Script knowledge constrains ellipses in fragments – Evidence from production data and language modeling

Robin Lemke, Lisa Schäfer, Heiner Drenhaus and Ingo Reich
SFB 1102, Saarland University
robin.lemke@uni-saarland.de

Abstract

We investigate the effect of script-based (Schank and Abelson 1977) extralinguistic context on the omission of words in fragments. Our data elicited with a production task show that predictable words are more often omitted than unpredictable ones, as predicted by the Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007). We take into account effects of linguistic and extralinguistic context on predictability and propose a method for estimating the surprisal of words in presence of ellipsis. Our study extends previous evidence for UID in two ways: First, we show that not only local linguistic context, but also extralinguistic context determines the likelihood of omissions. Second, we find UID effects on the omission of content words.

Background In order to communicate a message, speakers can choose between a full sentence (1a) and nonsentential utterances, or *fragments* (Morgan, 1973) (1b). Fragments can convey the same meaning as the corresponding sentence, but lack words that are obligatory in the sentence, like a finite verb. We investigate why people omit particular words in fragments and hypothesize that the choice between omitting and realizing a word is driven by the Uniform Information Density (UID) hypothesis (Levy and Jaeger, 2007), which has been applied to other omissions, like that of relative pronouns (Levy and Jaeger, 2007) and complementizers (Jaeger, 2010).

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

Uniform Information Density UID states that *information* is best distributed uniformly across the utterance. Following Shannon (1949), the information, or *surprisal* (Hale, 2001), of a word w_i is defined as the negative logarithm of its likelihood to appear in context (2).

(2) $$S(w_i) = -\log_2 p\,(w_i \mid context)$$

Surprisal indexes processing effort (Hale, 2001; Levy, 2008), and a uniform distribution makes the most efficient use of the hearer's limited cognitive resources. Previous research has shown that the optional omission of function words reflects optimization with respect to UID (e.g. Levy and Jaeger, 2007; Jaeger, 2010). Optimization consists in two strategies that contribute to a uniform distribution of information: First, omitting uninformative words avoids inefficient local surprisal minima. Second, words that reduce the surprisal of very informative, i.e. unpredictable, following words are more likely to be inserted. If this reasoning also applies to content words like *pizza* in (2), UID can explain why speakers sometimes use a (specific) fragment rather than a sentence: The fragment is preferred over the sentence if it results from omitting predictable words that are obligatory in the corresponding full sentence.

Materials and method Investigating whether omissions are subject to UID requires (i) a set of linguistic data containing the relevant omissions and (ii) surprisal estimates for both the omitted and realized words in this data set. Given these surprisal estimates, logistic regressions can show whether information-theoretic predictors like surprisal affect the likelihood of a word's omission.

Although the term *context* in (2) in principle comprises both linguistic and extralinguistic context (Levy, 2008), most of the previous information-theoretic studies on omissions (like the ones cited above) estimated the surprisal of words from corpora with *n*-gram language models. Such models take only (part of) the linguistic context of the target word into account. How-

ever, fragments often occur discourse-initially, so that predictability depends on extralinguistic context that cannot be retrieved from text corpora. Therefore we collected a data set of utterances for tightly controlled script knowledge-based contexts (Schank and Abelson, 1977) with a production task. This data set allows to quantify the effect of both extralinguistic and linguistic context.

Subjects read a story like (3) (original materials in German) and produced the utterance that they considered most likely in that context. Since scripts prime upcoming events (see e.g. Delogu et al., 2018), they should raise expectations about what will be said in a script-based situation. For instance, in (3), a request to pour the pasta into the pot or to give the speaker the pasta is probable.

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

In order to use empirically motivated script knowledge representations as stimuli, we based our materials on event chains extracted from DeScript (Wanzare et al., 2016), a crowd-sourced corpus of script knowledge that contains about 100 descriptions of the stereotypical time-course of everyday activities, such as cooking pasta. Following Manshadi et al. (2008), we defined an event as the finite verb and its nominal complement, e.g. `put pot` in (3). After dependency-parsing the corpus (Stanford parser, Klein and Manning (2003)) we extracted these event representations from it. We estimated the likelihood of an event given the previous one with bigram language models trained on the manually preprocessed data for each script with the SRILM toolkit (Stolcke, 2002). We then extracted sequences of three events that were most likely to follow each other and used these event chains to construct our materials. The first sentence in each item introduces the script (cooking pasta), and the next three ones elaborate the event chain (`put pot`, `turn on stove`, `boil water`). For each of 24 items, we collected responses from 100 participants recruited on the crowdsourcing platform Clickworker.

Production data preprocessing As there was a high degree of variation both between scripts and between subjects in the data collected with the production task, we preprocessed the data by manually resolving pronouns and ellipses, lemmatizing the remaining words and finally pooling synonyms to a single lemma. Because we are interested in content words, we removed all function words and adverbials. Removing function words is necessary because e.g. articles and prepositions cannot be freely omitted in standard German (Lemke, 2017; Reich, 2017) and adaptation to UID occurs only "within the bounds defined by grammar" (Jaeger, 2010, 25). Prepositions and distinctive case morphology were annotated on the noun (see (4) for an example), as these features can be important cues towards the meaning intended by the speaker. Adverbials were removed because they can remain implicit in regular sentences and therefore are not involved in the generation of fragments (even though it might be interesting to investigate whether this is subject to UID as well). For the utterance in (4a), preprocessing yields the abstract representation in (4b).

(4) a. Schütte die Nudeln in den Topf!
pour the pasta in the.ACC pot
Pour the pasta into the pot!
b. `pour pasta in.pot`

Investigating the effect of surprisal on omission requires surprisal estimates for both realized and omitted words, therefore we reconstructed all ellipses in the original data. We added those expressions that are minimally required in a full sentence, that is, missing verbs and/or their arguments. This ensures that the outcome of the independent variable, surprisal, is not affected by the dependent variable, omission. The data set for analysis comprises a total of 2.409 sentences consisting in 6.816 primitive expressions ("words" in what follows), 1.052 (15.43%) of these words had been omitted in the original data set.

Surprisal estimation We investigate potential effects of three measures of surprisal: (i) *unigram surprisal*, (ii) *context-dependent surprisal* that takes into account preceding linguistic material within the utterance and (iii) *surprisal reduction*, i.e. how much inserting a word before a target word reduces its surprisal.

We estimate the *unigram surprisal* of each word in the preprocessed data with unigram language models with Good-Turing discount on the preprocessed data that we trained using the SRILM toolkit (Stolcke, 2002). We trained an individual language model on the data for each script sepa-

rately, because this allows to interpret surprisal as conditioned on the script-based situation, i.e. on the extralinguistic context (5):

$$(5) \qquad S(w_i) = -\log_2 p(w_i \mid context_{extraling.}).$$

We use a novel method based on Hale (2001) to estimate *context-dependent surprisal*, that considers preceding words in addition to extralinguistic context. The default method to quantify effects of linguistic context on surprisal are bigram or higher order *n*-gram models. However, training *n*-gram models on elliptical data brings along a circularity issue observed by Levy and Jaeger (2007, 852): If predictable words are omitted more often than unpredictable ones, their corpus frequency is not proportional to their predictability. This problem could be addressed by ellipsis resolution, but training *n*-gram models on the enriched data set is also not realistic. A trigram model trained on the enriched data set estimates the surprisal of `pot` in a fragment `pour pot`, where `pasta` has been omitted from $p(pot \mid pour\ pasta)$. Crucially, this is psychologically implausible, because `pasta` is not included in the actual linguistic context.

Therefore we estimate context-dependent surprisal (and surprisal reduction, see below) with a method based on the approach by Hale (2001). Hale (2001) derives surprisal from the work done by the human parser, that consists in rejecting all parses that are compatible with the input before but not after processing a word. The larger the total probability mass of the rejected parses is, the more informative is a word. This approach requires to know the likelihood of each parse, i.e. each complete structure, which in our case is equivalent to its relative frequency in the enriched data set. Hale (2001) calculates the surprisal of a word w_i as the log ratio between the prefix probability α, i.e. the total probability mass of the parses compatible with an input, before and after processing w_i:

$$(6) \qquad S(w_i) = \log \frac{\alpha_{\text{i-1}}}{\alpha_{\text{i}}}$$

We modify Hale's approach by allowing for arbitrarily many omissions before and after each word in the input string in order to account for the possibility of ellipses when calculating a word's effect on the set of maintained parses and consequently on α_i. For instance, processing `pour` in the fragment `pour pot` rules out all parses that do not contain `pour`. Processing `pot` now excludes all parses that do not contain `pot` somewhere after `pour`, independently of whether there is a word like `pasta` between `pour` and `pot`. Surprisal is calculated as (6) based on the prefix probabilities before and after these processing steps. Our approach circumvents the circularity issue because it relies on nonelliptical representations. It is also psychologically realistic because it quantifies the work done by the parser incrementally.

Predictors	r^2	***t*-value**	***p*-value**
Unigram, context	.65	70.06	$< .001$
Unigram, reduction	.48	37.99	$< .001$
Context, reduction	.62	54.0	$< .001$

Table 1: Correlations between surprisal predictors.

Finally, we calculate *surprisal reduction*, i.e. how much inserting w_i reduces the surprisal of w_{i-1}, for all non-final words. For this purpose, we calculate the ratio between the prefix probability at w_{i+1} if w_i has been realized and the prefix probability at w_{i+1} if w_i has been omitted. In case of the example, how much the surprisal of `pot` is reduced by inserting `pasta` is calculated as (7).

$$(7) \qquad S\ reduction(pot,pasta) = \frac{\alpha_{put\ \ldots\ pot}}{\alpha_{put\ \ldots\ pasta\ \ldots\ pot}}$$

Results We analyzed the data with mixed effects logistic regressions (`lme4`, Bates et al. (2015)) that predict the omission of a word in the enriched data set from the surprisal measures. We first conducted separate analyses of unigram and context-dependent surprisal on the complete data set and then an analysis that considers both unigram surprisal and surprisal reduction for non-final words. In principle it would have been desirable to include all three surprisal measures as predictors in a single regression analysis, but, as table 1 shows, in particular context-dependent surprisal is highly correlated with the other two measures.

The models in the analyses of unigram surprisal[1] and context-dependent surprisal[2] contained by-script random intercepts and slopes for surprisal and by-subject random intercepts. In both analyses there are significant main effects of the respective predictor, that confirm our hypothesis that predictable words are more likely to be omitted. The effect for unigram surprisal ($\chi^2 = 7.39, p < .01$) is stronger than that of context-

[1] Ellipsis ~ UnigramS + (1+UnigramS|Script) + (1|Subj)
[2] Ellipsis ~ ContextS + (1+ContextS|Script) + (1|Subj)

dependent surprisal ($\chi^2 = 4.86, p < .05$).
The analysis that includes surprisal reduction and unigram surprisal[3] was conducted on a subset of the data that contained those non-final words that were not followed by an ellipsis (55.51% of the total data). The final model has random intercepts for subjects and scripts and contains significant main effects of both predictors. The effect of unigram surprisal ($\chi^2 = 10.39, p < .01$) replicates the analysis of the full data set, and the effect of surprisal reduction ($\chi^2 = 27.03, p < .001$) shows that words that reduce the surprisal of the following word more strongly are more likely to be realized. There is no significant interaction between both predictors ($\chi^2 = 0.01, p > .9$).

Discussion Our study confirms the predictions of UID on omissions in fragments: Predictable words are more often omitted in fragments, and words that reduce the surprisal of following ones are more often realized. This extends previous evidence for UID in two ways: First, we find UID effects on the omission of content words. Second, we show that not only local linguistic context, but also extralinguistic context determines the likelihood of omissions. UID however seems not to be the only factor in determining whether fragments are used, as the ratio of fragments varies even between scripts with a similar mean surprisal.

Our study also shows that event probabilities estimated from a corpus of script knowledge provide a reasonable model of extralinguistic context, to which subjects adapt their lingustic behavior. We also propose a method for estimating by-word surprisal in partially elliptical data in a psychologically realistic way. In our study this required a data set that we collected specfically for this purpose and a large amount of manual preprocessing. Future work could show inhowfar our results can be replicated on larger and less constrained data sets when preprocessing steps like reference and ellipsis resolution as well as the standardization of the production data are automatized.

[3] Ellipsis ~ UnigramS * SReduction (1|Script) + (1|Subj)

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Francesca Delogu, Heiner Drenhaus, and Matthew W. Crocker. 2018. On the predictability of event boundaries in discourse: An ERP investigation. *Memory & Cognition*, 46(2):315–325.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL (Vol. 2)*, pages 159–166.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Robin Lemke. 2017. Sentential or not? – An experimental study on the syntax of fragments. In *Proceedings of Linguistic Evidence 2016*. University of Tübingen, online publication system.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger P. Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.

Mehdi Manshadi, Reid Swanson, and Andrew S Gordon. 2008. Learning a Probabilistic Model of Event Sequences from Internet Weblog Stories. In *Proceedings of the Twenty-First International FLAIRS Conference*.

Jerry Morgan. 1973. Sentence fragments and the notion 'sentence'. In Braj B. Kachru, Robert Lees, Yakov Malkiel, Angelina Pietrangeli, and Sol Saporta, editors, *Issues in Linguistics. Papers in Honor of Henry and Renée Kahane*, pages 719–751. University of Illionois Press, Urbana.

Ingo Reich. 2017. On the omission of articles and copulae in German newspaper headlines. *Linguistic Variation*, 17(2):186–204.

Roger Schank and Robert Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Enquiry into Human Knowledge Structures.* Erlbaum, Hillsdale.

Claude E. Shannon. 1949. The mathematical theory of communication. In Claude E. Shannon and Warren Weaver, editors, *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. DeScript: A crowdsourced corpus for the acquisition of high-quality script knowledge. In *Proceedings of LREC 2016*, pages 3494–3501, Portoroz, Slovenia.