B1: Information Density in Scientific Text

Methodological Considerations

Hannah Kermes, Noam Ordan, Elke Teich,
Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen,
Anna Currey, Jonathan Poitz, Sarah Thiery
&
Peter Fankhauser (IDS, Mannheim)

Inaugural Colloquium SFB 1102 - Saarbrücken
April 17, 2015

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

Research Questions

Is Information Density (ID) a driving force in the diachronic development of written scientific English?

Is there a relation between linguistic encoding and ID?

Assumptions

- Higher average information density in written scientific English 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 linguistic encoding and ID
  - Linguistic features marking specialization and conventionalization serve optimizing ID in scientific texts
    - longer, expanded ling forms — less predictable, more informative
    - shorter, reduced ling forms — more predictable, less informative

Contemporary Example

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. Inhibition of BCR-ABL signalling either by Imatinib or by RNAi silencing BCR-ABL reduces PRL-3 and increases cleavage of PARP.

- Complex nominal groups
- Simple clause structure: X <verb> Y
- Omission of determiners
- Global effects: high TTR, high lexical density, large number of technical terms

Contemporary Example

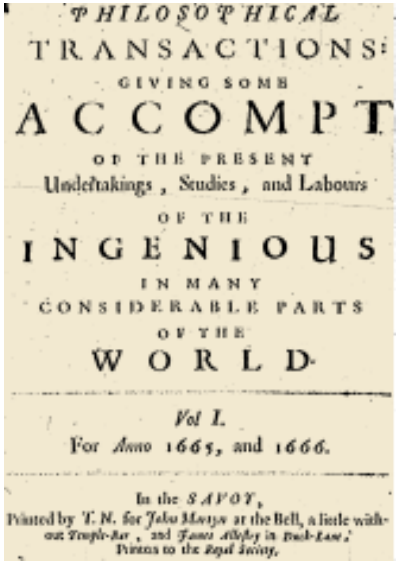
We report the discovery of a novel downstream target of BCR-ABL signalling, PRL-3 (PTP4A3), an oncogenic tyrosine phosphatase. Analysis of CML cancer cell lines and CML patient samples reveals the upregulation of PRL-3. Inhibition of BCR-ABL signalling either by Imatinib or by RNAi silencing BCR-ABL reduces PRL-3 and increases cleavage of PARP.

- Complex nominal groups
- Simple clause structure: X <verb> Y
- Omission of determiners
- Global effects: high TTR, high lexical density, large number of technical terms

Historical Example

Now those Colours argue a diverging and separation of the heterogeneous Rays from one another by means of their unequal Refractions, as in what follows will more fully appear. [. . .] 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.

(Newton, 1704)


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

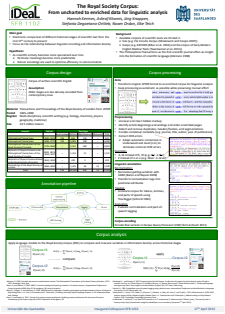
Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Pall-Mall, Printers to the Royal Society.

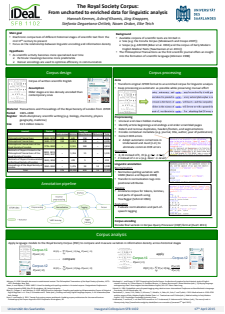
- Complex nominal groups
- Simple clause structure: X <verb> Y

Methodology

- Data: Royal Society Corpus, 33 M words, 1665 - 1869, discipline mix

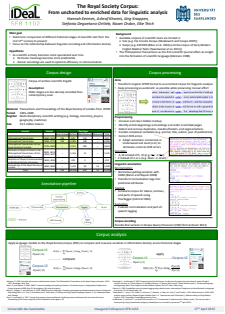


Methodology



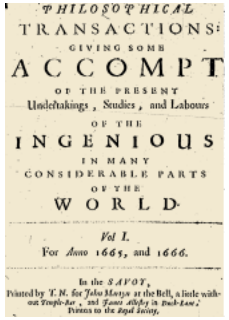
- Data: Royal Society Corpus, 33 M words, 1665 - 1869, discipline mix
- Corpus-based approach:
  - Compare linguistic feature distributions across time slices

Methodology



- Data: Royal Society Corpus, 33 M words, 1665 - 1869, discipline mix
- Corpus-based approach:
  - Compare linguistic feature distributions across time slices
- LM approach:
  - Measure average ID
  - Measure ID/surprisal

Corpus-based Approach


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 Pall-Mall, Printers to the Royal Society.

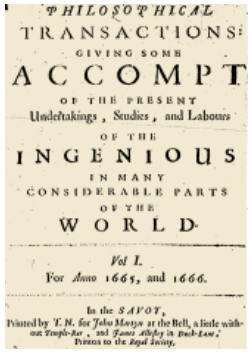
simple clause structure
complex NPs
determiner omission
TTR
lexical density
...

Corpus t1

Corpus t2

compare

LM Approach


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 Pall-Mall, Printers to the Royal Society.

Corpus t1

$P(unit|t1)$

$$H(t1) = -\sum_i P(unit_i|t1) log_2 P(unit_i|t1)$$

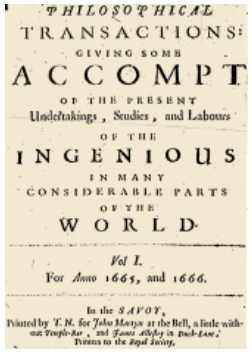
$\longrightarrow$ compare

Corpus t2

$P(unit|t2)$

$$H(t2) = -\sum_i P(unit_i|t2) log_2 P(unit_i|t2)$$

LM Approach


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 Pall-Mall, Printers to the Royal Society.

Corpus t1

$P(unit|t1)$

entropy

$$\boxed{H(t1)} = -\sum_i P(unit_i|t1) log_2 P(unit_i|t1)$$

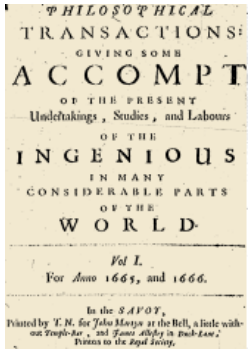
$\longrightarrow$ compare

Corpus t2

$P(unit|t2)$

$$H(t2) = -\sum_i P(unit_i|t2) log_2 P(unit_i|t2)$$

LM Approach


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 Pall-Mall, Printers to the Royal Society.

perplexity

$$2^{H(t1)}$$

entropy

Corpus t1

$P(unit|t1)$

$$H(t1) = -\sum_i P(unit_i|t1) log_2 P(unit_i|t1)$$

compare

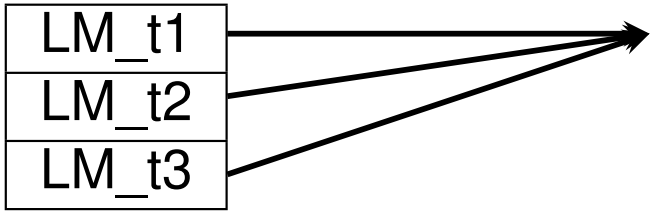
Corpus t2

$P(unit|t2)$

$$H(t2) = -\sum_i P(unit_i|t2) log_2 P(unit_i|t2)$$

LM Approach: Sample Analysis 1

- Divide Royal Society Corpus in 50 years slices
  - Keep training sets equal in size
- Train language models (3-grams) on one time slice
- Apply trained model to another time slice
  - Test how well the models can distinguish the time slices
  - Lower perplexity shows the right "class assignment"



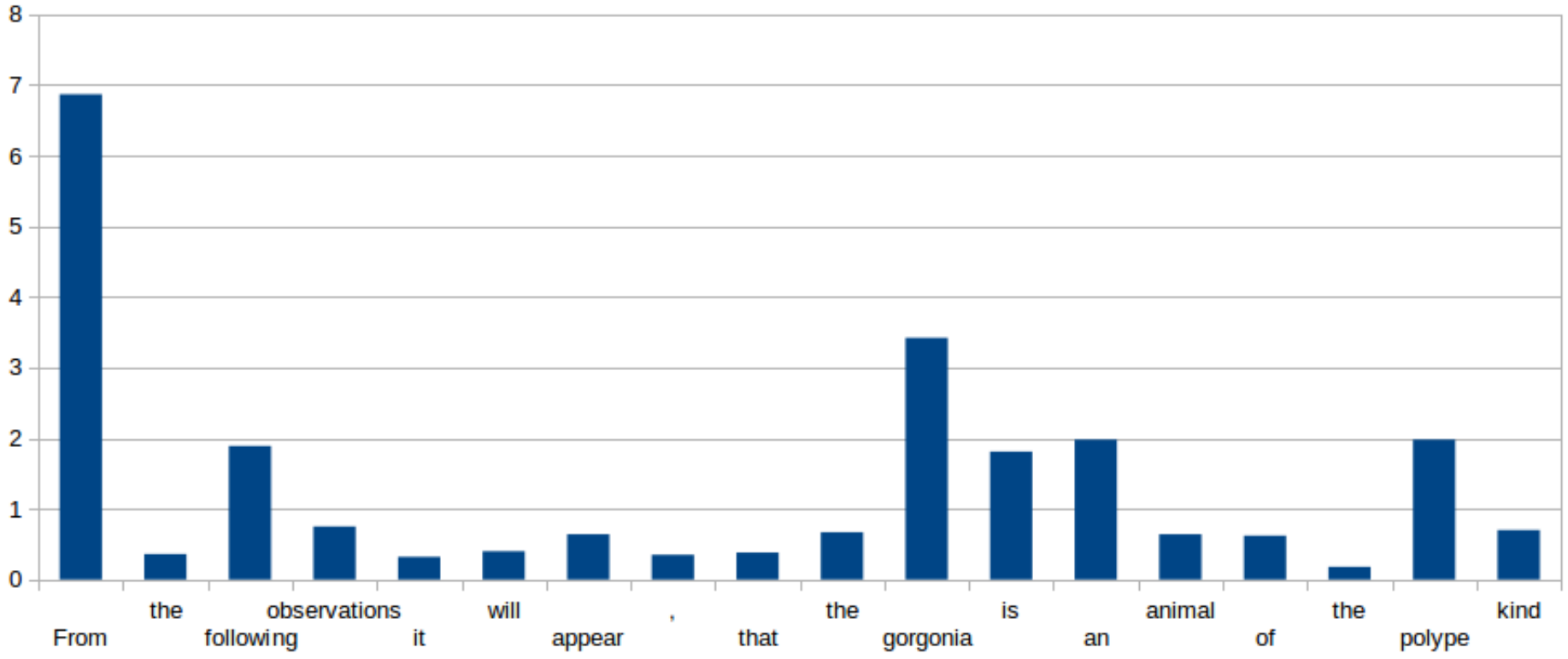
	LM_t1	LM_t2	LM_t3
text_1	PPL_t1	PPL_t2	PPL_t3
text_2	PPL_t1	PPL_t2	PPL_t3
text_3	PPL_t1	PPL_t2	PPL_t3
...	...	...	...

Results

	1665–1715	1716–1765	1766–1815	1816–1869
1665–1715	**507 (95%)**	26	1	1
1716–1765	37	**554 (89%)**	32	1
1766–1815	4	91	**791 (88%)**	16
1816–1869	0	0	188	**911 (83%)**

- Earliest and latest period well distinguished from one another
- Relation between close time periods, especially to the preceding time period

LM Approach: Sample Analysis 2



Outlook

- Elaborate experiment designs using LMs (B4)
  - Words, PoS, word+PoS, phrases, syntactic structures (Roark 2001)
  - Cache, document (Eisenstein/Barzilay 2008), corpus (Genzel/Charniak 2002), web (Microsoft n-grams)
  - In-domain vs. general

$\Rightarrow$ To measure average ID in scientific text over time

$\Rightarrow$ To find correlations between linguistic features and ID