Optimal encoding! – Information Theory constrains article omission in newspaper headlines*

Robin Lemke, Eva Horch, Ingo Reich
Universität des Saarlandes
Postbox 15 11 50
D-66041 Saarbrücken, Germany
robin.lemke@uni-saarland.de
{e.horch,i.reich}@mx.uni-saarland.de

Abstract

In this paper we argue that the distribution of article omission in newspaper headlines is constrained by information-theoretical principles (Shannon 1948). To this effect, we present corpus data and results from an acceptability rating study. Both point in the same direction: In our corpus, articles are significantly more frequent, when they precede a less predictable head noun. And subjects perceive article omission as more acceptable, if the head noun is (comparably) more predictable. This is in line with the information-theoretical prediction that article omission should be preferred over the overt realization of an article (provided that article omission is grammatical in the first place), if the head noun is comparably predictable in its local context.

1 Introduction

Functional deletion, that is the non-realization of, for example, complementizers (1), or articles (2), is a frequent phenomenon across text types.

(1) My boss thinks (that) I'm absolutely crazy. (Jaeger 2010:31)

(2) Gündogan set to miss ∅ rest of ∅ season with ∅ cruciate injury. (guardian.co.uk, 16.12.2016)

As the brackets in example (1) indicate, functional deletion is typically optional. However, if it is in fact an optional process (in a given genre), this raises the question why functional expressions are overtly realized in some cases, but not in others. In this paper, we want to argue that Information Theory is at least part of the story. This has already been shown in Jaeger (2010) with respect to the phenomenon of complementizer deletion, and we would like to add further evidence in support of this hypothesis from article omission.

In contrast to standard written German, see (4), newspaper headlines in German (and many other languages) allow for bare singular noun phrases (NPs), see for example the headline in (3) from the online newspaper *Zeit.de* (2016/12/01); for a more thorough overview over the phenomenon, see e.g. Sandig (1971), Stowell (1996), or Reich *in press* as well as the references cited therein.

(3) ∅ Niederlage für die ganze Gesellschaft
∅ defeat for the whole society

(4) Er berichtet von *(einer) Niederlage für die ganze Gesellschaft
he reports of *(a) defeat for the whole society

Like complementizer deletion, article omission in newspaper headlines is an optional process. Both the attested *Niederlage für die ganze Gesellschaft* and the constructed *Eine Niederlage für die ganze Gesellschaft* are, at least in principle, grammatical / acceptable newspaper headlines in German.

Previous research on article omission focused on specific structural constraints (e.g. to account for the structural asymmetry in article omission observed in Stowell 1996), and on specific constructions (like article omission in the complement of a preposition; see Kiss 2010), but less so on the question why in a given utterance token in a specific context an article is or is not realized. A notable exception is the work by De Lange and colleagues (see for example De Lange 2008, De Lange et al. 2009). De Lange and colleagues, however, investigate article omission in newspaper headlines primarily from a typological perspective and relate omission frequencies (on the

* We would like to thank four anonymous reviewers for valuable comments and suggestions. All remaining errors are, of course, ours.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 131–135,
Valencia, Spain, April 3-7, 2017.

basis of Information Theory) to the overall complexity of the respective article systems along the following lines: The more complex an article system is, the less predictable is a given article (like German *der*, *die* or *das*, for example); and the less predictable a given article is, the more pressure there is to overtly realize the article. Like De Lange and colleagues, we will also argue in the following that information-theoretical considerations are relevant in the description and analysis of article omission. In contrast to De Lange and colleagues, however, we consider article omission as a function of the predictability of the following head noun in a given local linguistic context (rather than as a function of the predictability of an article relative to a given article system).

2 Background: Information Theory and functional deletion

Information Theory relies on a probabilistic notion of information, whereby the amount of information conveyed by some unit is derived from its probability to occur given the previous context. Applied to sentence comprehension, the information, or surprisal (Hale 2001), of a word α in a given context c is calculated as the negative logarithm of the probability of α in c, in short $Surprisal(\alpha) = -log_2 P(\alpha|c)$. Hence, highly predictable words are less informative while highly unpredictable words are more informative. Communication is modeled as occurring through a noisy channel with limited capacity, which speakers should approximate in order to communicate efficiently. Exceedance of channel capacity is to be avoided and penalized with additional processing load. Consequently, speakers tend to distribute information uniformly across an utterance at a transmission rate close to channel capacity. This is argued for by Aylett & Turk (2004), De Lange et al. (2009), Genzel & Charniak (2002), Levy & Jaeger (2007), among others. In Jaeger (2010) the principle guiding the speaker in choosing between grammatical alternatives is called the *Uniform Information Density Hypothesis* (UID):

> **Uniform Information Density (UID)**
> Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).
>
> (Jaeger 2010: 25)

To get an idea of how the UID might relate to article omission, consider figure 1. Figure 1 illustrates the surprisal profiles of three different encodings of one and the same message (that tomorrow the judge pronounces the sentence). These encodings only differ in the (non-)realization of the relevant articles. As is apparent from the surprisal profiles, the low surprisal values of the articles *der* and *das* create substantial troughs. As a result, the surprisal profile of the encoding with overt articles is significantly less uniform than the surprisal profile without articles. The UID thus predicts that, other things being equal, the latter encoding should be preferred over the former encoding.

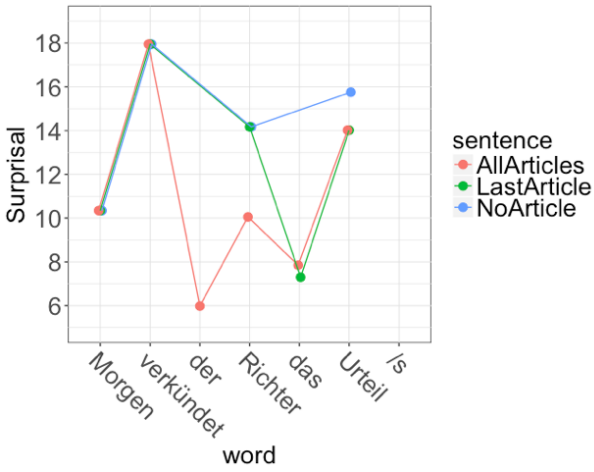

Figure 1: The surprisal profile of the headline *Morgen verkündet der Richter das Urteil*is more uniform in case of article omission across the board (based on trigrams calculated on the FraC corpus)

Jaeger (2010) argues, based on a corpus study, that the UID constrains the distribution of complementizer deletion in English. He shows that the insertion of a complementizer systematically reduces the surprisal on the following word(s). Thus, if the occurrence of a complement clause is highly unpredictable, the insertion of a complementizer might lead to a more uniform surprisal profile by significantly reducing the high surprisal of the word(s) to follow. On the other hand, if a complement clause is highly predictable and its onset less informative, dropping the complementizer might be the better option with respect to the UID.

A similar reasoning could apply to article omission: Again, speakers have to choose between grammatical alternatives which convey essentially the same proposition, which however differ in the way they distribute the relevant information across the utterance. Horch & Reich (2016) argue, based on language models trained on POS tags, that the insertion of an article systematically lowers the surprisal of the following noun. Now, given the results in Jaeger (2010), it seems straightforward to assume that speakers also exploit this kind of variation in order to optimize the surprisal profiles of their utterances. Specifically, speakers are expected to prefer overt articles if they precede nouns with rather high surprisal, and to prefer article omission, if they precede nouns with rather low surprisal (in order to raise the surprisal on the noun and to distribute the information encoded more uniformly across the utterance).

3 Corpus study

If speakers (and writers) try to optimize their utterances w.r.t. information-theoretic constraints, this should be reflected in production preferences and therefore in corpora of text types which allow for the respective omissions. However, accurately finding all instances of article omission is not a trivial issue, as there are several special cases of singular nouns which allow for or even require article omission even in standard written German, e.g. predicative (5a) or mass nouns (5b). The distinction between those cases and "genuine" cases of article omission thus requires a corpus, in which the relevant cases are explicitly annotated.

(5) a. Ich bin (ein) Student.
I am a student.
I am a student.

b. Wir brauchen noch (*ein/#das)
We need still a/the
Mehl.
flour
We still need flour.

Therefore, we tested our hypothesis on the FraC corpus (Horch 2016), which is text type-balanced and has been annotated by hand for different types of ellipses. Omitted articles are annotated with a placeholder `NoArt` in the corpus. The corpus contains about 17 different text types (2.000 sentences each) ranging from prototypically written (e.g. newspaper articles) to prototypically spoken (e.g. dialogues) text types.

We pre-processed the corpus by removing all articles and lemmatizing it. Then we computed each word's surprisal by training a bigram language model using Kneser-Ney smoothing (Kneser & Ney 1995) in an interpolated backing-off scheme (Katz 1987) with the SRILM toolkit (SRI International). Bigram surprisal was chosen in order to obtain a sensible measure given the small size of the corpus.

For reasons of comparison, we restricted our investigation to noun phrases that immediately follow a finite verb. The (bigram) surprisal of a noun is then equivalent to $-log_2 p(noun|verb)$. Due to the elimination of the articles from the training set, this figure only reflects the subcategorization preferences of the verb lemma in question and is not affected by the occurrence of an article in the original corpus. We take this to be a psychologically sensible measure of noun informativity.

For the analysis, we extracted all 131 postverbal nouns from the corpus. 50 of these are headed by an overt article, while the remaining 81 are not. The histogram in figure 2 shows the distribution of article omission across surprisal values and indicates that article omission is preferred more strongly for less informative nouns. We analyzed the data with a mixed effects logistic regression with random intercepts for noun lemmata and verb lemmata using the `lme4` (Bates et al., 2015) package in `R` (R Core Team, 2016). The integration of random slopes into the model were not appropriate due to the small size of the data set. A likelihood ratio test computed with the `anova` function in `R` shows that the model containing SURPRISAL as main effect fits significantly better to the data than a baseline model with random effects and the intercept only ($\chi^2 = 9.7, p < 0.01$). The main effect of SURPRISAL indicates that, as predicted by the UID, article omission is more likely the less informative the corresponding noun is.

4 Experimental study

The corpus study provides first support for our hypothesis, but the amount of appropriate data in the FraC headlines is rather small in absolute terms. It would be desirable to test the validity of the hypothesis on a larger corpus, but this is complicated by the reasons discussed in the previous section.

If speakers have a general preference for encodings conforming to UID though, these are proba-

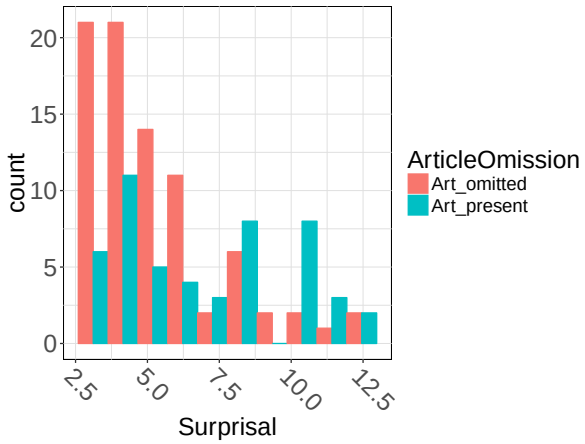

Figure 2: Histogram of NPs with and without overt articles in the headlines in FraC.

bly not only reflected in their production choices but also in the perception of well-formedness. We therefore shifted towards investigating our hypothesis with an acceptability rating study, which compared the acceptability of ARTICLEOMISSION as a function of SURPRISAL of a postverbal noun in constructed newspaper headlines a 2 × 2 design.

In order to obtain verb subcategorization preferences from a larger corpus, in this occasion we used the German Reference Corpus DeReKo (Kupietz & Keibel 2009), which contains mostly written text of different text types, e.g. scientific literature, fiction and newspaper articles. The corpus is accessible and searchable with the COSMAS II web interface, which we used to extract around 3.1 M instances of immediately postverbal nouns from the corpus. By "immediately postverbal" we understand such nouns that are at most separated by an article and/or one adjective from the preceding verb. The data set was pre-processed by removing all intervening articles and adjectives between noun and verb and lemmatized. After that, we computed surprisal as $Surprisal = -log_2 p(noun|verb)$. Our measure of surprisal is hence identical to the one used in the corpus study and reflects the subcategorization preference of the verb.

A sample item is given in (6). We constructed versions of the items with and without article omission and with a low (*Projekt* in (6)) and a highly informative noun (*Klage*), yielding 4 conditions. While surprisal was treated as a binary variable for distributing the materials across subjects, in the statistical analysis it was a numeric predictor in order to account for relative differences between more and less informative nouns.

(6) *Papst Franziskus unterstützt (das|∅)*
pope Francis supports (das|∅)
(Projekt|Klage) gegen Kinderarbeit.
(project|claim) against child.labor
'Pope Francis supports the project/claim against child labor.'

74 subjects rated 28 items (7 per condition) which were mixed with 92 unrelated fillers (constructed headlines as well) in a web-based questionnaire on a 7-point Likert scale. Subjects participated in a lottery of 10 × 30 euros as a reward. The rolling averages plot in figure 3 provides an overview of the distribution of ratings across the range of surprisal values tested and indicates that article omission is preferred for uninformative nouns.

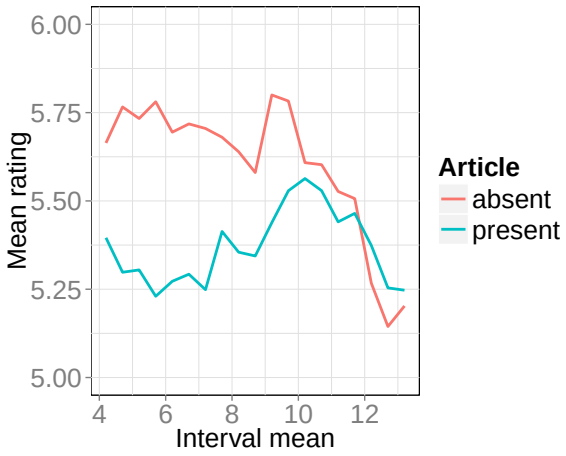

Figure 3: Rolling averages plot for the rating data. The plot shows mean ratings for all items contained in an interval of size 3, whose mean is displayed on the x-axis of the plot. For instance, the value at $x = 6$ is equivalent to the mean rating of all items ranging from a noun surprisal of 4.5 to 7.4. This smoothing technique allows to observe a general trend by averaging over individual values.

We analyzed the data with Cumulative Link Mixed Models for ordinal data with the `ordinal` package in R (Christensen, 2015). Besides a general preference for article omission across our items in fillers which is in line with the preference for article omission in the postverbal NPs in the corpus and is thus not of theoretic interest to us on itself, there is a significant interaction ($z = 2.9$, $p < 0.01$) between ARTICLEOMISSION and NOUNPREDICTABILITY indicating that article omission is specifically preferred for low

informative nouns, while the difference between conditions vanishes for informative nouns. This indicates that the article is specifically redundant in the context of uninformative nouns.

5 Discussion and outlook

Starting from the observation that the insertion of articles lowers the surprisal of the following noun (Horch & Reich 2016), we investigated in this paper whether article omission is the more preferred the less informative the following head noun is, as predicted by Information Theory. We modeled the linguistic context by falling back on the subcategorization preferences of verbs and confirmed our hypothesis with both a corpus study on article omission in German newspaper headlines and an acceptability rating study. The rating study suggests that subjects are in fact aware of the subtle and gradient contrasts in terms of information density and indicates that their preferences mirror the corpus data. Our results are thus in line with Jaeger's (2010) study on complementizer deletion and provide further evidence for the usefulness of applying information-theoretical concepts to the analysis of natural language.

It would be desirable, of course, to confirm these results with larger corpora and for a larger variety of contexts. This, however, requires high quality automatic annotation of article omissions in large-scale corpora, which is to the best of our knowledge currently not yet available.

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.

R. H. B. Christensen. 2015. ordinal—regression models for ordinal data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

Eva Horch and Ingo Reich. 2016. On 'Article Omission' in German and the 'Uniform Information Density Hypothesis'. In Stefanie Dipper, Felix Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13thConference on Natural Language Processing (KONVENS)*.

Eva Horch. 2016. Article missing? Talk at the 38th DGfS annual meeting, Konstanz.

SRI International. SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/.

T. Florian Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, and T. Hoffman, editors, *Advances in neural information processing systems*, pages 849–856. MIT Press.

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

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on Acoustics, Speech, and Signal Processing*, ASP-35(3).

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *IEEE transactions on Acoustics, Speech, and Signal Processing*.

Marc Kupietz and Holger Keibel. 2009. The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. pages 53–59.

Joke De Lange, Nada Vasic, and Sergey Avrutin. 2009. Reading between the (head)lines: A processing account of article omission in newspaper headlines and child speech. *Lingua*, 119:1523–1540.

Joke De Lange. 2008. *Article omission in child speech and headlines: a processing account*. Ph.D. thesis, Utrecht University, Utrecht.

R Core Team, 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ingo Reich. in press. On the omission of articles and copulae in German newspaper headlines. In D. Massam and Tim Stowell, editors, *Register variation and syntactic theory. Special issue of Linguistic Variation*. Benjamins.

Barbara Sandig. 1971. Syntaktische Typologie der Schlagzeile. In *Linguistische Reihe*, volume 6. Hueber Verlag, Ismaning.

Claude Shannon. 1948. A mathematical theory of communications. *The Bell System Technical Journal*, 27:379–423.

Tim Stowell. 1996. Empty heads in abbreviated english. In *GLOW 1991 (revised 1996)*. de Gruyter.