Computational Psycholinguistics

Integrating Computational and Behavioral Methods to Study Human Language Processing

T. Florian Jaeger (BCS)
Daniel Gildea (CS)
Lenhart K. Schubert (CS, LIN)
Michael K. Tanenhaus (BCS, LIN)
The Minds Behind the Project ...

Austin Frank (BCS)
Carlos Gomez Gallo (CS/LIN)
Celeste Kidd (BCS)
Matt Post (CS)
Ting Qian (BCS)
Benjamin Van Durme (CS/LIN)
Computational Psycholinguistics

- **Proposal:** Combine Rochester’s strengths in CS and BCS to...
  
  ... employ computational *theory* and *methods* to understand how the mind/brain acquires, maintains, and employs probability distributions to efficiently process language.
Probabilities in Language Processing

- **Frequency** of words $\sim$ production & recognition latencies

spoken word recognition

‘Click on the bench’

More frequent $\rightarrow$ more first fixations

[DahanETAL01]
Probabilities in Language Processing

- **Contextual probability** of words ~ production & recognition

![Graph showing word duration in spontaneous speech](image)

- Paid jobs degrade the **mind**
- I don’t **mind** going to a wedding

More probable (=less information) → shorter
Pres. Clinton did *not* have ...
Pres. Clinton did *not* have ...
‘Choices’ at many levels in production

Utterance level: Move the triangle to the left.
Select the triangle. Move it to the left.

Phrasal level: She gave {him the key/the key to him}
She already ate (dinner)
She stabbed him (with a knife).

Word level: I read a book (that) she wrote.

Morphological level: I’ve\have gone there.

Phonological level: t/d-deletion; vowel weakening

Phonetic level: formant energies, F1/F2 ratio, speech rate
Across Levels of Linguistic Processing: More Information $\rightarrow$ More Form
Across Levels of Linguistic Processing: More Information ➔ More Form
Across Levels of Linguistic Processing: More Information → More Form

Information content of HAVE: I(HAVE | upcoming context)
Across Levels of Linguistic Processing: More Information → More Form

Information content of SRC onset: \( I(\text{SRC} | \text{preceding noun}) \)

Information content of SRC onset: \( I(\text{SRC} | \text{participle}) \)
What Cues do we Track?

And he gave me the same medication.
What Cues do we Track?

Predictive features
trigram

UNIVERSITY OF ROCHESTER
What Cues do we Track?

Predictive features
- trigram
- syntax

And he gave me the same medication
What Cues do we Track?

Predictive features

- trigram
- syntax
- trigger

medication..... And he gave me the same medication
Cue Integration

- **Seeded:** NSF BCS-0844472 (w/ Gibson, MIT)
  Collaborative Research: Bayesian Cue Integration in Probability-Sensitive Language Processing
Probabilistic Counting

- But how do you track all that information?

- Over 400 million common trigrams in the English language (~ 5.6 GByte gzipped) → ~2GByte to store counts for all trigrams

- Limiting memory to about 5% of that, we can still get approximate counts with an error rate of ~12%
Estimating Probabilities

- Traditional databases:
  - Expensive to create ...
  - ... and to maintain (language changes)
  - Data sparsity
  - Limited to a few languages

- The web as data source could potentially overcome some of these short-comings
WWW-based Corpora

- Estimated 25+ billion web pages in many languages.

- E.g. Google 5-gram database of English
  - $10^3$-times larger than traditional corpora:
    - Number of words: $1,024,908,267,229$
    - Number of unique words: $13,588,391$

- Similar data sets available or cheap to create for a larger number of languages
WWW can Improve Estimates

- Lexical decision
- Word naming
- Picture naming

Combined estimates

log(RT) (centered) vs. log(Frequency) (centered)
Better estimates of language experience of young folks
Summary

- Speakers use probabilistic information to efficiently produce language
- We’re investigating
  - What cues do speakers track to derive probability estimates?
  - How speakers manage to keep track of large numbers of cues?
- Web may help us to overcome methodological challenges in this line of work
Thanks