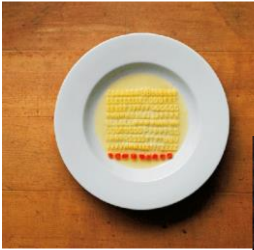
Use and usability of data mining for linguistic analysis

March 13, 2015
organized by Project
B1: Information density and scientific literacy in English

SFB 1102
Information Density and Linguistic Encoding

- **Language variation** acc. to context: situation, time, region; across languages
- **Data mining/text analytics**: machine learning, exploratory data analysis, language modeling



Research Focus in SFB 1102

Language Use

- Language offers a wide range of options of how to encode a message

Linguistic Variation

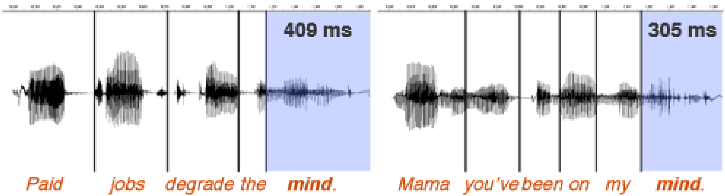
- Variation is an inherent property of the linguistic system

(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?

Hypothesis

- Language processing relies on predictability in context
- Contextually determined predictability can be appropriately indexed by Shannon’s notion of information

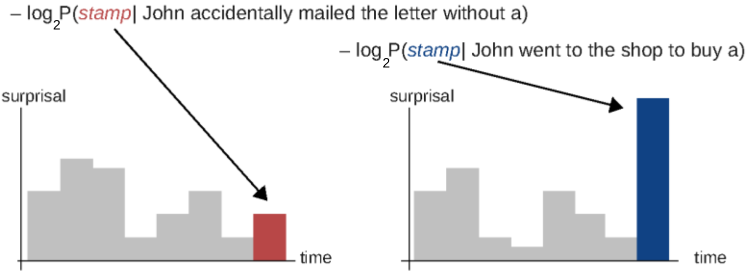
Information Density (ID)
Surprisal

$$Surprisal(unit) = \log_2 \frac{1}{P(unit \,|\, \mathrm{Context})} = -\log_2 P(unit \,|\, \mathrm{Context})$$

(5) a. *John accidentally mailed the letter without a stamp .*

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

$$Surprisal(unit) = -\log_2 P(unit \,|\, \text{Context})$$



– $\log_2$P(*stamp*| John accidentally mailed the letter without a)

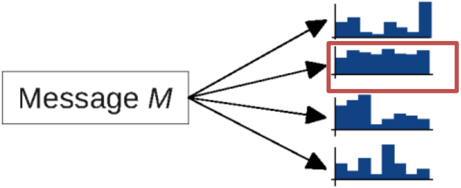
– $\log_2$P(*stamp*| John went to the shop to buy a)

John mailed....without a stamp *John went... buy a stamp*

$$Effort(unit) \propto Surprisal(unit)$$

Uniform Information Density

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



preferred encoding

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

$$\begin{aligned} Surprisal(unit) &= -\log_2 P(unit \mid \mathrm{Context}) \\ &= -\log_2 P(word \mid \mathrm{Script}) \\ &= -\log_2 P(syntactic\_unit \mid \mathrm{Discourse}) \\ &= -\log_2 P(phone \mid \mathrm{Collocation}) \end{aligned}$$

Linguistic Variation

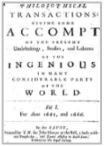
- ID and use of fragments (Project B3)

  Wohin mit dem Zeugs?
  Wo soll ich das Zeugs hintun?

- Cross-linguistic ID: Slavic languages (Project C4)

🇨🇿	🇵🇱
Základním posláním	*Podstawowym zadaniem*
Česko-polského fóra	*Forum Polsko-Czeskiego*

- Diachronic ID: English scientific language (Project B1)



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.

1665

An Account of Some Observations Concerning Tides, Made by Mr. Samuel Colepresse at and nigh Plimouth, An. 1667.

CTLA4 overexpressing adipose tissue-derived mesenchymal stem cell therapy in a dog with steroid-refractory pemphigus foliaceus

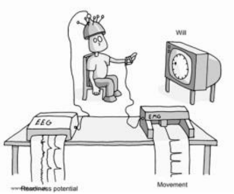
BioMed Central
The Open Access Publisher

Articles

Cases Database

2015

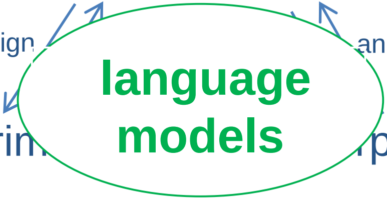
Methods



ex

us

P(*unit* | Context)



design

analyze

language models

experim

rpus

production/comprehension

register, languages, diachrony

Uses of LM (DM)

- Capture linguistic variation
- Measure ID locally and for whole texts/corpora
- Compare ID across texts/corpora within a language and across languages
- Tease apart different linguistic levels (e.g. lexical vs. syntactic) wrt their contributions to ID
- Help us find out which linguistic features (if any) are mainly involved in modulation of ID

Usability of LM (DM)

- How do LMs measure ID/surprisal?
- How can language models be compared?
  → relative ID
- How best to build language models?
  How to avoid mistakes?
  → e.g. different background models, smoothing techniques
- How to make language models accessible for linguistic interpretation (by human, by machine)?
  → e.g. visualization

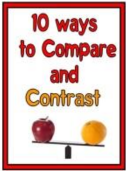
Program

09:20-10:00	**Jon Dehdari**, *An Overview of Language Modeling and its Applications*
10:00-10:40	**Dietrich Klakow**, *Practical Applications for Language models*
10:40-11:10	*Coffee break*
11:10-11:50	**Peter Fankhauser**, *Observing Surprisal through the Blurry Lense of Language Models*
11:50-12:30	**Jilles Vreeken**, *Mining Sequential Patterns*
12:30-13:30	*Lunch*
13:30-14:10	**Steffen Koch**, *Trends and Topics: How Visual Approaches Foster Synergetic Effects by Combining Linguistic Analyses and Interactive Exploration*
14:10-14:50	**Stefan Evert**, *A Multivariate Approach to Linguistic Variation and Distribution*
14:50-15:20	*Coffee break*
15:20-15:40	**B1**: *Information Density and Scientific Literacy in English*
15:40-16:00	**B3**: *Information Density and Fragments in German*
16:00-16:20	**C4**: *Modeling mutual intelligibility between Slavic languages*
16:20-17:30	*Discussion*

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 choice in language encoding across linguistic levels