Information Density and Linguistic Encoding

www.sfb1102.uni-saarland.de
Information Density and Linguistic Encoding

Assumptions (cf. Zipf, 1949; Shannon, 1949, 1951)

- language approximates an optimal code for human communication (close to channel capacity)

Information Density

\[ -\log_2 P(\text{unit} \mid \text{Context}) \]

- choice of a particular linguistic encoding depends on context
Surprise = Information

- **Expected low surprise** → little information
- **Not expected high surprise** → much information
$S(\text{unit}) = -\log_2 P(\text{unit} | \text{Context})$

**John accidentally mailed the letter without a *stamp*.**

**John went to the shop to buy a *stamp*.**

\[ -\log_2 P(\text{stamp} | \text{John accidentally mailed the letter without a}) \]

\[ -\log_2 P(\text{stamp} | \text{John went to the shop to buy a}) \]

---

**Context**

\[ (\log_2 2) \text{unit} = P(\text{unit}) \]

---

John accidentally mailed the letter without a *stamp*.

John went to the shop to buy a *stamp*.

John mailed....without a *stamp*  

more information (e.g. 4 bits)

less information (e.g. 1 bit)

John went... buy a *stamp*
Overall goals and structure

- Linguistic processes are driven by informational concerns
- Linguistic variation and change serve communication
- Focus: reduced vs. expanded linguistic encodings

Area A: Situational Context and World Knowledge
Area B: Discourse and Register
Area C: Variation in Linguistic Encoding
Project B1
Information Density in Scientific Writing: A Diachronic Perspective

Stefania Degaetano-Ortlieb, Stefan Fischer, Hannah Kermes, Ashraf Khamis, Jörg Knappen, Elke Teich

*UdS, Saarbrücken*

Peter Fankhauser

*IDS, Mannheim*
Assumptions: Scientific language over time

- more diversified (across domains)
- more specialized (within a domain)

Diversification, specialization
  → linguistic creativity
  → linguistic density

- more distinguishable as scientific
- more homogeneous

Conventionalization
  → linguistic uniformity
  → entrenchment

Optimal use of channel
This I did with much solicitude further inquire into; whereupon I found not only one hollowness, but as often as I cut the Nerve asunder, the hollowness still continued therein, and I found in some places not only one cavity, but two or three cavities at once;

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.

http://www.biomedcentral.com/
Data

**Scientific**
- 1506–1700
- 1665–1869
- 1970s & 2000s

**Mixed**
- Helsinki Corpus
  - 1500–1700
  - 1641–1700
- Archer
  - 1600–1999
  - 1640–1740
- Google Books

**EME (1500–1700)**
**LME (1700–1900)**
**ME (1900–present)**

**Brown Family**
Royal Society Corpus (RSC)

- Book reviews
- Articles
- Miscellaneous
- Obituaries

<table>
<thead>
<tr>
<th>Journal</th>
<th>Type</th>
<th>brv</th>
<th>fla</th>
<th>mis</th>
<th>nws</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philosophical Transactions (1665-1678)</td>
<td>124</td>
<td>641</td>
<td>154</td>
<td>-</td>
<td>919</td>
<td></td>
</tr>
<tr>
<td>Philosophical Transactions (1683-1775)</td>
<td>154</td>
<td>3903</td>
<td>338</td>
<td>-</td>
<td>4395</td>
<td></td>
</tr>
<tr>
<td>Philosophical Transactions of the Royal Society of London</td>
<td>-</td>
<td>2531</td>
<td>283</td>
<td>-</td>
<td>2814</td>
<td></td>
</tr>
<tr>
<td>Abstracts of Papers Printed ...</td>
<td>-</td>
<td>1316</td>
<td>15</td>
<td>-</td>
<td>1331</td>
<td></td>
</tr>
<tr>
<td>Abstracts of Papers Communicated ...</td>
<td>-</td>
<td>429</td>
<td>5</td>
<td>-</td>
<td>434</td>
<td></td>
</tr>
<tr>
<td>Proceedings of the Royal Society of London</td>
<td>-</td>
<td>1476</td>
<td>38</td>
<td>14</td>
<td>1528</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>278</td>
<td>10296</td>
<td>833</td>
<td>14</td>
<td>11421</td>
<td></td>
</tr>
</tbody>
</table>

Size: ca. 35 million tokens, source: XML (JSTOR)
Sample analyses

Analysis 1: Typical features over time
Grammatical patterns

Analysis 2: Linguistic encodings over time
N compounds vs. N-PREP-N

Macroanalysis: data-driven entropy

Microanalysis: hypothesis-driven surprisal
A1: Typical features over time

Data

• Royal Society Corpus (1665-1869)

Method

• Relative entropy + feature ranking
  Kullback-Leibler Divergence (KLD)

\[ D_{KL}(T1||T2) = - \sum_i p(\text{trigram}_i|T1) \log_2 \frac{p(\text{trigram}_i|T1)}{p(\text{trigram}_i|T2)} \]

• 1184 POS 3-grams
  (approximation of grammatical patterns)

(Fankhauser et al. 2014. Exploring and Visualizing Variation in Language Resources. LREC 2014)
### A1: Typicality in single time periods

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1650 vs</th>
<th>1750 vs</th>
<th>1800 vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP.VVZ.DT</td>
<td>he gives a</td>
<td></td>
<td>1750/1800</td>
<td></td>
</tr>
<tr>
<td>IN.PP.VVD</td>
<td>as I said</td>
<td></td>
<td>1800</td>
<td></td>
</tr>
<tr>
<td>PP.VVD.PP</td>
<td>I found it</td>
<td></td>
<td>1850</td>
<td></td>
</tr>
<tr>
<td>PP.VVD.DT</td>
<td>I found a</td>
<td></td>
<td>1850</td>
<td></td>
</tr>
<tr>
<td>IN.PP.VVD</td>
<td>as it appeared</td>
<td></td>
<td></td>
<td>1800</td>
</tr>
<tr>
<td>IN.PP.VBD</td>
<td>that it was</td>
<td></td>
<td></td>
<td>1800</td>
</tr>
<tr>
<td>PP.VVD.DT</td>
<td>I found the</td>
<td></td>
<td>1800/1850</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1700 vs</th>
<th>1850 vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN.DT.NN</td>
<td>on an inch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.NN</td>
<td>degree of heat</td>
<td>1650</td>
<td></td>
</tr>
<tr>
<td>DT.NN.IN</td>
<td>the quantity of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP.VVD.DT</td>
<td>I found the</td>
<td>1850</td>
<td></td>
</tr>
<tr>
<td>PP.VVD.TO</td>
<td>it seemed to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.NN</td>
<td>inch in diameter</td>
<td>1650/1700</td>
<td></td>
</tr>
<tr>
<td>IN.DT.NN</td>
<td>of an inch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT.NN.IN</td>
<td>the action of</td>
<td>1650/1700</td>
<td></td>
</tr>
<tr>
<td>DT.JJ.NN</td>
<td>the same time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.DT</td>
<td>part of the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN.DT.NN</td>
<td>inch in diameter</td>
<td>1650/1700/1750</td>
<td></td>
</tr>
<tr>
<td>NN.IN.NN</td>
<td>of an inch</td>
<td>1650/1700</td>
<td></td>
</tr>
<tr>
<td>DT.JJ.NN</td>
<td>the same time</td>
<td>1700</td>
<td></td>
</tr>
<tr>
<td>DT.NN.IN</td>
<td>the number of</td>
<td>1700</td>
<td></td>
</tr>
<tr>
<td>JJ.NN.IN</td>
<td>small quantity of</td>
<td>1750</td>
<td></td>
</tr>
</tbody>
</table>
# A1: Typicality in single time periods

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1650 vs</th>
<th>1750 vs</th>
<th>1800 vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP.VVZ.DT</td>
<td><em>he gives a</em></td>
<td>1750/1800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN.PP.VVD</td>
<td><em>as I said</em></td>
<td>1800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP.VVD. PP</td>
<td><em>I found it</em></td>
<td>1850</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP.VVD.DT</td>
<td><em>I found a</em></td>
<td>1850</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN.PP.VVD</td>
<td><em>as it appeared</em></td>
<td>1800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN.PP.VBD</td>
<td><em>that it was</em></td>
<td>1800/1850</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1700 vs</th>
<th>1850 vs</th>
<th>1800 vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN.PP.VVD</td>
<td><em>as it appeared</em></td>
<td>1800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN.PP.VBD</td>
<td><em>that it was</em></td>
<td>1800/1850</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- from **verbal** (PersPron + V) to **nominal** trigrams (rel. clauses + complex N groups (cf. Halliday 1988))
- If verbal, past tense (VVD) prevails

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1650 vs</th>
<th>1750 vs</th>
<th>1800 vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN.DT.NN</td>
<td><em>on an inch</em></td>
<td>1650</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.NN</td>
<td><em>degree of heat</em></td>
<td>1650</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.NN</td>
<td><em>the quantity of</em></td>
<td>1650</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT.NN.IN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT.JJ.NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN.IN.DT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>trigram</th>
<th>example</th>
<th>1650/1700/1750</th>
<th>1650/1700</th>
<th>1700</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN.DT.NN</td>
<td><em>of an inch</em></td>
<td>1650/1700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT.JJ.NN</td>
<td><em>the same time</em></td>
<td>1700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT.NN.IN</td>
<td><em>the number of</em></td>
<td>1700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JJ.NN.IN</td>
<td><em>small quantity of</em></td>
<td>1750</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A1: Diachronic tendencies

Verbal trigrams with personal pronoun

Nominal trigrams
### A1: Increasing typicality

<table>
<thead>
<tr>
<th>trigrams</th>
<th>example</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT.JJ.JJ</td>
<td><em>the same general</em></td>
<td>nominal</td>
</tr>
<tr>
<td>NN.TO.DT</td>
<td><em>respect to the</em></td>
<td></td>
</tr>
<tr>
<td>TO.DT.JJ</td>
<td><em>to the same</em></td>
<td>prepositional</td>
</tr>
<tr>
<td>IN.VVG.DT</td>
<td><em>for determining the</em></td>
<td>gerund</td>
</tr>
<tr>
<td>DT.NN.VBZ</td>
<td><em>the latter is</em></td>
<td>verbal; BE</td>
</tr>
<tr>
<td>VV.DT.JJ</td>
<td><em>produce the same</em></td>
<td>verbal; base form</td>
</tr>
<tr>
<td>VV.IN.DT</td>
<td><em>account for the</em></td>
<td></td>
</tr>
<tr>
<td>MD.VB.VVN</td>
<td><em>will be found</em></td>
<td></td>
</tr>
<tr>
<td>VB.VVN.IN</td>
<td><em>be considered as</em></td>
<td>verbal; passive</td>
</tr>
<tr>
<td>VBD.VVN.IN</td>
<td><em>were made with</em></td>
<td></td>
</tr>
<tr>
<td>VBZ.VVN.IN</td>
<td><em>is composed of</em></td>
<td></td>
</tr>
<tr>
<td>VVN.TO.VV</td>
<td><em>found to contain</em></td>
<td>verbal; to-inf</td>
</tr>
</tbody>
</table>

Information Density in Scientific Writing

SFB 1102
A1: Diachronic tendencies

Scientific language (RSC)

- Modal passive
- BE passive
- Past passive
- Present passive

General language (CLMET)

(cf. Diller, De Smet & Tyrkkö 2011)
A2: Linguistic encodings over time

Analysis

• Alternation between N compounds and N-PREP-N
  example: copper alloy  alloy of copper

Background

• No distinction between pre- and postmodification patterns in NP encoding density

• **Denser encoding mechanisms**: clausal subordination in speech (Biber, 1988); NP ‘heaviness’ in writing (Aarts, 1992: 83; Biber et al., 1999: 578)

Data

• N-N lemma sequences whose heads occur ≥10 times and their exact N-PREP-(DET)-N counterparts in RSC

• 53 matching heads and 139 variation patterns in 1665–1869
A2: Linguistic encodings over time

Methods

• Frequency distribution
• Average surprisal
• Syntagmatic context

Frequency distribution

Hypothesis
More N-N compounds, fewer N-PREP-N counterparts over time

Findings
• Increase in frequency for both constructions
• No predominance of compounds over time
A2: Frequency distribution

Frequency N-N compounds and exact N-PREP-N counterparts (normalized on the basis of nouns)
A2: Linguistic encodings over time

Average surprisal (AvS)

• Word-based model using three preceding words for context (cf. Genzel & Charniak, 2002)

Aims

• Comparison of mean of AvS in N-N and N-PREP-N constructions
• Individual investigation of all 139 variation patterns

Hypothesis

• Compounds are informationally denser than their prepositional counterparts

Findings

• Overall greater mean of AvS in N-N (6.91b) than in N-PREP-N (4.58b)
• Higher mean of AvS for c. 90% of compounds (depending on head noun)
A2: Average surprisal

Average surprisal (in bits) for *alloy* and *aperture* counterparts
A2: Linguistic encodings over time

Syntagmatic analysis

Aim

• Effect of lemma representation and surrounding parts of speech on mean AvS for N-N and N-PREP-N constructions

Focus

• 10 compounds with a lower mean of AvS than their N-PREP-N counterparts

ocean attraction, temperature correction, inch distance, mile distance, oxygen gas, copper ore, copper plate, brass ring, carbon vapour, copper wire
A2: Linguistic encodings over time

Lemma representation

• Plural head noun (e.g. ocean attractions, corrections of temperature) higher AvS than its singular counterpart

• Preposition other than of (e.g. gas with oxygen) element with highest AvS in N-PREP-N (counterintuitive)

Surrounding parts of speech

• N-N compound in a coordinated structure (e.g. mountain and ocean attraction) lower mean of AvS (more distributed information)

• N-N compound modified by an adjective versus a determiner higher mean of AvS (the opposite for N-PREP-N constructions)

Higher information value for (a) premodifying noun in N-N compounds (e.g. copper alloy) and (b) head noun in N-PREP-N constructions (e.g. alloy of copper)

First noun element has highest AvS
Envoi

Findings (so far)

• support of previous studies (e.g. Biber/Gray)
• additional features
• alternative encodings have different information densities

Uses of ID/surprisal

• method vs. linguistic property
• data-driven vs. hypothesis-driven
• macro- vs. micro-analysis

ID/surprisal as method

• context: $P(\text{unit}) \rightarrow P(\text{unit} | \text{context})$
• frequency ranges: $P(\text{unit}) \rightarrow \log_2 P(\text{unit} | \text{context})$
• Units of all linguistic levels: phone, morpheme, word, phrase, clause, ...
Publications


