Does Information Theory constrain the usage of fragments? – An experimental study

Robin Lemke Lisa Schäfer Ingo Reich

IDeaL SFB 1102

Universität des Saarlandes

Background: Fragments and Information Theory

Fragment (following Morgan, 1973)

Antecendentless overtly nonsentential utterance which expresses propositional content and has illocutionary force

(1) a. [Passenger to taxi driver:] "To the university, please."
b. [Ticket inspector to passenger:] "Your ticket, please."

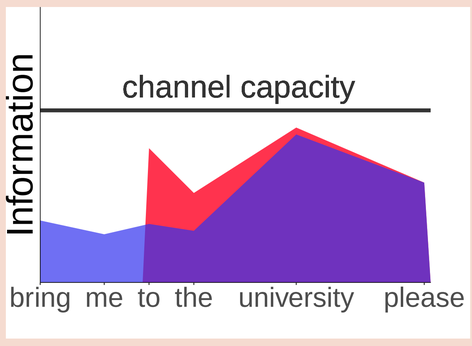
When are fragments licensed?

- Delete up to recoverability (Barton, 1998)
- When a salient antecedent is available (Stainton, 2006)
- When there is a high probability for the hearer to correctly interpret it (Bergen and Goodman, 2014)

Uniform Information Density (Levy and Jaeger, 2007))

- Information (or Surprisal, Hale 2001) of a word can be measured as $\mathbf{S = -\log p(word \mid context)}$
- Speakers intend to communicate at a uniform rate without exceeding the capacity of the communicative channel

Hypothesis



- Very (un)predictable words can cause in peaks/troughs in the information density profile
- Fragments are preferred over sentences when they are more well-formed w.r.t. UID

Estimating event probabilities from a script corpus

DeScript corpus (Wanzare et al., 2016)

- Crowd-sourced corpus of script knowledge
- 100 individual descriptions for 40 scenarios
- Description in list form, many elliptical utterances

Method

1. Parse corpus with Stanford parser (Klein and Manning, 2003)
2. Extract events (verb + noun, Manshadi et al., (2008))
3. Manually annotate ellipses and disambiguate pronouns
4. Manually unify synonym events
5. Calculate bigram language models (SRILM, Stolcke, (2002))

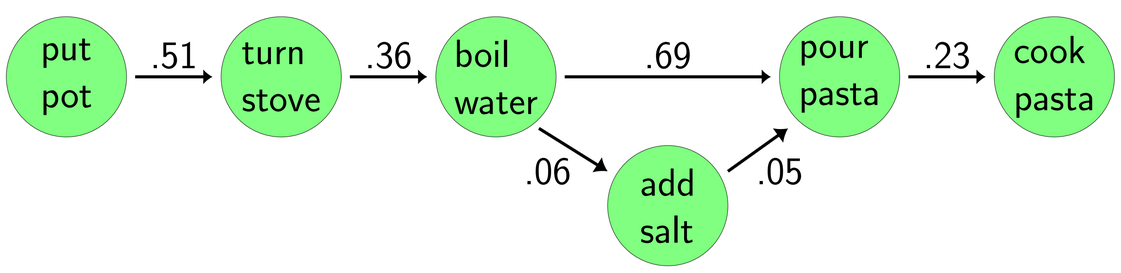


Figure 1: Sample event probabilities in the pasta scenario in DeScript.

Sample item (original in German)

(2) Today Annika and Jenny want to cook a large serving of pasta. Annika put a pot with water on the stove. Then she turned the stove on. After a few minutes, the water started to boil.
Now Annika says to Jenny:

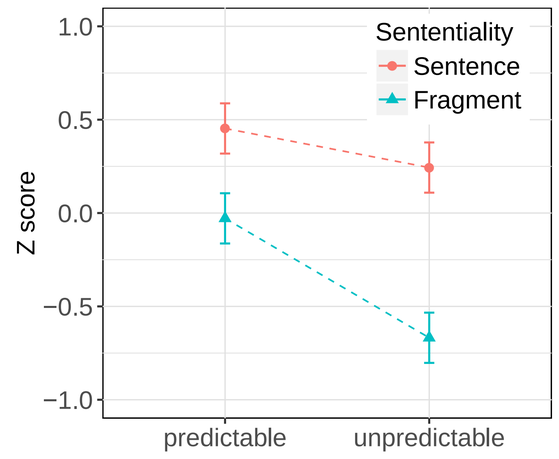
a.	Pour the pasta into the water, please!	Predictable
b.	Set the table, please!	Unpredictable
c.	The pasta, please!	Predictable
d.	The table, please!	Unpredictable

Acceptability rating study

Method

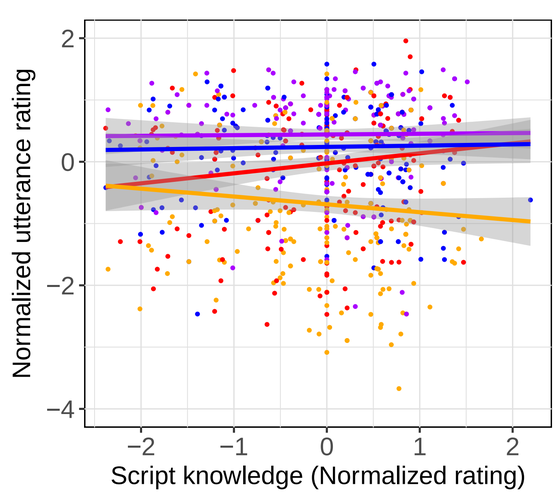
- 2 × 2 design, Sententiality × Predictability
- Naturalness rating on 7-point Likert scale
- Crowd-sourced via clickworker, 48 subjects, German
- 24 items + 44 fillers + 21 items from another experiment
- Analysis with CLMMs (`R`, `ordinal` (Christensen, 2015))

Predictability improves ratings for fragments . . .



- Significant main effect of Sententiality ($|z| = 6.01, p < .001$)
- Significant main effect of Predictability ($|z| = 4.54, p < .001$)
- Significant interaction ($|z| = 3.13, p < .01$)

. . . specifically when people have script knowledge



- ScriptKnowledge: Predictability is significant for for predictable/unpredictable fragments ($|z| = 3.34$, $p < .001$), but not for predictable/unpredictable sentences

Discussion and next steps

- As expected, predictable fragments are significantly better than unpredictable ones, but even those are worse than full sentences → Politeness constraint?
- The more script knowledge subjects have, the more they differentiate between predictable and unpredictable fragments
- Production study: Gather naturalistic data on what people would *say* in described situations
- Rating study with production data: Are specifically the informative, entropy-reducing, parts of the utterance maintained?

Selected References Levy, Roger P and T Florian Jaeger (2007). "Speakers optimize information density through syntactic reduction". In: *Advances in neural information processing systems*. Ed. by B Schlökopf, J Platt, and T Hoffman, pp. 849–856. • Manshadi, Mehdi, R Swanson, and A. S. Gordon (2008). "Learning a probabilistic model of event sequences from Internet weblog stories.". In: *FLAIRS Conference*, pp. 159–164. • Morgan, J L (1973). "Sentence fragments and the notion sentence". In: *Issues in linguistics: Papers in honor of Henry and Rene Kahane*. Ed. by B Kachru et al., pp. 719–751. • Wanzare, Lilian D. A. et al. (2016). "DeScript: A crowdsourced corpus for the acquisition of high-quality script knowledge". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)*. Portorož, Slovenia, pp. 3494–3501.