SFB 1102
Information Density and Linguistic Encoding

Elke Teich

Inaugural Colloquium
April 17, 2015
Language offers a wide range of options of how to encode a message.

Variation is an inherent property of the linguistic system.
1. a. *My boss confirmed* that he is absolutely crazy.
   b. *My boss confirmed* he is absolutely crazy.

2. a. *Wo soll ich das Zeugs hintun?*
   b. *Wohin mit dem Zeugs?*

3. a. *If this method of control were to be used,* trains would operate more safely.
   b. *The use of this control method leads to safer train operation.*
4. a. Paid jobs degrade the *mind*.  
b. Mama you’ve been on my *mind*.
Observations and Main Question

- Options are available at all levels of the linguistic system: phonetic, morphological, lexical, syntactic, discourse
- Choices are dependent on different kinds of context: local (e.g. syntactic, phonetic) vs. global (e.g. situation, text type)

Is there a unifying explanation?
Language processing relies on \textit{predictability in context}.

Contextually determined predictability can be appropriately indexed by the notion of information.

\[
\text{Surprisal}(\text{unit}) = \log_2 \frac{1}{P(\text{unit} | \text{Context})} = -\log_2 P(\text{unit} | \text{Context})
\]
5. a. John accidentally mailed the letter without a **stamp**.
   b. John went to the shop to buy a **stamp**.

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

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

\[
\text{Effort(unit)} \propto \text{Surprisal(unit)}
\]
Uniform Information Density

- Speakers exploit linguistic variation to avoid peaks and troughs in information density
- Speakers modulate the order, density and specificity their linguistic encoding

Message $M$
Goals

- Investigate the extent to which the notion of optimal distribution of information offers a common explanation of patterns of variation
- Investigate the role of different kinds of context as determinants of predictability
- Investigate linguistic encodings at different linguistic levels

\[
\text{Surprisal}(\text{unit}) = -\log_2 P(\text{unit} \mid \text{Context}) \\
= -\log_2 P(\text{word} \mid \text{Script}) \\
= -\log_2 P(\text{syntactic } \text{unit} \mid \text{Discourse}) \\
= -\log_2 P(\text{phone} \mid \text{Collocation})
\]
Methods

experiment/corpus

derive

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

design

validate

compare

analyze

experiment

corpus

production/comprehension

register, languages, diachrony
Contributions

Information theory for linguistic inquiry

- Find communicative explanations for aspects of language use, variation and change

- Transcend disciplinary boundaries through one unifying approach: psycholinguistics, computational linguistics, phonetics, socio-linguistics, contrastive linguistics, historical linguistics, semantics
Research areas

A Situational Context and World Knowledge
Brings non-linguistic context into characterizations of surprisal

B Discourse and Register
Examines the relation between encoding and information density at the level of text

C Variation in Linguistic Encoding
Offers information density explanations for encoding choices across linguistic levels and languages
Today’s Program

11:00 Guest talk: **Ted Gibson** (MIT)
*Language for communication: Language as rational inference*

lunch break

13:00 **Francesca Delogu** (Projekt A1)
*Script-based surprisal: Evidence from event-related potentials*

13:20 **Ekaterina Kravtchenko** (Projekt A3)
*The processing of predictable events in a script context*

13:40 **Hannah Kermes** (Projekt B1)
*Information density and scientific literacy in English – preliminary analyses using language modeling*

14:00 **Vera Demberg** (Projekt B2)
*On the information conveyed by discourse connectives*

14:20 **Zofia Malisz** (Projekt C1)
*The relationship between information rate and speech rate in several European languages*

coffee break

15:00 SFB 1102 Poster Session

16:00 Guest talk: **Florian Jaeger** (Rochester)
*Processing efficiency shapes language: Natural languages have lower than expected information density*
Evening event