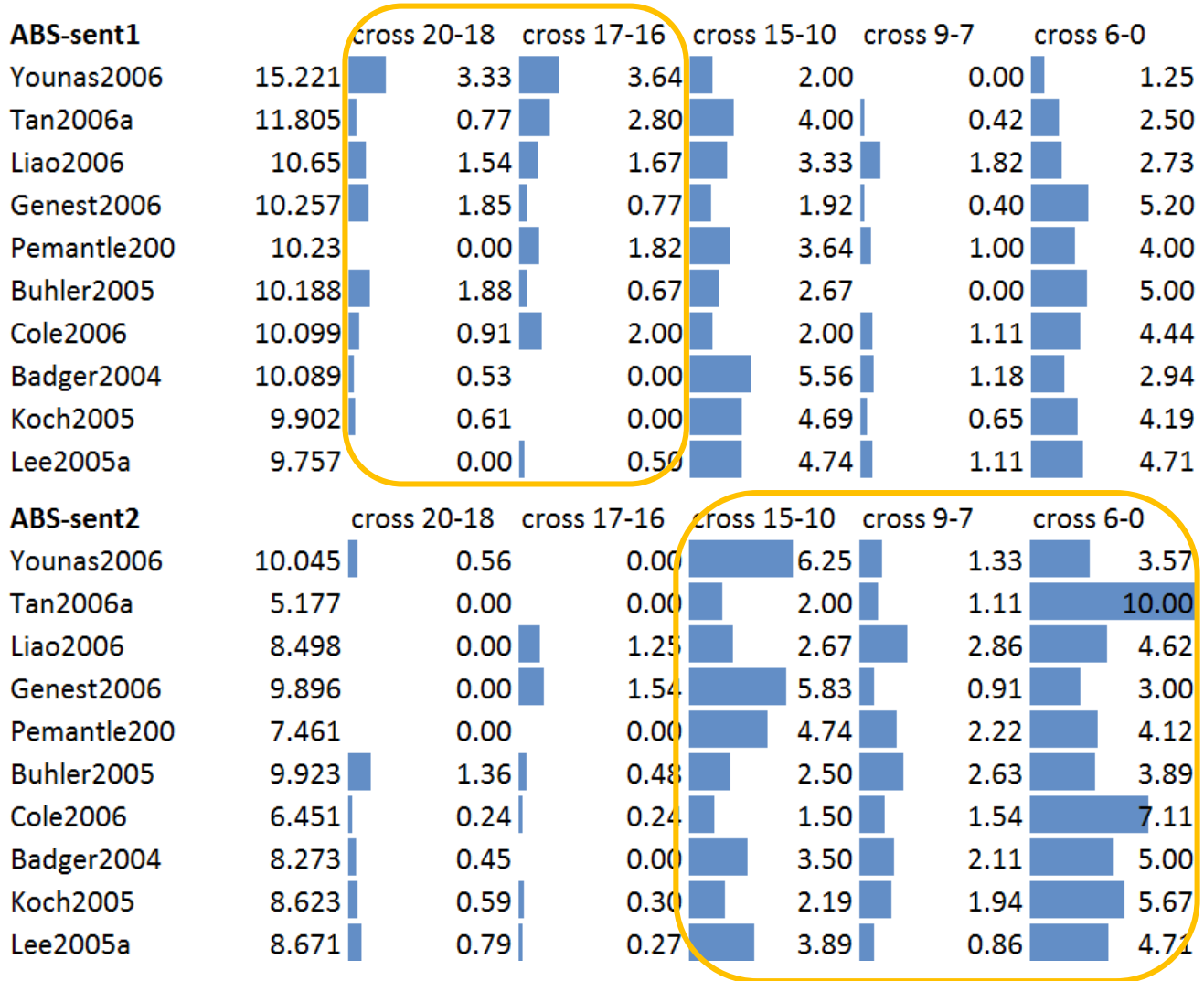
ABS-sent2	cross_sent									
Younas2006	10.045	Such	transactions	:	are	generally	complex	,	require	longer
Tan2006a	5.177	A	number	of	techniques	have	been	proposed	to	address
Liao2006	8.498	Password	authentication	is	one	of	the	simplest	and	the
Genest2006	9.896	They	usually	represent	incomplete	specifications	of	required	or	forbidden
Pemantle2005a	7.461	Milenkovic	and	Compton	in	2002	gave	an	analysis	of
Buhler2005	9.923	To	perform	large	comparisons	efficiently	,	BLAST	(	Methods
Cole2006	6.451	We	consider	a	model	of	selfish	routing	in	which
Badger2004	8.273	This	state	of	affairs	is	partly	due	to	the
Koch2005	8.623	A	security	policy	framework	specifies	a	set	of	(
Lee2005a	8.671	Because	of	the	unpredictability	of	the	order	of	update

cross 6-0
cross 9-7
cross 15-10
cross 17-16
cross 20-18

Analysis 2 – Word Cross-entropy Rates Abstracts-CompSci

Further inspection

- By other visualization options



Analysis 3
LMs on Abstracts vs RAs

Research design and Considerations

- LMs build based on pos trigrams
  (done by **Jonathan Poitz** based on his BSc-thesis)
- LMs build based on word trigrams
  (done by **Anna Currey**)

DASCITEX Corpus																	**Set / Instances**
Abstracts									RAs								
A Aggarwal2006 (sent 10) Andrews2004 (sent 18) …	B1	B2	B3	B4	C1	C2	C3	C4	A Aggarwal2006 (sent 10) Andrews2004 (sent 18) …	B1	B2	B3	B4	C1	C2	C3	C4	Training set ~ 2000 each
¼ of each discipline for equal distribution									¼ of each discipline for equal distribution								Test set ~ 500 each

Analysis 3
LMs on Abstracts vs RAs

Use of LMs

1. Train model on Abstracts/RAs and
   a. Test how well the models distinguish between the two (similar to a classification task)
      → are they distinct in terms of Inf Theory

   b. Compare relative ID of Abstracts and RAs to determine which have a higher ID
      → which is more informationally dense

2. Train model on Abstracts and test on RAs and train model on RAs and test on Abstracts
   → to identify linguistic patterns that are different

Analysis 3
LMs on Abstracts vs RAs

Preliminary analysis (done by Jonathan Poitz)

- Tested LM_Abs and LM_RAs on Abstracts/RAs
- Lower perplexity shows the right "class assignment"

abstracts	PPL ABS	PPL RAs	diff
Aaltonen2007_C3	7.26	8.29	1.03
Abney2004_B1	9.53	10.35	0.82
Adachi2006_C4	5.93	6.69	0.77
Adger2005_C1	7.82	8.11	0.29
Aggarwal2006_A	7.12	7.89	0.76
Aguilar2005_B3	6.29	8.32	2.03

RAs	PPL ABS	PPL RAs	diff
Aaltonen2007_C3	9.26	8.26	0.99
Abney2004_B1	12.72	10.41	2.31
Adachi2006_C4	9.02	7.89	1.13
Adger2005_C1	11.65	9.89	1.76
Aggarwal2006_A	10.17	8.89	1.28
Aguilar2005_B3	9.98	8.73	1.25

class	Abstracts	RAs
abs_A	46	10
abs_B1	51	5
abs_B2	56	0
abs_B3	53	3
abs_B4	56	0
abs_C1	43	12
abs_C2	54	1
abs_C3	53	2
abs_C4	55	0
RAs_A	0	56
RAs_B1	0	56
RAs_B2	0	56
RAs_B3	0	56
RAs_B4	0	56
RAs_C1	0	55
RAs_C2	0	55
RAs_C3	0	55
RAs_C4	0	55

→ Why 100% for RAs?

- First analyses are always very valuable!
- Help to reconsider research design!

For example:

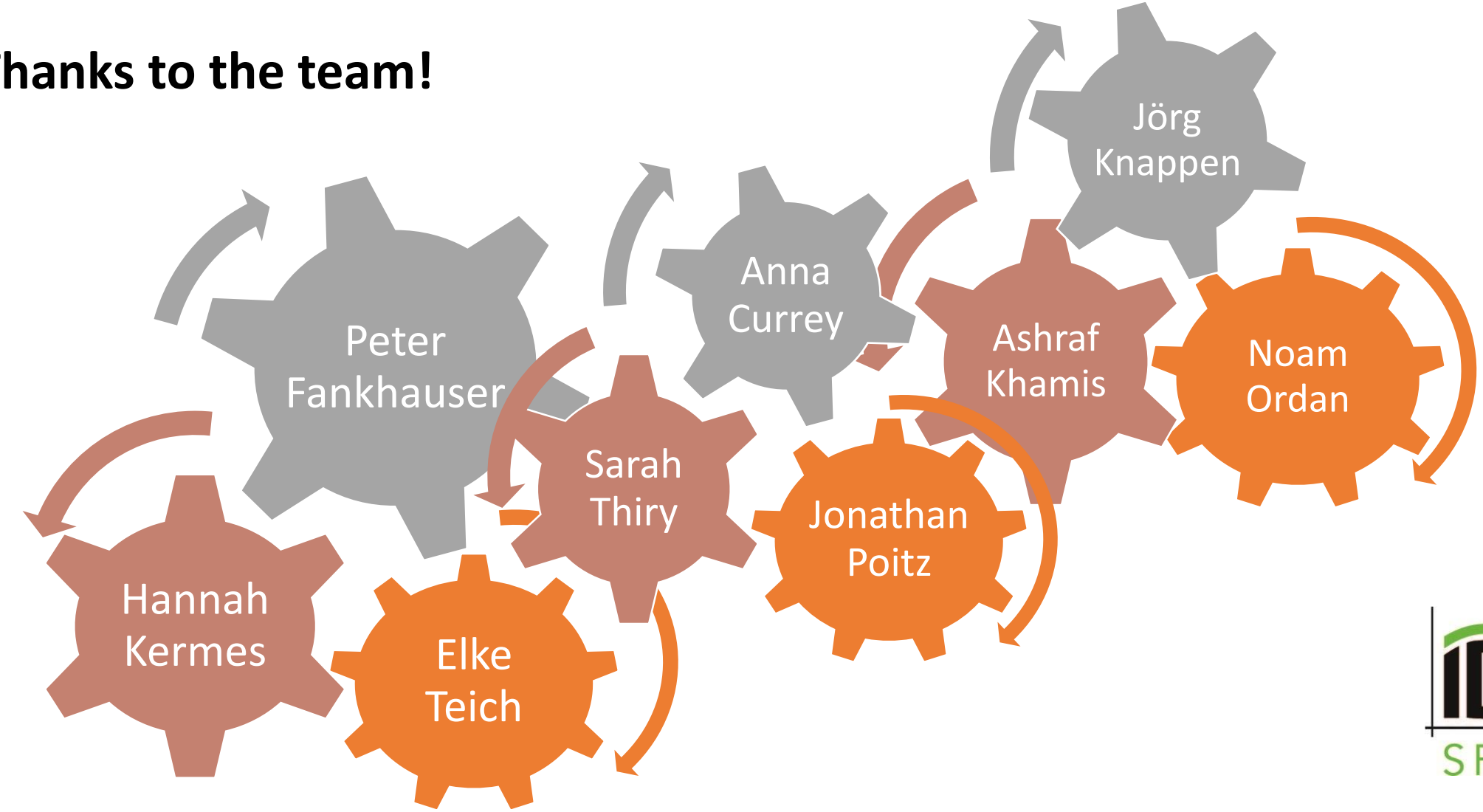
- Equal size of Abstracts and RAs
- Equal proportion of disciplines
- Similar data composition (include/exclude headlines, etc.)

Conclusion

- First analyses showed that Abstracts differ in terms of information density from RAs
- BUT: Research design and interpretation of the data is not trivial!
  - Have a good knowledge of our data
  - Gain a better understanding of what the methods can provide us with to ask the right question that can be answered by the appropriate methodology
  - Need of good visualizations
  - Etc.

Thank you for your attention!

Thanks to the team!



IDeaL
SFB 1102