Language – Text – Information

Elke Teich
Institut für Angewandte Sprachwissenschaft
sowie Übersetzen und Dolmetschen
My brief

- innovation
- interdisciplinarity
- science and humanities in the digital world
- myself
- 30 minutes
My approach

• **Reflection** on conditions of and dispositions in doing science

• **A semi-personal (gist) history** of linguistics

• **Some kind of conclusions**
Reflection

Science as a vocation
(1917)

→ Profession: specialization
→ Calling: passion, serve the scientific cause, work hard

„Das akademische Leben ist [ ] ein wilder Hasard“

(1864-1920)
Linguistics

• Research questions:
  – How is language structured?
  – How do we acquire language?
  – How do we use language?
  – How does language evolve?
  – How do we process language?

• Perspectives:
  – nature or nurture? → Language as biological or social system

• Methodological dispositions: rationalist or empiricist?
Language

S $\rightarrow$ NP VP
NP $\rightarrow$ Det N
VP $\rightarrow$ V NP PP
PP $\rightarrow$ P NP

(1957)

(*1928)
Language

- Language as *knowledge* (grammar rules, grammaticality), favors the *nature* perspective ("universal grammar")
- BUT: What about language use and variation?

> Gifscht mir e rood

> Gebbschd mier e roodi
A register is a cluster of associated features having a greater-than-random (or rather, greater than predicted by their unconditioned probabilities) tendency to co-occur.” (Halliday, 1988:162)

The use of this control method leads to a safer and faster train operation in the most adverse weather conditions.

(written)
You can control the trains this way and if you do that you can be quite sure that they’ll be able to run more safely and more quickly than they would otherwise, no matter how bad the weather gets.

(spoken)
Text

Language as contextualized choice

“A register is a cluster of associated features having a greater-than-random (or rather, greater than predicted by their unconditioned probabilities) tendency to co-occur.” (Halliday, 1988:162)

• **Linguistics**: comparison of frequencies in texts of different language varieties

• **Computational Linguistics**: statistical modeling of (aspects of) linguistic structure
Corpus: a large, structured collection of texts/utterances

A. PRESS: Reportage (44 texts) Political, Sports, Society Spot New, Financial, Cultural
B. PRESS: Editorial (27 texts) Institutional Daily Personal Letters to the Editor
C. PRESS: Reviews (17 texts) theatre books music dance
D. RELIGION (17 texts) Books, Periodicals, Tracts
E. SKILL AND HOBBIES (36 texts) Books, Periodicals
F. POPULAR LORE (48 texts) Books, Periodicals
G. BELLES-LETTRES - Biography, Memoirs, etc. (75 texts) Books, Periodicals
H. MISCELLANEOUS: US Government & House Organs (30 texts) Government Documents, Foundation Reports, Industry Reports, College Catalog, industry House organ
J. LEARNED (80 texts) Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences. Political Science, Law, Education, Humanities, Technology and Engineering
K. FICTION: General (29 texts) Novels Short Stories
L. FICTION: Mystery and Detective Fiction (24 texts) Novels, Short Stories
M. FICTION: Science (6 texts) Novels, Short Stories
N. FICTION: Adventure and Western (29 texts) Novels, Short Stories
P. FICTION: Romance and Love Story (29 texts) Novels, Short Stories
R. HUMOR (9 texts) Novels, Essays, etc.
• “Brown corpus”: **9,750** publications (Google Scholar)
• Brown **family** of corpora
• Studies on varieties, register, language change

<table>
<thead>
<tr>
<th></th>
<th>1960s</th>
<th>1990s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>Frown</td>
<td></td>
</tr>
<tr>
<td>LOB</td>
<td>FLOB</td>
<td></td>
</tr>
</tbody>
</table>

• For instance: (cf. Mair & Leech, 2006)
  – regularized morphology
    \[dreamt \rightarrow dreamed\]
  – elimination of \textit{shall} as future marker
  – increase in phrasal verbs
    \[ride \rightarrow give/take a ride\]
Text

- Language as *experience*, favors the *nurture* perspective
- BUT: Frequency alone

\[ P(\text{item}) \]

is not good enough…
A system for the grammatical annotation of natural language receives natural language text and annotates each word with a set of tags indicative of its possible grammatical or syntactic uses.
An empirical probability of collocation function defined on pairs of tags is iteratively extended to a selected set of tag sequences of increasing length so as to select a most probable tag for each word of a sequence of ambiguously-tagged words.
Information

- Language use relies on expectations
- Predictability in context can be appropriately indexed by Shannon’s (1948) notion of information

\[ P(item \mid Context) \]

\[ Surprisal(item) = -\log_2 P(item \mid Context) \]
Information

John accidentally mailed the letter without a *stamp*. 

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

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

- \( \log P(stamp | \text{John went to the shop to buy a}) \)
Information

• Perspective of information can help to find a unifying explanation for language use:
Language strives for an optimal code

• Challenges:
  – Data with **contextual information** (local linguistic and extra-linguistic context)
    $P(item \mid context)$
  - **Interesting** and **relevant** linguistic **features**
    $P(item \mid context)$
→ To provide more contexts is a **humanistic** task

American English

British English

→ To provide methods is a **computational** task
Conclusions (1)

Only by combining the three perspectives can we advance our understanding of how language functions as a part of modeling human experience.

This is clearly an interdisciplinary endeavor.
Conclusions (2)

• Do you have a disposition for science?

• If so, become a specialist, then work in an interdisciplinary fashion

• **Innovation** is most likely to happen in interdisciplinary contexts

• Look for places to work that welcome this spirit: **Saarbrücken** is such a place!
Kunst und Wissenschaft, Forschung und Lehre sind frei.

(Grundgesetz, Artikel 5, Absatz 3, Satz 1)
The end