Script-based discourse expectations trigger omissions in fragments – Evidence from production data

Robin Lemke, Lisa Schäfer, Heiner Drenhaus & Ingo Reich (Saarland University)
robin.lemke@uni-saarland.de

Subsentential utterances, so-called *fragments* (Morgan, 1973) (1a), are often used to communicate the same meaning as full sentences (1b). Fragments have been studied extensively from a syntactic perspective, however, the question of why specific omissions (e.g. that of *would you like*, but not of *slice* in (1)) occur is relatively unexplored. We investigate the hypothesis that such omissions are constrained by the Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007), which predicts that words that are *predictable* in context are likely to be omitted. While previous evidence for UID is based on effects of (local) linguistic context, we extend these results to script-based discourse expectation as a source of predictability.

(1) a. [Ann and Bill are sharing a pizza. Ann asks Bill:] Another slice?
b. [Ann and Bill are sharing a pizza. Ann asks Bill:] Would you like another slice of pizza?

Uniform Information Density The central idea of UID is that information – defined as the negative logarithm of a word's probability: $I = -log_2 p\,(word|context)$ (Shannon, 1948) – is best distributed uniformly across an utterance. As information indexes processing effort (Levy, 2008), this makes the most efficient use of the limited cognitive resources available to the hearer. Previous research has shown that speakers can optimize their utterances with respect to UID by omitting uninformative function words, as complementizers (Jaeger, 2010) and relative pronouns (Levy and Jaeger, 2007). If this finding could be extended to content words as verbs and nouns like *like*, *would* and *pizza* in (1), UID would also explain the choice between fragments and full sentences: A fragment as (1a) is preferred over the corresponding full sentence (1b) if UID favors the omission of highly predictable words that are obligatory in the full sentence.

Production study Investigating this empirically requires first a set of linguistic data containing the relevant omissions and, second, information estimates for both the omitted and realized words in that data set. Logistic regressions can show whether lower information significantly increases the likelihood of a word's omission. Most of the previous information-theoretic studies on omissions used corpus data. However, as fragments often occur discourse-initially, the predictability of words in fragments depends to a large extent on extralinguistic context that cannot be retrieved from corpora. Therefore, we used a production study to collect a data set of utterances for a set of tightly controlled contexts referring to script knowledge (Schank and Abelson, 1977). We presented subjects 24 short stories like (2) (originals in German) and asked them to produce the utterance that they considered most likely to be said by the specified character in that context.

(2) Annika and Jenny want to cook 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:

Since scripts prime upcoming events (Bower et al., 1979; Delogu et al., 2018; van der Meer et al., 2002), they should shape expectations about what will be said in a script-based situation. For instance, in (2) a request to pour the pasta into the pot or to give the speaker the pasta seems probable. In order to use scripts as a model of extralinguistic context, we based our materials on event chains (Manshadi et al., 2008) extracted from the crowd-sourced DeScript corpus (Wanzare et al., 2016) of script knowledge. All stories had the same structure: The first sentence of the story introduces the script, whereas the next three ones refer to events likely to follow each other. For each story, we collected responses from 100 participants recruited via the crowdsourcing platform Clickworker. As there was a high degree of variation in the data and we are interested in content words, i.e. nouns and verbs, we preprocessed the data by resolving pronouns and ellipses, removing all function words, lemmatizing the remaining words and finally pooling synonyms to a single lemma. This transforms utterances like (3a) to an abstract representation (3b).

(3) a. pour the pasta into the pot!
b. put pasta pot

Estimating information We used the *unigram information* $I = -log_2 p(word)$ of each expression in the preprocessed data as measure of its likelihood. As we calculated information by training individual language models on the data for each script (with the SRILM toolkit, Stolcke (2002)), in this case unigram information reflects the likelihood of a word to appear in an utterance in context of the script-based story representing extralinguistic context, i.e. $I = -log_2 p(word|context)$. Training n-gram language models on data containing omissions like fragments potentially introduces a circularity issue (Levy and Jaeger, 2007: 852): If UID is correct, predictable words are omitted more often than unpredictable ones, so that their measured frequency is not proportional to their predictability. In order to prevent this, we trained our language models on an enriched version of the data. We added those expressions that are minimally required in a full sentence, that is, missing verbs and/or their arguments, and recorded whether each of the expressions in the enriched data set had been omitted or realized in the original data set.

Results Figure 1 shows the distribution of information estimates for omitted and realized words. We used mixed effects logistic regressions (`lme4`, Bates et al. (2015)) predicting the OMISSION of an expression in the enriched data set described in the preceding paragraph from its unigram INFORMATION. The model also contained two control predictors which might also constrain omission (and actually do so): PARTOF-SPEECH (noun/verb) of the word and its POSITION in the utterance (numeric). All binary predictors were sum-coded. Models contained by-script and by-subject random intercepts and by-script random slopes for INFORMATION. The final model contains a significant main effect of INFORMATION ($z = -2.7, p < 0.01$) that shows that, as our hypothesis predicts, predictable words are indeed more likely to be omitted. In our presentation, we also discuss similar results obtained with a different method of estimating surprisal, which is based on the approach by Hale (2001) and takes the linguistic material preceding a word into account.

Discussion Our analysis shows that, as UID predicts, words that are more likely in context are more often omitted. UID however seems not to be the only factor in determining whether fragments are used. For instance, the ratio of fragments varies strongly across scripts (between 0% and 80%, mean = 24.2%). Even some scripts with a similar mean information, as the sauna script (mean $I = 3.28$) and the dentist script (mean $I = 3.16$), strongly differ in fragment ratio (60.5% vs. 1.9%). In fact, if only those scripts with an overall ratio of fragments above the median (15.9%) are analyzed as described above, the UID effect is even more pronounced ($z = -3.6, p < .0001$). Taken together, our data support the hypothesis that omissions in fragments are driven by expectations arising from script-based context. This extends previous evidence for UID in two ways: First, we find UID effects not only in function, but also in content words, and, second we show that not only (local) linguistic context but also discourse expectations based on script knowledge determine predictability and consequently the likelihood of omissions.

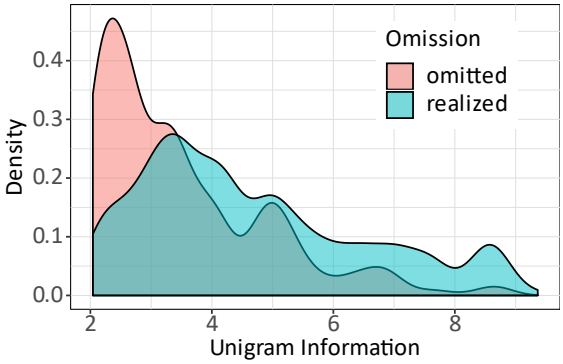

Figure 1 Distribution of omitted and realized words as a function of unigram information.

Selected references Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. Cognitive Psychology, 11(2). 177–220. • Delogu, F., Drenhaus, H., and Crocker, M. W. (2018). On the predictability of event boundaries in discourse: An ERP investigation. Memory & Cognition, 46(2). 315–325. • Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. Proceedings of NAACL (Vol. 2). 159–166. • Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3). 1126–1177. • Levy, R. P. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schlökopf, B., Platt, J., and Hoffman, T., editors, Advances in Neural Information Processing Systems. 849–856. • Morgan, J. (1973). Sentence fragments and the notion 'sentence'. In Kachru, B., Lees, R., Malkiel, Y., Pietrangeli, A., and Saporta, S., editors, Issues in Linguistics. Papers in Honor of Henry and Renée Kahane. 719–751. • Wanzare, L. D. A., Zarcone, A., Thater, S., and Pinkal, M. (2016). DeScript: A crowdsourced corpus for the acquisition of high-quality script knowledge. LREC 16. 3494–3501.