Lisa Schäfer*

Topic drop in German: Empirical support for an information-theoretic account to a long-known omission phenomenon

Abstract: German allows for *topic drop* (Fries 1988), the omission of a preverbal constituent from a V2 sentence. I address the underexplored question of why speakers use topic drop with a corpus study and two acceptability rating studies. I propose an information-theoretic explanation based on the Uniform Information Density hypothesis (Levy and Jaeger 2007) that accounts for the full picture of data. The information-theoretic approach predicts that topic drop is more felicitous when the omitted constituent is predictable in context and easy to recover. This leads to a more optimal use of the hearer's processing capacities. The corpus study on the FraC corpus (Horch and Reich 2017) shows that grammatical person, verb probability and verbal inflection impact the frequency of topic drop. The two rating experiments indicate that these differences in frequency are also reflected in acceptability and additionally evidence an impact of topicality on topic drop. Taken together my studies constitute the first systematic empirical investigation of previously only sparsely researched observations from the literature. My information-theoretic account provides a unifying explanation of these isolated observations and is also able to account for the effect of verb probability that I find in my corpus study.

Keywords: topic drop, ellipsis, information theory, Uniform Information Density hypothesis, predictability, recoverability, corpus study, acceptability judgments

1 Introduction

It is known since at least Reis (1982) and Ross (1982) that German allows for the omission of a preverbal constituent from a declarative verb-second (V2) sentence. A constituent from the so-called *Vorfeld* ('prefield') in terms of the topological field model (Drach 1937) is left out, so that the sentence superficially starts with the fi-

*Corresponding author: Lisa Schäfer,** Department of Modern German Linguistics, Saarland University, Saarbrücken, Germany, e-mail: lisa.schaefer@uni-saarland.de

nite verb in the *linke Satzklammer* ('left bracket'). This phenomenon is usually referred to as *topic drop* but it is also known under other terms such as *pro zap* (Ross 1982), *null topic* (Fries 1988; Cardinaletti 1990), *uneigentliche Verbspitzenstellung* ('improper verb first clause') (Auer 1993; Imo 2014), *Vorfeld-Analepse* (Zifonun et al. 1997) or *Vorfeld-Ellipse* ('prefield ellipsis') (Frick 2017). Topic drop is illustrated in (1), an example taken from the *FraC* fragment corpus (Horch and Reich 2017). The 1st person singular subject pronoun *ich* 'I' has been omitted from the prefield (indicated by Δ), leaving the main verb *kann* 'can' in the sentence-initial position.

(1)   Δ *Kann die Verabredung leider   nicht einhalten*
       Δ can   the appointment unfortunately not   keep
       'Unfortunately, (I) cannot keep the appointment'
       (FraC S1127)

In a first detailed analysis of the phenomenon, Fries (1988) notes its apparent register dependency: Topic drop appears exclusively or at least preferably in spoken language (see Auer 1993; DUDEN 2016) and in conceptually spoken (in the terminology of Koch and Oesterreicher 1985) text types such as telegrams (see Reis 1982; Barton 1998, cf. Frick 2017), personal letters, diaries and certain literary texts (Fries 1988).[1]

     There has been a variety of theoretical work on topic drop over the last 30 years (e. g. Huang 1984; Fries 1988; Cardinaletti 1990; Zifonun et al. 1997; Trutkowski 2016) but this work has mainly focused on the licensing and the grammatical properties of this phenomenon and how to model them adequately. What previous research has hardly investigated is the question of when topic drop is actually *used* (but see e. g. the interactional linguistic approach by Helmer 2016). I aim at answering this question and explore speaker and hearer preferences beyond the grammatical properties that license topic drop. This article makes two main contributions: Firstly, I provide the first systematic empirical investigation of claims from the theoretical literature according to which grammatical person, verbal inflection and topicality influence the usage of topic drop. Secondly, I propose an information-theoretic account of the usage of topic drop which is based on the Uniform Information Density (UID) hypothesis (Levy and Jaeger 2007). My account provides a unifying explanation to previously isolated findings and can

---

1 The restriction to informal registers has also been observed for omission phenomena in other languages: According to Haegeman (1997: 233), English allows for subject omission in colloquial speech and, like French, in what she calls "abbreviated written registers" such as diaries and instructions. Similarly, subject omission in Russian mainly occurs in informal registers and colloquial speech (Zdorenko 2010).

additionally account for an effect of the matrix verb's frequency-based probability on the usage of topic drop that I find in a corpus study.

In my empirical investigations I systematically and jointly test central aspects of the usage of topic drop according to the previous literature. First, I focus on the grammatical person and review previous corpus linguistic studies that report a preference for topic drop of the 1st person singular subject pronoun *ich* 'I' over the other grammatical persons. I discuss two explanations provided in the theoretical literature: an *inflectional hypothesis* by Auer (1993) and a *pragmatic hypothesis* by Imo (2014). Second, as a testing ground to distinguish between the two accounts I explore verbal inflection. As a third factor I consider the effect of topicality on topic drop which is debated controversially in the theoretical literature. I propose a unifying information-theoretic account that incorporates these isolated claims and is based on predictability and recoverability. Following this account, I suggest the probability of the verb (termed *verb surprisal*) as a predictor whose effect only UID can explain. I test the predictions from the theoretical literature and of my information-theoretic account in three empirical studies. In a corpus study I investigate the frequency of topic drop in the text message subcorpus of FraC considering effects of grammatical person, verbal inflection and of the probability of the verb in the left bracket. In two rating studies I systematically investigate the role of topicality and again take into account grammatical person and verbal inflection.

This article is structured as follows: In Section 2, I discuss previous research on the usage of topic drop with respect to the three central factors grammatical person, verbal inflection and topicality. Section 3 presents my unifying information-theoretic account based on UID and its predictions with respect to the usage of topic drop. In Section 4, I present a corpus study that evidences in line with the literature and with previous corpus studies that topic drop of the 1st person singular is more frequent than of the 3rd person singular. Additionally, it shows that topic drop is less frequent before unpredictable verbs, but that a distinct inflectional marking leads to more topic drop even before unpredictable verbs which provides genuine support for the information-theoretic account. The two acceptability rating studies are presented in Section 5. They show in line with the pragmatic hypothesis and the information-theoretic account but against the inflectional hypothesis that topic drop of the 1st person singular subject pronoun is also more acceptable regardless of whether the following verb has a distinct inflectional ending or not. Furthermore, they suggest that clear inflectional marking and topic continuity together improve the ratings for topic drop. In Section 6, I summarize the results of the three studies and show that the information-theoretic approach is suitable to explain the data on the usage of topic drop in a uniform way.

2 Previous evidence on why topic drop is used

In this section, I focus on three central factors that impact topic drop according to previous literature: First, I present three corpus linguistic studies that have evidenced an influence of grammatical person: Topic drop of the 1st person singular subject pronoun seems to be particularly frequent. Second, I discuss distinct inflectional marking on the verb as a possibility to distinguish between the predictions of two hypotheses that aim at explaining the prevalence of topic drop of the 1st person singular. Third, I review a debate in the literature on the role of topicality for topic drop, i. e. whether the topic status of a constituent is a prerequisite for its omission as the name of the phenomenon suggests.

2.1 Grammatical person

Previous literature has attested an influence of grammatical person on topic drop in such a way that the 1st person singular subject pronoun *ich* 'I' is particularly often omitted. Without providing numbers, Auer (1993) reports that in his corpus of spoken German the subject and object pronoun *das* 'that' is dropped most frequently, but he also states that *ich* is frequently omitted.[2] More recently, Androutsopoulos and Schmidt (2002) analyzed a corpus of 934 text messages and found an omission rate of 60 % for the 1st person singular subject pronoun, followed by one of 51 % for the 3rd person singular subject pronouns (the majority instances of *das* 'that' and *es* 'it') and of 30 % for the 1st person plural subject pronoun. The 2nd person singular pronouns had an omission rate of 26 %, whereas the 2nd and 3rd person plural were never omitted. These results are in line with Frick's (2017) corpus study on 3999 Swiss German text messages: She found that of all text messages with 1st person singular subject pronouns again about 60 % were elliptical, followed by 53 % omissions of 2nd person singular subject pronouns, 41 % of 3rd person singular subject pronouns and around 20 % of the plural subject pronouns. In both studies *ich* is more often omitted from the prefield of text messages than it is realized and its omission rate is the highest.

　　In order to account for this observation, Auer (1993: 198) proposes what I will call the *inflectional hypothesis*: He suggests that the omission of the 1st person singular subject pronoun is easily possible because, if the subject pronoun is left

---

2 The study by Auer (1993) is rather exploratory in nature. He does neither characterize his data in detail – he only states that he has extracted about 100 instances of proper and improper verb first clauses from a corpus of spoken conversations (Auer 1993: 195) – nor does he present frequency tables, omission rates or a statistical analysis of his results.

out, the German verbal morphology in present tense singular is still sufficiently differentiated to express the grammatical person only by inflection. However, this argument is limited to the present tense of weak and strong verbs. In the preterite and for preterite present verbs even in present tense the forms for the 1st and 3rd person singular are syncretic (e. g. *ich spielte* 'I played' vs. *sie spielte* 'she played' and *ich weiß* 'I know' vs. *sie weiß* 'she knows'). Hence, the verb form that agrees with the 1st person singular is not necessarily distinctly marked so that the preference for topic drop of the 1st person singular cannot be explained by recoverability through distinct inflectional marking alone.

Imo (2014) borrows the inflectional hypothesis from Auer (1993) but adds a further explanation that I name *pragmatic hypothesis*: He states that the omitted 1st person singular subject pronoun is easy to process because "the default 'origo' of speaking, i. e. 'I-here-now', can be activated in most cases so that the recipients can assume that the 'missing' element is the unmarked 'I'" (Imo 2014: 153–154). Hence, two factors seem to facilitate the identification of the omitted constituent which increases the likelihood that a speaker leaves out a constituent and that the hearer may successfully recover it: the distinct verbal morphology that Auer (1993) proposes in the inflectional hypothesis and the easier recoverability of the pronoun referring to the situationally prominent speaker that Imo (2014) suggests in the pragmatic hypothesis. Although Imo (2014) bases his argumentation on both hypotheses simultaneously, I will consider them separately in this paper because they partly make opposing predictions as the next section shows.

The empirical studies in this paper first investigate whether the corpus linguistic findings on grammatical person can be replicated and extended using the FraC corpus and second whether they are reflected in acceptability preferences.

2.2 Verbal inflection

In order to test Auer's (1993) inflectional hypothesis I look at the impact of verbal inflection. I compare topic drop of subjects before full verbs with a distinct inflectional marking in present tense to topic drop before verbs that have syncretic forms for the 1st and the 3rd person singular like the preterite present modal verbs (e. g. *ich kann* 'I can' vs. *sie kann* 'she can') and full verbs in preterite using corpus frequencies and acceptability judgments. As the syncretic verb forms do not allow to distinguish between the 1st and the 3rd person singular, the inflectional but not the pragmatic hypothesis predicts that topic drop of a 1st person singular subject pronoun has no longer an advantage over topic drop of the 3rd person singular when the inflectional marking is not distinct. The inflectional hypothesis furthermore clashes with the claim by Zifonun et al. (1997) that topic drop is

actually preferred with modals and auxiliaries compared to full verbs. This claim is supported by corpus linguistic evidence: Androutsopoulos and Schmidt (2002) report that the most frequent verbs that appear with topic drop are the auxiliaries *sein* 'to be' and *haben* 'to have' and the modals *wollen* 'want', *müssen* 'must' and *können* 'can'. Modal verbs exhibit the highest omission rate of more than 70 %, for the auxiliaries the rate is about 60 % and for the most frequent full verbs *gehen* 'to go', *kommen* 'to come' and *sitzen* 'to sit' only at about 40 %. This tendency is also present in Frick's (2017) corpus study where topic drop is relatively more frequent with copulae and modal verbs than with full verbs. The fact that in both studies topic drop is more frequent with modal verbs that do not have a distinct inflectional ending for the 1st and the 3rd person singular in present tense than with most full verbs already strongly questions the inflectional hypothesis by Auer (1993). In my empirical studies I investigate whether there are differences in frequency and acceptability of topic drop depending on whether the following verb is distinctly marked for inflection.

2.3 Topicality

Previous research widely agrees that contextual salience is a prerequisite of topic drop (Fries 1988; Trutkowski 2016; Reich 2018), as Cardinaletti (1990: 75) puts it, "the reference of the null argument must be recoverable either from the linguistic or extra-linguistic context". However, it is still a matter of debate whether salience necessarily coincides with the topic status of the constituent. By topic – which is a rather vague information-structural term (see Musan 2002) – I understand following Reinhart's (1981) influential definition "the expression whose referent the sentence is about". In the literature, there is not only no agreement on the role of topicality for topic drop but in most of the cases, there has not even been a clear positioning of the authors. Nevertheless, since many of them use the term topic drop and notions like "topic position" (Huang 1984; Auer 1993)[3] or determine the expression targeted by topic drop as "most thematic" (Oppenrieder 1987; Zifonun

---

3 This suggests that the prefield is a genuine topic position in German. However, topics may also be placed in different positions (Molnár 1998; Jacobs 1999; Frey 2000). Frey (2000) proposes a special topic position in the middle field and argues that the prefield also allows for elements that may not be topics, such as expletives. Hence, the theoretical literature suggests that the prefield is neither the only position where topics may be placed in German nor may only topics be placed in this position. Speyer (2010) presents an optimality-theoretic model of how the prefield is filled according to which phrases that serve the purpose of scene-setting or contrast are ranked higher than those that represent the topic.

et al. 1997) or "non-rhematic" (Fries 1988), this suggests that they at least implicitly share the view of Sternefeld (1985: 407) and Helmer (2016: 25) who hypothesize that only topics may be omitted. In contrast, there are also authors who postulate a purely structural account to topic drop: Trutkowski (2016) presents introspective counter-examples like (2) which show that topic drop is not restricted to topics but can also target semantically empty elements. Frick (2017) supports this view with corpus examples from Swiss German text messages like (3) for different types of *es* expletives. She even explicitly states that the term *topic* in topic drop should not be confounded with the information-structural concept topic (Frick 2017: 67).

(2)   Δ *Regnet grad'.*
        Δ rains   now
        (Trutkowski 2016: 120)

(3)   *[...] Δ hät ales scho   zue.*
        [...] Δ has all   already closed
        (Frick 2017: 146)

The fact that semantically empty elements may be omitted from the prefield position clearly questions the assumption that topicality is a necessary prerequisite for topic drop. My investigations will show that this assumption is doubtful even for referential expressions which can be topicalized.

3 Predictability and recoverability – towards an information-theoretic account of topic drop

The review of central aspects on topic drop in the theoretical literature and the controversies emerging from the diverging positions already clearly illustrated that there is not yet a unifying account of the usage of topic drop. For instance, the inflectional and the pragmatic hypotheses are just isolated claims that aim at explaining the prevalence of topic drop of the 1st person singular. However, they do not explain why speakers use topic drop at all given that the corresponding full forms would also be available to them. In this paper, I propose that the choice between topic drop and the corresponding full form depends on both predictability and recoverability of the potentially omitted preverbal constituent: A prefield constituent is more likely to be omitted if it is predictable given the preceding context and / or if it can be easily recovered given the subsequent verb. I model this idea by means of an information-theoretic account based on the Uniform Information Density (UID) hypothesis (Levy and Jaeger 2007) which has been successfully employed to account for a variety of omission phenomena (e. g. Levy and Jaeger 2007; Jaeger 2010; Kravtchenko 2014; Lemke et al. 2017).

3.1 The Uniform Information Density (UID) hypothesis

In information theory, the *information* of a word is defined as the negative binary logarithm of its conditional probability given context, i. e. $-log_2\ p(word\ |\ context)$ (Shannon 1948). Following Shannon (1948), communication is modeled as the transmission of information from a sender to a receiver over a channel. This channel has a limited capacity, i. e. there is an upper bound to the amount of information that can be successfully transmitted and it is most efficient to send at a rate close to but not exceeding this channel capacity. Psycholinguistic research has established that information, also termed *surprisal*, indexes processing effort (Hale 2001; see also Levy 2008; Demberg and Keller 2008). The channel capacity can consequently be interpreted as an upper bound to the processing capacities of the hearer. The central idea of UID is that successful communication consists in distributing surprisal uniformly across the utterance avoiding minima and maxima in the information density profile. UID predicts that from a set of alternative *grammatical* encodings of a message speakers choose the encoding that conforms best to this principle (Jaeger 2010: 25). This entails that topic drop is only an alternative encoding when it is grammatical, i. e. when it is licensed.

Speakers can optimize their utterance with respect to UID in two ways. First, they omit predictable words which have low surprisal. Since such words would cause undesirable minima in the information density profile, omitting them makes a more efficient use of the processing resources of the hearer. Second, speakers smooth surprisal maxima by inserting words before very unpredictable words that are hard to process. If the inserted words increase the likelihood of the unpredictable words, this reduces the processing effort of the latter.

With respect to topic drop, UID makes two predictions: First, topic drop should be the more preferred, the more predictable the omitted expression is given previous linguistic or extra-linguistic context because this avoids surprisal minima. Second, the full form should be preferred over topic drop when the insertion of the prefield constituent reduces high processing effort associated with the following verb because this prevents a surprisal maximum. In this case the context *following* the prefield constituent impacts its omission.

3.2 Avoid surprisal minima

According to the first prediction, a surprisal minimum, which corresponds to an inefficient use of the hearer's processing resources, is caused by a highly predictable expression and can be avoided by omitting this expression. Making use
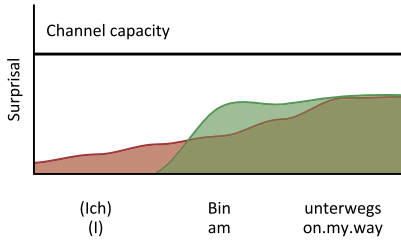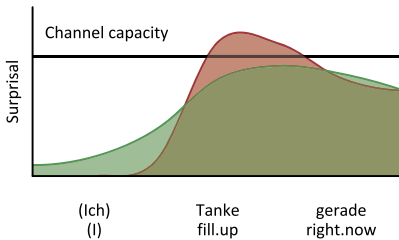
Figure 1: Hypothetical ID profile: *ich* creates a surprisal minimum in the full form (red). The utterance with topic drop is more uniform (green).

of topic drop results in a more uniform information density profile. This idea is illustrated graphically in Figure 1.[4] On the y-axis the plot shows hypothetical surprisal values for the words of the utterance *(Ich) bin unterwegs* '(I) am on my way' produced as the answer to the question 'Hey, where are you?'. In the full form the 1st person singular pronoun *ich* that refers to the speaker is very predictable because the speaker is both linguistically and extra-linguistically prominent. *Ich* hence creates a surprisal minimum as indicated by the red curve, i. e. the surprisal of *ich* is very far below the (hypothetical) channel capacity. Omitting *ich*, i. e. using topic drop as depicted in the green curve, thus leads to a more efficient distribution of surprisal across the utterance without exceeding the hearer's processing capacities.

The tendency to avoid surprisal minima can straightforwardly capture the isolated observations on topic drop from the previous literature that I discussed in Section 2: If my empirical studies confirm an effect of grammatical person, i. e. that the 1st person singular referring to the speaker is indeed more likely in a speech situation than the 3rd person singular, this result can be explained by UID: The redundant pronoun *ich* causes a trough in the information density profile that can be smoothed by omitting the pronoun. This also captures the prediction of the pragmatic hypothesis because when the speaker is part of the default origo of speaking, she or he is very predictable. A similar line of reasoning can be employed to explain a potential effect of topicality. UID predicts that the omission of a topic is more felicitous provided that the topic is more likely to be talked about than other referents. Therefore the topic is more predictable, i. e. less informative, and thus more likely to be omitted in order to avoid a surprisal minimum.

---

4 All graphics in this paper except for the screenshot in Figure 6 were produced in R (R Core Team 2018) using the package ggplot2 (Wickham 2016).

Figure 2: Hypothetical ID profile: *tanke* creates a surprisal maximum in the utterance with topic drop (red). The full form (green) is more uniform.

3.3 Avoid surprisal maxima

If my empirical investigations evidence that the usage of topic drop is constrained by the tendency to avoid surprisal maxima, this would provide additional evidence and explanatory power for an information-theoretic account. Topic drop should be less felicitous when a verb with high surprisal is left in the sentence-initial position as illustrated in Figure 2: If a speaker uttered the topic drop *Tanke gerade* 'Fill up right now' instead of *Bin unterwegs* and under the premise that *tanken* 'to fill up' is less predictable than the auxiliary *sein* 'to be', this would lead to a maximum in the information density profile as indicated by the red curve, i. e. a region that causes high processing effort. In this situation, realizing the preverbal constituent *ich*, i. e. not using topic drop, would lead to a more uniform information density profile. Provided that inserting *ich* increases the likelihood of *tanke*, the pronoun smooths the surprisal maximum on the verb with surprisal close to but not exceeding channel capacity as shown by the green curve. Following from this second prediction, I expect that the surprisal of the verb following the prefield constituent predicts topic drop. If I find an effect of this predictor, this could provide genuine evidence for UID.

    The strategy to avoid surprisal minima is determined by predictability, i. e. the preverbal constituent is predictable given some preceding linguistic or extralinguistic context. In contrast, the processing effort on the verb is influenced not only by the predictability of the verb itself but also by resolving ellipsis. Since ellipsis can only be resolved *after* the material following the ellipsis site has been encountered, when the hearer notices that something has been omitted, recoverability needs to be taken into account. Assuming an incremental parser that uses any incoming information immediately for a parsing decision (Marslen-Wilson 1973, 1975; Altmann and Kamide 1999), it is reasonable to assume that topic drop is resolved immediately on the following verb. For instance, a distinct verbal inflection on the verb is a cue that is used to recover the omitted preverbal constituent. The

more difficult this recovery is, the higher is the additional processing load on the verb. My information-theoretic account hence predicts that topic drop of a subject is more felicitous before a verb with distinct inflectional marking. In German, the distinct inflection of a verb indicates the grammatical person of the subject. If the subject is omitted, a distinct inflection can be a cue towards recovering the omitted constituent and can reduce the processing effort of resolving ellipsis which prevents a surprisal maximum on the verb. This way a distinct verbal inflection may in particular help to process verbs with a high surprisal as it reduces the overall processing load on the verb and therefore reduces the probability of a peak in the information density profile that exceeds channel capacity.

In the corpus study and the experiments I will empirically test the claims made in the theoretical literature, replicate the ones that have already been evidenced by previous corpus linguistic studies and test whether my information-theoretic account can explain the findings in a unifying way. I will also consider verb surprisal as UID-specific factor that constrains the usage of topic drop.

4 Corpus study

I conducted a study on the fragment corpus (FraC) (Horch and Reich 2017) to test whether grammatical person, verbal inflection and verb surprisal influence the frequency of topic drop. Since FraC is not annotated for information-structural categories, it is unsuitable to investigate topicality. I will focus on this factor in the acceptability rating studies presented in Section 5. The corpus study partly replicates previous research, i. e. the work by Androutsopoulos and Schmidt (2002) and the study by Frick (2017) on Swiss German text messages but extends it substantially: First, I interpret the results with respect to my information-theoretic account and take into account verb surprisal as additional predictor. Second, I use logistic regressions, a more statistically elaborate method than the purely descriptive approach in Androutsopoulos and Schmidt (2002) and the chi-squared tests used in Frick (2017). Logistic regressions allow me to also consider interactions between the predictors that could influence the usage of topic drop.

4.1 Topic drop in the fragment corpus FraC

The data set is based on FraC (Horch and Reich 2017), a German text type-balanced corpus consisting of 17 different text types with about 2,000 utterances each, i. e. a total of about 34,000 utterances. The text types in FraC range from prototypically

written ones like newspaper articles to prototypically spoken ones like dialogues to written but conceptually spoken (Koch and Oesterreicher 1985) ones like text messages. The corpus is annotated for several omission phenomena including topic drop, object omission, article omission and copula omission. In the corpus there are a total of 967 instances of topic drop which I extracted along with the text type they occur in. I manually annotated a possible reconstruction of the omitted constituent, its syntactic function and its grammatical person and number.

The distribution by grammatical person and syntactic function is illustrated in Figure 3:[5] Just like in the corpus studies by Androutsopoulos and Schmidt (2002) and Frick (2017), the most frequently omitted constituents are 1st person singular pronouns, followed by the 3rd person singular which includes in particular instances of *das* 'that' and *es* 'it', whereas there are few omissions of the remaining grammatical persons.[6] Mainly subjects are targeted by topic drop.

Figure 4 shows the distribution of the 967 topic drops across conceptually spoken and conceptually written text types and confirms the text type dependency of topic drop attested in the literature: It occurs preferentially in conceptually spoken text types like dialogues and blogs, but above all in text messages: With 385 instances, almost 20 % of all utterances in this subcorpus contain an instance of topic drop.

4.2 Data set creation

Because of the predominance of topic drop in the text type text messages[7] and due to the large variability between text types (for instance, there are almost no

---

5 27 instances of dropped adverbials like *da* 'there' or *dann* 'then' were omitted from the figure for reasons of comprehensibility.

6 The reported numbers are absolute, i. e. they do not take into account how many instances of 1st or 3rd person singular pronouns are realized and do not allow to infer the subsequent omission rates. In principle such omission rates in form of relative numbers would be desirable for the whole corpus. However, this would demand a high annotation effort because theoretically one would have to annotate all syntactically complete utterances. Therefore, such relative numbers are only provided for the subcorpus as described below.

7 Answering the interesting question *why* text messages exhibit the highest ratio of topic drop among the text types in the corpus is beyond the scope of this paper. However, it is worth noticing that Thurlow and Poff (2013: 173) have ascertained that the linguistic and stylistic devices used in text messages do not differ a lot from the ones being characteristic for similar longer existing text types like notes (see also Frick 2017: 13). Although FraC does not show this link, previous literature has recognized topic drop as typical stylistic device of telegrams (Reis 1982; Barton 1998) which might be considered as kind of predecessors of text messages.
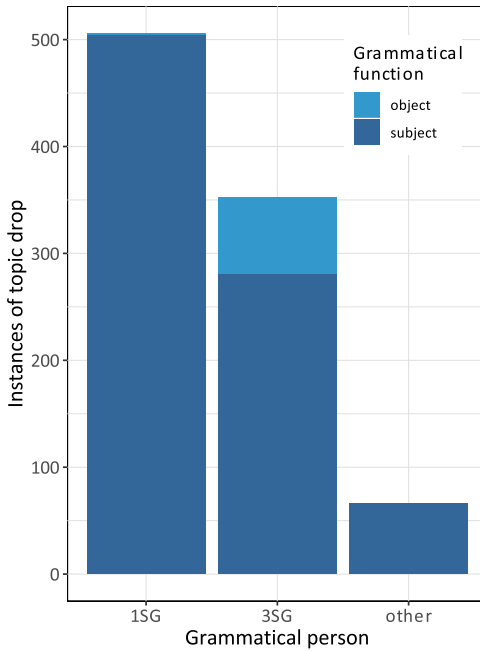
Figure 3: Instances of topic drop in FraC divided into grammatical person and grammatical function.

1st person singular pronouns in the ads subcorpus of FraC), I restrict my empirical investigations to topic drop in text messages. Furthermore, I limit it to the contrast between 1st and 3rd person singular subjects because topic drop mainly targets these persons and an effect of verbal inflection can only be observed for subjects, which inflectionally agree with the subject in German. For the statistical analysis of my predictors, I created a data set from the text messages subcorpus consisting in all 1st and 3rd person singular topic drop instances ($n_{\text{1SG,TD}}$ = 232, $n_{\text{3SG,TD}}$ = 42) and all syntactically complete utterances where a subject of the 1st or 3rd person singular occupies the preverbal position ($n_{\text{1SG,FULLFORM}}$ = 104, $n_{\text{3SG,FULLFORM}}$ = 57), i. e. the non-elliptical counterparts of the topic drop instances. I obtained these full forms with a semi-automatic approach: The corpus data were dependency-parsed and analyzed morphologically with the ParZu dependency parser for German (Sennrich et al. 2009) and lemmatized with the TreeTagger (Schmid 1994, 1999). I extracted all elements labeled as subject that occur before a finite verb and manually excluded noise, mainly instances where the preverbal constituent had falsely been classified as the subject. For each utterance I annotated the grammatical person of the (omitted or realized) preverbal constituent and whether
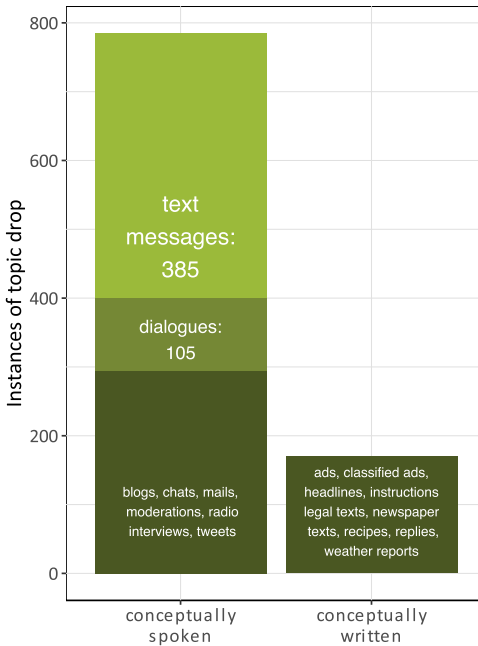
Figure 4: Topic drop in conceptually spoken ($n = 8$) vs. conceptually written ($n = 9$) text types in FraC.

the main verb is explicitly marked for inflection or not ($n_{explicit} = 328$, $n_{syncretic} = 107$).[8] Additionally, I extracted the unigram surprisal of the verb lemma from a language model trained on the lemmatized text messages subcorpus of FraC using the SRILM language modeling toolkit (Stolcke 2002). This unigram surprisal[9] measures the frequency of the verb lemma and is an approximation to a verb's probability in context. It is able to take into account properties of the text type text messages because the corresponding language model is trained on the text

---

8 The distribution of verb types was as follows: There were 203 full verbs, 180 auxiliaries and 52 modal verbs according to the parsing results. The syncretic forms mainly stem from modal verbs and auxiliaries in the preterite *war* 'was' and *hatte* 'had'.

9 This surprisal measure is a coarse approximation to a psychologically realistic surprisal measure. However, the fact that for many text messages no context is available, as well as the limitations of standard n-gram models that cannot take into account preceding sentences or extralinguistic context do not allow for the computation of more realistic estimates. A promising approach for future research could be to manipulate the verb surprisal experimentally, e. g. by presenting topic drop in a predictive and in an unpredictive context.

messages subcorpus only. Moreover, it approximates the likelihood of verbs comparatively well because finite verbs, unlike nominal expressions, are restricted in their syntactic distribution to the left bracket in most German declarative main clauses. Consequently, hearers can anticipate that the finite verb usually either follows the prefield constituent or appears sentence-initially in sentences with topic drop. When parsing a finite verb in the left bracket, it is hence only necessary to assess how likely the verbs are relative to each other. This likelihood is precisely quantified as unigram surprisal. The measure however remains an approximation since it is computed on lemmas and it also considers verbs in the right bracket and it may be the case that certain verbs occur more frequently in the right than in the left bracket. Still, it is the best available approximation for the verb's likelihood in this situation. I included it in my analysis to test the UID prediction that topic drop is more felicitous when the following verb is unexpected as this avoids surprisal maxima.

4.3 Analysis

I performed logistic regressions in R (R Core Team 2018) to predict topic drop from the predictors discussed above. To find the final model, I performed backwards model selection using model comparisons with likelihood-ratio tests (anova, R Core Team 2018): A model with the interaction or main effect in question was compared to a model without this effect. The same strategy was used to obtain p-values for the effects in the final model. The full model consisted of effects for the sum-coded predictors Person (1st vs. 3rd person singular), Inflection (distinct inflectional marking vs. syncretic) and unigram Surprisal of the verb lemma, as well as all two-way interactions between these predictors. The final model (Table 1) contained significant effects for Person, Surprisal and the interaction between Surprisal and Inflection, as well as a marginal effect of Inflection.[10] Its predictions are visualized in Figure 5. The main effect of Person shows that topic drop of the 1st person singular subject pronoun is more frequent than topic drop of the 3rd person singular subject pronoun (χ^2 = 27.63, p < .001). The main effect of the unigram Surprisal (χ^2 = 14.21, p < .001) indicates that topic drop is more frequent when the verb surprisal is lower. The significant interaction between Surprisal and Inflection shows that topic drop before verbs with a higher surprisal is more frequent when the verb has a distinct inflectional marking (χ^2 = 4.86, p < .05). There is a marginal main effect of Inflection (χ^2 = 3.07,

10 Omission ~ Person + Surprisal + Inflection + Inflection:Surprisal.

Table 1: Fixed effects in the final glm for the corpus study.

Predictor	Estimate	SE	χ^2	p-value	
PERSON	0.64	0.12	27.63	< 0.001	***
SURPRISAL	−0.23	0.06	14.21	< 0.001	***
INFLECTION	−0.93	0.54	−1.72	0.08	.
INFLECTION:SURPRISAL	0.14	0.06	4.86	< 0.05	*
number of observations					435
Pseudo R^2 (Nagelkerke)					0.11
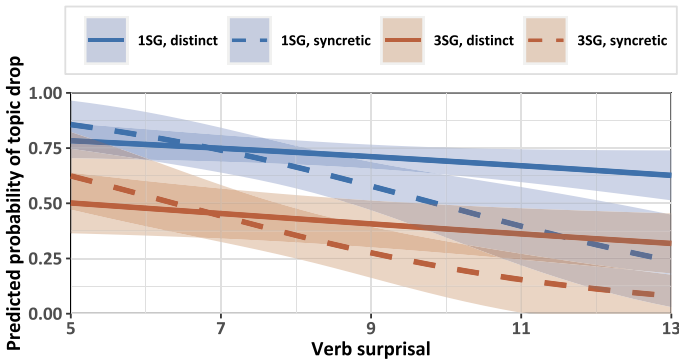
Figure 5: Probability of topic drop predicted by the final glm as a function of verb SURPRISAL, grammatical PERSON and INFLECTION.

$p = 0.08$) which suggests a slight trend towards topic drop being less frequent before verbs with distinct inflectional marking. The interaction between INFLECTION and PERSON is not significant ($\chi^2 = 1.3$, $p = 0.25$).

4.4 Discussion

The results of the corpus study support predictions from the theoretical literature and of my information-theoretic account. Concerning grammatical PERSON, they are in line with the previous corpus linguistic findings by Auer (1993), Androutsopoulos and Schmidt (2002) and Frick (2017): The 1st person singular is more frequently targeted by topic drop than the 3rd person singular. The inflectional hypothesis, however, cannot account for this finding, because there is neither a significant main effect of distinct verbal inflection nor an interaction between person and inflection present in the data. Topic drop is neither in general more fre-

quent when the following verb has a clear inflectional ending, rather there is an opposite tendency, nor is topic drop of the 1st person singular pronoun in particular more frequent before a verb with distinct marking. This strongly questions Auer's (1993) hypothesis that the prevalence of topic drop of the 1st person singular hinges on the easy reconstructability through distinct morphological marking. In contrast, both the pragmatic hypothesis and the information-theoretic account can provide an explanation for the higher frequency of topic drop of the 1st person singular pronoun. According to the pragmatic hypothesis, the speaker is the default origo of speaking, i. e. the 1st person singular pronoun is particularly easy to recover because its reference is clearly determined. From the UID perspective, the 1st person singular pronoun is more likely to be omitted because it is in general more likely to appear in prefield position than a 3rd person singular pronoun regardless of topic drop. In the data set used in my analysis there are 336 instances where the preverbal constituent — elided or not — is a 1st person singular pronoun and only 99 where it is a 3rd person singular pronoun. If the 1st person singular pronoun is in general more frequent, it becomes more likely that a speaker makes use of it, so the pronoun is less informative and more likely to be omitted.

The main effect of verb SURPRISAL provides genuine evidence for an information-theoretic account of the usage of topic drop. This evidence needs to be qualified by recalling that the surprisal measure used here is a coarse approximation rather than a psychological realistic estimate of a verb's real surprisal in context. It allows me however to take some form of context information into account namely the text type: A verb is more unpredictable when it occurs less frequently in text messages. Based on this, the main effect of verb SURPRISAL reflects the strategy to avoid surprisal maxima as predicted by the UID hypothesis: A preverbal constituent is less likely to be targeted by topic drop if it increases the likelihood of a subsequent highly unpredictable verb that would cause a peak in the information density profile. In this situation the preverbal constituent is a means to smooth the profile and is therefore more often realized.

The interaction between INFLECTION and SURPRISAL, i. e. that a distinct verbal inflection makes topic drop more frequent even with unexpected verbs, provides further evidence for an information-theoretic account. Topic drop before a verb with a high surprisal causes additional processing effort for the hearer: On the one hand, the high verb surprisal is more likely to exceed channel capacity, so that the amount of information is too high to be easily processed by the hearer. On the other hand, the hearer has to invest processing effort to recover the omitted constituent. A distinct inflectional marking on the verb provides information on the congruent subject. So if the subject is omitted from the prefield position, the clear verbal inflection can help to recover it. This way, the distinct verbal inflection

acts as a cue that facilitates recovering the omitted constituent and thus reduces the processing effort required to recover it and hence the total processing effort on the verb.

In sum, the corpus study replicated the effect of grammatical person that previous corpus linguistic studies found: Topic drop is more frequent with the 1st person singular than with the 3rd person singular. The absence of an interaction of INFLECTION with PERSON questions Auer's (1993) inflectional hypothesis as an explanation for this prevalence of the 1st person singular. Instead, both the pragmatic hypothesis and the information-theoretic account can explain the preference for 1st person singular topic drop with prominence or predictability of the speaker. What is more, the information-theoretic account can explain both the effect of verb SURPRISAL and its interaction with INFLECTION. This provides first evidence for the unifying character of the information-theoretic account.

5 Experimental investigations on topic drop in German

The corpus study provides first evidence for a unifying information-theoretic account of the usage of topic drop based on the factors grammatical person, verbal inflection and verb surprisal. The role of topicality, however, has not yet been investigated because FraC is not annotated for information-structural categories. Such an annotation would not only be costly, time-consuming and difficult to achieve due to the vague topic concept, but in part even impossible because for some text messages no pre-context is available. Therefore, I investigate topicality experimentally, which allows me to systematically control it using minimal pairs. This is also beneficial for the investigation of the effects of grammatical person and inflection as the 1st and 3rd person singular can be compared in an identical context.

5.1 Preliminary considerations

5.1.1 Setting the topic

Investigating topicality experimentally requires to determine and manipulate the topic of a sentence. Reinhart (1981: 62) notes that grammatical subjects may be considered "unmarked topics" although this relation is not obligatory because also objects and even non-NPs may serve as topics. Lambrecht (1994: 132)

supports this claim by stating that there is a strong cross-linguistic correlation between the syntactic category *subject* and the information-structural category *topic*. If an element is the subject of a sentence such as *Julia* in (4), then it should also be the "unmarked topic" and it should be more likely that it will as well be the subject and topic of the next sentence ((4-a) vs. (4-b)).

(4)   *Julia geht mit   mir essen.*
       Julia goes with me eat
       'Julia will be dining out with me.'

   a.   *Sie lädt   mich diesmal ein.*
         she invites me   this.time PART
         'She invites me this time.'   (topic (C_b) = preverbal constituent)
   b.   *Ich lade sie diesmal ein.*
         I   invite her this.time PART
         'I invite her this time.'   (topic (C_b) ≠ preverbal constituent)

This intuition is captured by the framework of centering theory (Grosz et al. 1995; Walker et al. 1998) that was originally developed to determine the reference of anaphora. I employ it as a mechanism to set the topic in my experimental items based on the previous utterance and on the grammatical function hierarchy. In centering theory, each utterance has a so-called *backward-looking center C_b* that corresponds to the concept of topic (Walker et al. 1998: 3). The C_b of an utterance U_n is chosen from the set of all referring expressions contained the previous utterance U_{n-1} (the so-called *forward-looking centers C_f*). In English and German these C_f are ordered based on grammatical function (see Walker et al. 1998 for English and Speyer 2007 for German; cf. Walker et al. 1998 for a different hierarchy in Japanese), i. e. subjects are ranked higher than objects and objects are ranked higher than adverbials. The C_b of U_n, i. e. its topic, is determined as the highest-ranked element from the C_f of the previous utterance U_{n-1} that is realized in U_n. In example (4), the C_b of both (4-a) and (4-b) is *Julia* pronominalized as *sie* 'she/her' because it is the subject of the context sentence U_{n-1} and hence the highest-ranked element of the set of forward-looking centers C_f of the previous utterance (ranked higher than the adverbial *mit mir* 'with me') that is mentioned in the target utterance. This means that in (4-a) the preverbal constituent is also the C_b, i. e. the topic, whereas in (4-b) the preverbal constituent and the C_b, i. e. the topic, are distinct. I use this difference as basis of my manipulation of topicality: I constructed conditions like (4-a) where the preverbal constituent of the target utterance is identical to the topic formalized as C_b and to the highest-ranked

element of C_f.[11] These are compared to conditions like (4-b) where the preverbal constituent of the target utterance is distinct from the topic formalized as C_b that is at the same time the highest-ranked element of C_f.

5.1.2 Materials

Based on this operationalization of topicality, I constructed 24 items like (5) for the two acceptability rating experiments. They were designed as short text message dialogues between two persons and have the following structure: The conversation starts with an unspecific question that serves the purpose of establishing a natural discourse (5-a). In the first sentence by the second conversation partner ((i) respectively) either a 3rd person like *Julia* or the speaker herself or himself is the subject.[12]

---

11 In these conditions topic continuity necessarily coincides with subject continuity. Previous research has already shown that subject continuity is preferred by German comprehenders when they resolve pronouns (Colonna et al. 2012). I expect a similar preference for resolving topic drop. This is not a problem for my information-theoretic approach because it could also account for a preference for topic drop based on subject continuity: The subject of the target utterance is known to the speaker and the hearer, so it is more predictable, less informative and should be more likely to be omitted in order to avoid a surprisal minimum. In future research however it is desirable to tease apart effects of topicality from effects of subject continuity. This might be done for example by testing pairs of context and target sentences like (i) where the subject of the context sentence is not picked up again in the target sentence at all. Therefore, the topic in form of the C_b of the target sentence cannot be retrieved via subject continuity. If topic continuity is the crucial factor for the acceptability of topic drop, I would expect that example (i) is rated better than (4-b). If subject continuity is the decisive factor, (i) should not be rated better than (4-b). And if both topic and subject continuity play a role, I would expect a three-part gradation of acceptability: (4-a) with topic and subject continuity should receive better ratings than (i) with only topic continuity and (i) should receive better ratings than (4-b) with neither topic nor subject continuity.

(i)   a.   *Julia geht mit mir essen.*
         b.   *(Ich) bestelle mir eine Pizza Sophia Loren.*
              I   order   me a   pizza Sophia Loren

12 In principle, it would be desirable to also look at the 2nd person singular because it exhibits relatively high omission rates in corpora too. However, it is hard to compare it straightforwardly to the 1st and 3rd person singular because assertive V2 sentences with the 2nd person like (i) appear to be marked. In most of the cases it is pragmatically odd to make statements using the 2nd person. This might explain why there are considerably less instances of both 2nd person singular full forms and topic drop in corpora.

(i)   ?*(Du) lädst sie / mich diesmal ein.*

The subject does not appear in prefield position, but the prefield is always filled with an adverbial. This is intended to rule out structural parallelism effects that could be expected since previous research found effects of parallelism on pronoun resolution, i. e. that pronouns are more likely to refer to a referent in a similar syntactic position (Smyth 1994; Chambers and Smyth 1998), should at least be alleviated.[13] The second character that is mentioned in the dialogue, i. e. the competing possible target of topic drop, is introduced in a prepositional phrase as an object or adverbial. This way, she or he occupies a less prominent position in the grammatical hierarchy *Subject > Object > Adverbial* (Walker et al. 1998; Speyer 2007) and is therefore less likely to be picked up (as topic) in the next sentence compared to the subject. The last utterance of the item, i. e. the target utterance ((ii) respectively), is produced by the conversation partner who has set the topic and either contains topic drop or not.

(5) a. A: *Was gibt's Neues?*
         A: 'What's the news?'

  b. (i) B: *Am Samstag geht Julia mit mir schick essen.*   TOPIC:
           B: 'On Saturday Julia dines out well with me.'   *identical*

     (ii) B: *(Sie) lädt mich diesmal ein.*   PERSON: *3SG*
           B: '(She) invites me this time.'   OMISSION: *omitted (realized)*

  c. (i) B: *Am Samstag gehe ich mit Julia schick essen.*   TOPIC:
           B: 'On Saturday I dine out well with Julia.'   *not identical*

     (ii) B: *(Sie) lädt mich diesmal ein.*   PERSON: *3SG*
           B: '(She) invites me this time.'   OMISSION: *omitted (realized)*

  d. (i) B: *Am Samstag geht Julia mit mir schick essen.*   TOPIC:
           B: 'On Saturday Julia dines out well with me.'   *not identical*

     (ii) B: *(Ich) lade sie diesmal ein.*   PERSON: *1SG*
           B: '(I) invite her this time.'   OMISSION: *omitted (realized)*

  e. (i) B: *Am Samstag gehe ich mit Julia schick essen.*   TOPIC:
           B: 'On Saturday I dine out well with Julia.'   *identical*

     (ii) B: *(Ich) lade sie diesmal ein.*   PERSON: *1SG*
           B: '(I) invite her this time.'   OMISSION: *omitted (realized)*

---

13 In order to fully rule out such an interference, it would be necessary to change the grammatical role of either the antecedent in the context sentence or of the prefield constituent in the target sentence – perhaps as sketched in footnote 11.

The experimental manipulation is based on three binary predictors: TOPIC, PERSON and OMISSION. TOPIC varies between *identical* and *not identical*. *Identical* means that the topic of the target utterance, i. e. the C_b, is identical to the preverbal constituent whereas *not identical* means that the preverbal constituent is not the C_b of the target utterance. For PERSON, there are the levels 1st person singular (*1SG*) and 3rd person singular (*3SG*) that refer to the grammatical person of the preverbal constituent. And OMISSION indicates whether the preverbal constituent is *omitted* or *realized*. In experiment 1, all target utterances contain a full verb in present tense with distinct morphological marking for grammatical person, i. e. the inflection and sometimes also parts of the verbal stem unambiguously indicate the grammatical person: In (5), the *e*-ending in *lade* clearly marks 1st person singular present tense, whereas the *-t* and the umlaut *ä* in *lädt* signal that it is the form of the 3rd person singular. In experiment 2, the full verbs were replaced by constructions with modal verbs that have syncretic forms for the 1st and the 3rd person singular (6). The modals were varied between items.[14] There was always an object pronoun referring to the competing referent which allowed the participants to unambiguously recover the subject of the sentence in cases with topic drop, even in absence of explicit inflectional marking. Therefore, potentially degraded ratings for utterances with syncretic verb forms cannot be attributed to the global ambiguity of the target utterances because the disambiguation due to the object pronoun takes place before the rating process.

(6) a. B: *(Sie)* **will** *mich diesmal einladen.* PERSON: *3SG*
          B: '(She) wants to invite me this time.' OMISSION: *omitted (realized)*
     b. B: *(Ich)* **will** *sie diesmal einladen.* PERSON: *1SG*
          B: '(I) want to invite her this time.' OMISSION: *omitted (realized)*

5.1.3 Presentation and procedure

In order to make topic drop more natural to the participants, I accounted for the register dependency of the phenomenon by presenting the material in a text messaging design (see Figure 6). The text type knowledge (Heinemann and Viehweger 1991) of the participants should be activated and it should be prevented that they

---

14 There were 8 sentences with *wollen* 'want', 5 sentences with *dürfen* 'may', 4 sentences with *sollen* 'shall', 3 sentences with *mögen* 'want' / 'would like', 3 sentences with *können* 'can' and 1 sentence with *müssen* 'must'.
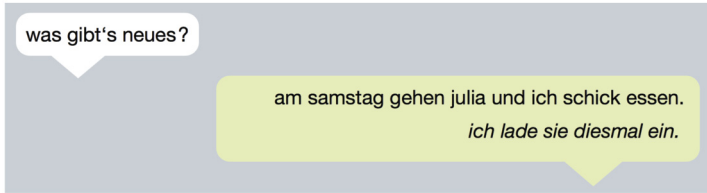
Figure 6: Text message design of the materials as presented to the participants.

only access their standard grammar and reject utterances with topic drop just because they are colloquial and typically restricted to specific text types. Furthermore, to keep the conditions comparable, I presented the whole text in lower case which is a common stylistic device in text messages (Schnitzer 2012). This way, the finite verb in the target utterance (i. e. *lade / lädt* in (5)) is written identically, i. e. with the initial letter in lower case, in the omitted and in the realized condition.

Both experiments were conducted over the Internet: The participants were recruited on the crowd sourcing platform *Clickworker*[15], and the actual survey was presented via the survey presentation software *LimeSurvey* (Limesurvey GmbH 2021).

5.2 Experiment 1

5.2.1 Hypotheses and predictions

In experiment 1, I investigate the acceptability of topic drop depending on topicality and grammatical person. The study tests whether topic drop is rated as more acceptable when the omitted constituent is the topic. The information-theoretic account that I propose predicts a significant interaction between TOPIC and OMISSION, i. e. that topic drop is more acceptable when the omitted constituent is the topic. Experiment 1 also investigates grammatical PERSON using minimal pairs that contrast the 1st and the 3rd person singular to see whether the differences in frequency that I found in the corpus study are reflected in differences in acceptability. The pragmatic and the inflectional hypotheses as well as the information-theoretic account predict a significant PERSON:OMISSION interaction, i. e. that topic drop is rated better for the 1st person singular.

---

15 https://www.clickworker.de

5.2.2 Design and method

Experiment 1 has the form of a $2 \times 2 \times 2$ within-subjects design crossing the 3 binary factors TOPIC, PERSON and OMISSION. This results in the 8 conditions illustrated in (5). The materials consist in 24 items and 80 fillers among which were 24 further utterances with omissions – instances of gapping and right node raising – to avoid that the topic drop items stand out as the only syntactically incomplete utterances. The materials were distributed among 8 lists using a Latin Square design. 48 self-reported native speakers of German between age 18 and 50 were paid 2.50 € respectively for participating in the study. Their task consisted in rating the naturalness of the last italicized utterance of each stimulus on a 7-point Likert scale (7 = completely natural).

5.2.3 Analysis

5 participants were excluded because they had exceeded a beforehand set threshold by having rated 4 or more (more than half) of 7 ungrammatical attention checks with 6 or 7 points on the scale, which indicates that they did not read all sentences carefully. The data of the remaining 43 participants were analyzed in R (R Core Team 2018) with cumulative link mixed models (CLMMs) for ordinal data (Christensen 2019). The same backwards-model selection procedure as in the analysis of the corpus study was performed. The full model contained the ratings as response variable and all three sum-coded variables and the corresponding two-way interactions between them. Starting from the maximal random effects structure justified by the data (Barr et al. 2013), I excluded those random effects that contributed least to explaining the data until the model converged. The final CLMM[16] included significant main effects for all three variables and for the interaction between PERSON and OMISSION, as well as for the interaction between TOPIC and OMISSION (Table 2). The random effects structure that was identical for the full and the final model consisted of random intercepts for participants and for items, by-participant random slopes for all predictors, the interaction between PERSON and TOPIC and for INDEX, i. e. the number of the item in the experiment, and by-item random slopes for all predictors.

---

16 RATING ~ TOPIC + PERSON + OMISSION + PERSON:OMISSION + TOPIC:OMISSION + (1 + TOPIC + PERSON + OMISSION + INDEX + PERSON:TOPIC | Subjects) + (1 + TOPIC + PERSON + OMISSION | Items).

Table 2: Fixed effects in the final CLMM for experiment 1.

| Predictor	Estimate	SE	χ^2	p-value	
TOPIC	0.41	0.24	8.26	< 0.05	*
PERSON	−0.51	0.2	7.11	< 0.01	**
OMISSION	−1.66	0.34	20.78	< 0.001	***
PERSON:OMISSION	−1.12	0.25	20.37	< 0.001	***
TOPIC:OMISSION	0.64	0.25	6.82	< 0.01	**
number of observations					1032
Pseudo R^2 (Nagelkerke)*					0.45

*The pseudo R^2 value was obtained by comparing the final CLMM to an intercept only CLM of the data (Mangiafico 2016).

Table 3: Mean ratings and standard deviations per condition for experiment 1.

OMISSION	PERSON	TOPIC	mean rating	standard deviation
full form	1SG	identical	5.64	1.44
topic drop	1SG	identical	5.38	1.4
full form	1SG	not identical	5.67	1.45
topic drop	1SG	not identical	4.21	1.67
full form	3SG	identical	5.58	1.43
topic drop	3SG	identical	4.65	1.6
full form	3SG	not identical	5.58	1.37
topic drop	3SG	not identical	4.91	1.46

5.2.4 Results

The final model revealed a significant main effect of PERSON (χ^2 = 7.11, p < 0.01) and a significant interaction between the predictors PERSON and OMISSION (χ^2 = 20.74, p < 0.001): Utterances with the 1st person singular were generally rated as better than utterances with the 3rd person singular but the 3rd person was particularly degraded in the topic drop conditions (cf. Figure 7). There was also a significant main effect of TOPIC, as well as a significant interaction between TOPIC and OMISSION (χ^2 = 7.97, p < 0.01): Utterances where the preverbal constituent was not the topic were overall degraded, but topic drop was rated particularly worse when topic and preverbal constituent were distinct. Additionally, I found a significant main effect of OMISSION (χ^2 = 20.78, p < 0.001): Syntactically complete utterances received higher ratings than utterances with topic drop.
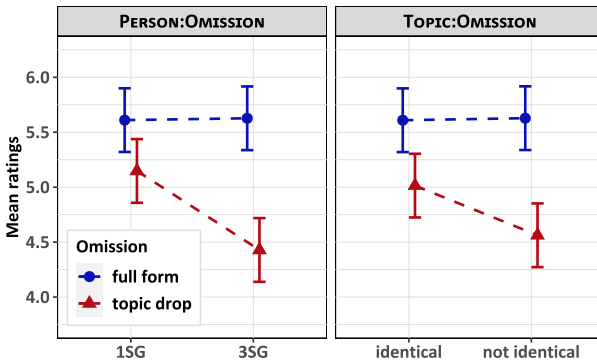
Figure 7: Mean ratings and 95 % confidence intervals for experiment 1. The complete data are once split per PERSON (including both TOPIC conditions), once per TOPIC (including both PERSON conditions). Topic drop of 1SG is rated significantly better than topic drop of 3SG. Topic drop of a topic constituent is rated significantly better than topic drop of a non-topic constituent.

5.2.5 Discussion

In experiment 1, I tested whether grammatical person and topic impact the acceptability of topic drop. For PERSON I found that the refusal of the 3rd person singular was particularly strong in the topic drop conditions although participants in general preferred utterances with the 1st person singular over utterances with the 3rd person singular. This result is in line with the finding of the corpus study. Topic drop of the 1st person singular pronoun is not only more frequent than topic drop of 3rd person singular pronouns but also more acceptable. As I showed in the discussion of the corpus study, a preference for topic drop of the 1st person singular is predicted not only by the inflectional and the pragmatic hypotheses but also by the information-theoretic account: I expect the 1st person singular not only to be more frequent in the corpus but in general in text messages. Since it is more likely that a speaker talks about herself or himself, the 1st person singular is more predictable, i. e. less surprising and more likely to be omitted because this avoids a surprisal minimum in the information density profile.[17]

The information-theoretic account also predicts the effect of TOPIC: Topic drop is more acceptable when the omitted constituent is also the topic, i. e. the C_b of the

---

17 In a future study it would be desirable to assess how likely either referent is to be the preverbal constituent of the target sentence with a production study. If the prefield is more often filled with the 1st person singular subject, this would provide additional evidence for the line of reasoning of the information-theoretic account. I thank an anonymous reviewer for pointing this out.

target utterance in the framework of centering theory. The subject of the previous utterance occupies the highest position in the ranking of the forward-looking centers C_f and therefore necessarily becomes the C_b of the target utterance if it is picked up in this utterance. This results in higher predictability and lower surprisal of the C_b of the target utterance. If the C_b is placed in the prefield, omitting it, i. e. using topic drop, leads to a more efficient distribution of surprisal across the utterance because a surprisal minimum is avoided.

The main effect of OMISSION according to which topic drop is degraded across the board compared to the full form has not been predicted by my account but can be straightforwardly accounted for: Topic drop is, as already described above, a phenomenon of informal and colloquial speech and text types. When participants are asked to rate the acceptability of utterances containing such a phenomenon it is likely that they orient themselves at least to a certain degree by standard grammar even though they were instructed to only use their intuitions. So it seems plausible that they gave lower ratings to topic drop just exactly because it is such a colloquial phenomenon despite the presentation as text messages.[18] Since my experiment shows *relative* differences between conditions it is not impacted by such independent reasons that influence the ratings for topic drop in general.

In sum, both the effects of PERSON and of TOPIC are in line with my proposed information-theoretic account to the usage of topic drop while the OMISSION effect can be explained by a recourse to standard grammar.

5.3 Experiment 2

Experiment 1 as well as the corpus study are in line with previous literature that has reported a higher frequency of topic drop of the 1st person singular: Topic drop of *ich* is not only more frequent but also more acceptable. Experiment 2 investigates the reason for this preference. In Section 2.1, I discussed two approaches from the theoretical literature that aim at answering this question: First, there is the inflectional hypothesis by Auer (1993), who attributes this preference to the distinct verbal inflection for the 1st person singular in the present tense. Second, Imo (2014) argues for a pragmatic hypothesis according to which the speaker as default origo of speaking is easy to recover. The information-theoretic account that I am proposing, shares both predictions: If the 1st person singular is the origo of

---

18 The mean rating of 4.79 ($sd = 1.59$) for all topic drop conditions compared to the mean rating of 2.88 ($sd = 1.83$) for the ungrammatical catch trials clearly excludes the possibility of a floor effect. Topic drop was degraded compared to the full forms but it still received quite high ratings.

speaking, it is more predictable, i. e. less surprising and more likely to be omitted following UID. A distinct verbal inflection can help to recover topic drop as it provides information about the grammatical person of the omitted subject and thus reduces the overall processing effort on the verb following topic drop.

Experiment 2 tests the inflectional hypothesis by using modal verbs instead of full verbs: In experiment 1, the full verbs in present tense had a distinct inflection for the 1st and the 3rd person singular. This allowed the participants to clearly identify already at the verb which constituent had been omitted in the topic drop conditions: *ich lad-e ein* 'I invite' vs. *sie läd-t ein* 'she invites'. For German modal verbs however, the forms for 1st and 3rd person singular are identical, e. g. *ich kann* 'I can' vs. *sie kann* 'she can'. My items are hence temporarily ambiguous[19] in the topic drop conditions. Participants may only recover the omitted constituent when they find the object pronoun of the competing referent as no distinct verbal morphology is available.

5.3.1 Hypotheses and predictions

Since Auer (1993) explains the prevalence of topic drop of the 1st person singular with the distinct inflectional marking on the verb in present tense, the reverse conclusion has to be that the prevalence is no longer present when this distinct marking is absent. Thus, if Auer's claim is correct, there should be no significant interaction between Person and Omission when there is no distinctive inflectional marking on the verb. The pragmatic hypothesis by Imo (2014), however, as well as the information-theoretic account, would still predict such an interaction to be present: The preference for topic drop with the 1st person singular hinges on the fact that the 1st person singular is more easily recoverable given that the speaker is either the origo of speaking (pragmatic hypothesis) or in general more predictable (information-theoretic account). Hence, the hypothesis to be tested in experiment 2 is whether topic drop of a 1st person singular pronoun is still preferred over topic drop of a 3rd person singular pronoun even when the verb forms are identical as

---

19 As the rating study is an offline task, the disambiguation is completed when the rating takes place. To observe the disambiguation process itself, an online method like self-paced reading would be necessary. It could be interesting to insert some material between the finite verb and the disambiguating pronoun (*can take a vacation* to visit the zoo *with me tomorrow*). The reading time on the pronoun and a spillover region could tell whether the disambiguation towards either 1st person singular or to 3rd person singular causes differences in processing effort. If the 1st person singular is more predictable, then the disambiguation towards the 1st person singular should be more likely, and this should be reflected in faster reading times.

it is the case for the partly syncretic modal verbs. Furthermore, experiment 2 tests again the impact of topicality on topic drop with the intention of replicating the result from experiment 1. Like in experiment 1, a significant interaction between TOPIC and OMISSION would support this hypothesis.

5.3.2 Design and method

Like experiment 1, experiment 2 had a 2 × 2 × 2 within-subjects design crossing TOPIC, PERSON and OMISSION with the same 8 conditions as in experiment 1. 48 native speakers of German between age 18 and 50 years who had not taken part in experiment 1 received 2.50 € for participating. The items from experiment 1 were adapted as sketched in Section 5.1.2, i. e. the full verbs were replaced by constructions with syncretic modal verbs. The same procedure, i. e. the collection of ratings on a 7-point Likert scale, and the same fillers as in experiment 1 were used.

5.3.3 Analysis

No participants were excluded because none of them had rated more than the previously established threshold of 4 of 7 ungrammatical attention checks with 6 or 7 points on the scale. The data of 48 subjects were analyzed with CLMMs, following the procedure described for experiment 1. The full model was identical to the full model in experiment 1. The final CLMM[20] contained the ratings as response variable, of the sum-coded predictors PERSON, OMISSION and an interaction between PERSON and OMISSION (see Table 4), as well as random intercepts for participants and for items and by-participant random slopes for all predictors, the interaction between PERSON and TOPIC and for index, and by-item random slopes for all predictors.

5.3.4 Results

Just like in experiment 1, there were a significant main effect of PERSON ($\chi^2 = 8.1$, $p < 0.01$) and a significant interaction between PERSON and OMISSION ($\chi^2 = 16.85$, $p < 0.001$): Again, there was a general preference for utterances with the 1st person singular and a specific preference for topic drop of the 1st person singular

---

20 RATING ~ PERSON + OMISSION + PERSON:OMISSION + (1 + TOPIC + PERSON + OMISSION + INDEX + PERSON:TOPIC | Subjects) + (1 + TOPIC + PERSON + OMISSION | Items).

Table 4: Fixed effects in the final CLMM for experiment 2.

Predictor	Estimate	SE	χ^2	p-value	
PERSON	−0.57	0.19	8.1	< 0.01	**
OMISSION	−1.06	0.22	18.83	< 0.001	***
PERSON:OMISSION	−0.93	0.23	16.85	< 0.001	***
number of observations					1152
Pseudo R^2 (Nagelkerke)[*]					0.32

[*]The pseudo R^2 value was again obtained by comparing the final CLMM to an intercept only CLM of the data (Mangiafico 2016).
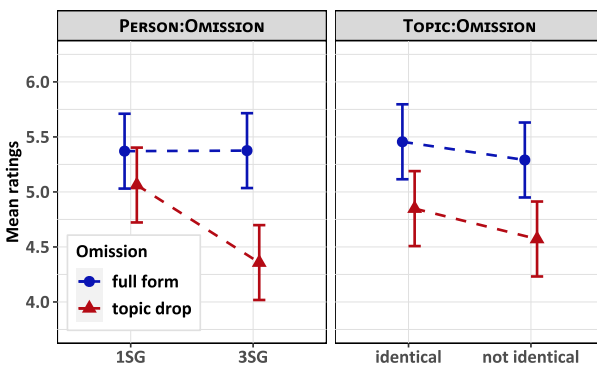
Figure 8: Mean ratings and 95 % confidence intervals for experiment 2. The complete data are once split per PERSON (including both TOPIC conditions), once per TOPIC (including both PERSON conditions). *Topic Drop* of 1SG is rated significantly better than topic drop of 3SG. There is no significant interaction between TOPIC and OMISSION.

as compared to the 3rd person singular. Furthermore, I found again a significant main effect of OMISSION (χ^2 = 18.83, p < 0.001): Syntactically complete utterances were rated better than utterances with topic drop. In contrast to experiment 1, I found no significant interaction between TOPIC and OMISSION (χ^2 = 2.03, p = 0.15): Utterances with topic drop were not rated as more acceptable when the omitted constituent was also the topic than when it was not.

5.3.5 Discussion

Experiment 2 tested the inflectional hypothesis (Auer 1993) as an explanation to the prevalence of topic drop with the 1st person singular. This hypothesis is

Table 5: Mean ratings and standard deviations per condition for experiment 2.

OMISSION	PERSON	TOPIC	mean rating	standard deviation
full form	1SG	identical	5.41	1.55
topic drop	1SG	identical	5.24	1.67
full form	1SG	not identical	5.25	1.74
topic drop	1SG	not identical	4.27	1.77
full form	3SG	identical	5.35	1.67
topic drop	3SG	identical	4.86	1.77
full form	3SG	not identical	5.52	1.43
topic drop	3SG	not identical	4.43	1.93

strongly questioned by the result that topic drop of the 1st person singular pronoun is rated better than topic drop of the 3rd person singular even in absence of distinct inflectional marking. The preference for the 1st person singular apparently does not hinge on the clear inflectional marking that facilitates the recovery of the omitted constituent but has to be motivated by other factors. Such a factor could be the prominence of the speaker as origo of speaking as proposed by Imo (2014) that makes her or him easy to recover. As stated already above, the information-theoretic account that I propose derives the same prediction from a more general line of reasoning: It predicts that topic drop is more acceptable the more predictable the omitted expression is. Consequently, if the 1st person singular is more predictable because it is the origo of speaking and hence more frequent in corpora, it is less surprising and more likely to be omitted.

For the main effect of OMISSION the same argumentation applies as for experiment 1, namely the hypothesis that participants oriented themselves by standard grammar when they rated the topic drop conditions. The missing significance of the interaction between TOPIC and OMISSION is unexpected under my information-theoretic account. If a topic is more predictable, it should be more likely to cause a surprisal minimum in the information density profile which should in turn increase the likelihood of its omission. However, it seems to be the case that topic drop of a topic is rated as better only when it occurs before a verb that is distinctly marked for inflection. My information-theoretic account based on both predictability and recoverability could explain this with a combined effect of topicality and explicit verbal inflection: While the omitted constituent is more predictable when it is topic, a distinct verbal inflection facilitates its recovery because it provides information about the grammatical person of the omitted subject. Both factors together hence seem to improve the acceptability of topic drop. In experiment 2, however, when the distinct verbal inflection is no longer

present, the recovery of the omitted constituent is not facilitated. The topicality of the omitted constituent alone does not seem to be sufficient to improve the acceptability of topic drop.

6 General discussion

In order to answer the question of when topic drop is used I conducted a corpus study and two acceptability rating experiments. My studies are the first joint systematic empirical investigation of claims made in the literature on the factors grammatical person, verbal inflection and topicality. I find empirical support for an information-theoretic account of topic drop based on the additional factor verb surprisal. My account provides a unifying explanation to previously isolated observations and needs to be extended in future research. There is first evidence that topic drop is more likely to be used and perceived as more acceptable when the omitted constituent is predictable in context and easy to recover. From an information-theoretic perspective, this distributes processing effort more efficiently across the utterance.

For grammatical person I showed that topic drop of the 1st person singular pronoun is more frequent in my text messages corpus than topic drop of 3rd person singular pronouns. This result is in line with previous corpus linguistic studies and with two hypotheses from the previous literature, the inflectional and the pragmatic hypothesis. The two rating experiments show that the higher frequency of topic drop with the 1st person singular is reflected in higher acceptability. My information-theoretic account provides an explanation for the prevalence of 1st person singular topic drop: The 1st person singular is overall more frequent in the text messages subcorpus, which makes it more predictable and less surprising. According to UID, omitting a constituent with low surprisal is preferable because this avoids a local surprisal minimum. While the frequency of the 1st person singular pronoun in the text messages subcorpus is a valid indicator of its predictability in a future study it would nevertheless be desirable to evidence that the 1st person singular is indeed more probable in the experimental items than the 3rd person singular.

This UID-based explanation for the prevalence of 1st person singular topic drop partly covers the line of reasoning of the pragmatic hypothesis by Imo (2014). This pragmatic hypothesis however is based on the extra-linguistic factor origo of speaking. Any additional factor one needs to assume to explain topic drop makes the respective account more complicated. It is hence an advantage of the

information-theoretic account that it can subsume the impact of the factor origo as part of the predictability of the omitted constituent.

My data on verbal inflection provide evidence against the inflectional hypothesis by Auer (1993). The corpus study revealed that topic drop was not more frequent before verbs that have a distinct inflectional marking and experiment 2 showed that topic drop of the 1st person singular was still preferred over topic drop of the 3rd person singular even without explicit inflectional marking. So a distinct inflectional marking cannot explain why topic drop of the 1st person singular is more acceptable.

Distinct verbal inflection however seems to play a role for the recovery of the omitted constituent. It provides a cue to the grammatical person of the omitted subject and thus reduces the processing effort caused by recovering the omitted constituent. In my corpus study, this is indicated by a higher ratio of topic drop before unpredictable verbs when the verbs are distinctly marked for inflection. Moreover, my experiments show that topic drop of the preverbal constituent is only rated better when it is topic *and* the following verb has a distinct inflectional ending. Put differently, I did only find a preference for omitting a topic when the following verb provided information on how to resolve ellipsis. This result suggests that topicality alone is not strong enough as a cue to favor topic drop.

The results so far are in line both with findings in the theoretical literature and with the information-theoretic account. However, only the latter predicts an effect of verb surprisal. The corpus study provides first evidence that the usage of topic drop is constrained by a strategy to avoid surprisal minima and maxima. Topic drop is less frequent in my data set when the main verb has a high unigram surprisal which means when the verb lemma occurs rarely in the text type text messages. In this case, realizing the preverbal constituent reduces the high processing effort on the unpredictable verb because the ellipsis does not have to be resolved. This leads to an overall more efficient distribution of surprisal across the utterance. In a future study it would be desirable to measure surprisal in a more psychological realistic manner so that the probability is constrained not only by the text type but also by the local linguistic context. This might be best achieved in an experiment where properties of the context are manipulated which impact the surprisal of either the preverbal constituent or the following verb.

My information-theoretic account integrates previously isolated observations from the theoretical literature into a unifying approach to the usage of topic drop: Topic drop is used to distribute surprisal, i. e. processing effort, efficiently across utterances. My account gains additional explanatory power as I provide first data on an effect of the following verb's surprisal on topic drop. Future research is needed to extend these findings by using a more sophisticated surprisal measure and by disentangling topicality and subject continuity.

References

Altmann, Gerry T. M. & Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73(3). 247–264. 10.1016/s0010-0277(99)00059-1.

Androutsopoulos, Jannis & Gurly Schmidt. 2002. SMS-Kommunikation: Ethnografische Gattungsanalyse am Beispiel einer Kleingruppe. *Zeitschrift für Angewandte Linguistik* 36. 49–79.

Auer, Peter. 1993. Zur Verbspitzenstellung im gesprochenen Deutsch. *Deutsche Sprache* 23. 193–222.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278. 10.1016/j.jml.2012.11.001.

Barton, Ellen L. 1998. The grammar of telegraphic structures: Sentential and nonsentential derivation. *Journal of English Linguistics* 26(1). 37–67. 10.1177/007542429802600103.

Cardinaletti, Anna. 1990. Subject/Object asymmetries in German null-topic constructions and the status of specCP. In Joan Mascaró & Marina Nespor (eds.), *Grammar in progress*, 75–84. Berlin & New York: De Gruyter Mouton. 10.1515/9783110867848.75.

Chambers, Craig G. & Ron Smyth. 1998. Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language* 39(4). 593–608. 10.1006/jmla.1998.2575.

Christensen, Rune Haubo Bojesen. 2019. Ordinal—Regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Colonna, Saveria, Sarah Schimke & Barbara Hemforth. 2012. Information structure effects on anaphora resolution in German and French: A crosslinguistic study of pronoun resolution. *Linguistics* 50(5). 991–1013. 10.1515/ling-2012-0031.

Demberg, Vera & Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2). 193–210. 10.1016/j.cognition.2008.07.008.

Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*. Frankfurt / Main: Diesterweg.

DUDEN. 2016. *Duden. Die Grammatik. 9.*, vollständig überarb. und aktual. Auflage. Berlin: Dudenverlag.

Frey, Werner. 2000. Über die syntaktische Position der Satztopiks im Deutschen. *ZAS Papers in Linguistics* 20. 137–172.

Frick, Karina. 2017. *Elliptische Strukturen in SMS: Eine korpusbasierte Untersuchung des Schweizerdeutschen* (Empirische Linguistik = Empirical Linguistics 7). Berlin & Boston: De Gruyter. 10.1515/9783110517859.

Fries, Norbert. 1988. Über das Null-Topik im Deutschen. *Sprache & Pragmatik* 3. 19–49. 10.18452/9478.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225.

Haegeman, Liliane. 1997. Register variation, truncation, and subject omission in English and in French. *English Language and Linguistics* 1(2). 233–270. 10.1017/S1360674300000526.

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 159–166. Pittsburgh, PA: Association for Computational Linguistics. 10.3115/1073336.1073357.

Heinemann, Wolfgang & Dieter Viehweger. 1991. *Textlinguistik: Eine Einführung* (Reihe Germanistische Linguistik 115). Tübingen: Niemeyer. 10.1515/9783111376387.

Helmer, Henrike. 2016. *Analepsen in der Interaktion: Semantische und sequenzielle Eigenschaften von Topik-Drop im gesprochenen Deutsch* (Oralingua 13). Heidelberg: Universitätsverlag Winter.

Horch, Eva & Ingo Reich. 2017. The Fragment Corpus (FraC). In *Proceedings of the 9th International Corpus Linguistics Conference*, 392–393. Birmingham (UK): University of Birmingham.

Huang, C.-T. James. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15(4). 531–574.

Imo, Wolfgang. 2014. Elliptical structures as dialogical resources for the management of understanding. In Susanne Günthner, Wolfgang Imo & Jörg Bücker (eds.), *Grammar and Dialogism*, 139–176. Berlin & Boston: De Gruyter. 10.1515/9783110358612.139.

Jacobs, Joachim. 1999. The dimensions of topic-comment. *Linguistics* 39(4). 641–681. 10.1515/ling.2001.027.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62. 10.1016/j.cogpsych.2010.02.002.

Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36. 15–43.

Kravtchenko, Ekaterina. 2014. Predictability and syntactic production: Evidence from subject omission in Russian. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 785–790. Quebec City, Canada: Cognitive Science Society.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents* (Cambridge Studies in Linguistics 71). Cambridge University Press. 10.1017/CBO9780511620607.

Lemke, Robin, Eva Horch & Ingo Reich. 2017. Optimal encoding! – Information Theory constrains article omission in newspaper headlines. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 131–135. Valencia, Spain: Association for Computational Linguistics.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. 10.1016/j.cognition.2007.05.006.

Levy, Roger & Florian T. Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John Platt & Thomas Hofmann (eds.), *Advances in Neural Information Processing Systems 19*, 849–856. Cambridge, MA: The MIT Press. 10.7551/mitpress/7503.003.0111.

Limesurvey GmbH. 2021. LimeSurvey: An Open Source survey tool. http://www.limesurvey.org.

Mangiafico, Salvatore S. 2016. *Summary and analysis of Extension Program Evaluation in R, version 1.18.1*. rcompanion.org.

Marslen-Wilson, William D. 1973. Linguistic structure and speech shadowing at very short latencies. *Nature* 244(5417). 522–523. 10.1038/244522a0.

Marslen-Wilson, William D. 1975. Sentence perception as an interactive parallel process. *Science* 189(4198). 226–228. 10.1126/science.189.4198.226.

Molnár, Valéria. 1998. Topic in focus. On the syntax, phonology, semantics and pragmatics of the so-called "contrastive topic" in Hungarian and German. *Acta Linguistica Hungarica* 45(1–2). 89–166. 10.1023/A:1009609008756.

Musan, Renate. 2002. Informationsstrukturelle Dimensionen im Deutschen. Zur Variation der Wortstellung im Mittelfeld. *Zeitschrift für Germanistische Linguistik* 30(2). 198–221. 10.1515/zfgl.2002.015.

Oppenrieder, Wilhelm. 1987. Aussagesätze im Deutschen. In Jörg Meibauer (ed.), *Satzmodus zwischen Grammatik und Pragmatik* (Linguistische Arbeiten 180), 161–189. Tübingen: Niemeyer.

R Core Team. 2018. R: A language and environment for statistical computing. Tech. rep. Vienna, Austria.

Reich, Ingo. 2018. Ellipsen. In Frank Liedtke & Astrid Tuchen (eds.), *Handbuch Pragmatik*, 240–251. Stuttgart: J. B. Metzler. 10.1007/978-3-476-04624-6_24.

Reinhart, Tanya. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27. 53–94.

Reis, Marga. 1982. Zum Subjektbegriff im Deutschen. In Werner Abraham (ed.), *Satzglieder im Deutschen: Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung* (Studien zur deutschen Grammatik 15), 171–211. Tübingen: G. Narr.

Ross, John Robert. 1982. Pronoun deleting processes in German. Paper presented at the annual meeting of the Linguistic Society of America, San Diego, California, 27–30 December.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging using decision trees. In *International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Schmid, Helmut. 1999. Improvements in Part-of-Speech Tagging with an application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora*, 13–25. Dordrecht: Springer. 10.1007/978-94-017-2390-9_2.

Schnitzer, Caroline-Victoria. 2012. *Linguistische Aspekte der Kommunikation in den neueren elektronischen Medien*: Ludwig-Maximilians-Universität München dissertation. doi: 10.5282/edoc.14779.

Sennrich, Rico, Gerold Schneider, Martin Volk & Martin Warin. 2009. A new hybrid dependency parser for German. In Christian Chiarcos, Richard de Castilho & Manfred Stede (eds.), *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From form to meaning: Processing texts automatically. Proceedings of Biennial GSCL Conference 2009*, 115–124. Tübingen: Narr. 10.5167/uzh-25506.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(4). 623–656. 10.1002/j.1538-7305.1948.tb00917.x.

Smyth, Ron. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research* 23(3). 197–229. 10.1007/BF02139085.

Speyer, Augustin. 2007. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft* 26(1). 83–115. 10.1515/ZFS.2007.011.

Speyer, Augustin. 2010. Filling the German Vorfeld in written and spoken discourse. In Sanna-Kaisa Tanskanen, Marja-Liisa Helasvuo, Marjut Johansson & Mia Raitaniemi (eds.), *Discourse in interaction*, 263–290. Amsterdam: Benjamins. 10.1075/pbns.203.19spe.

Sternefeld, Wolfgang. 1985. Deutsch ohne grammatische Funktionen. Ein Beitrag zur Rektions- und Bindungstheorie. *Linguistische Berichte* 99. 394–439.

Stolcke, Andreas. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 901–904.

Thurlow, Crispin & Michele Poff. 2013. Text messaging. In Susan C. Herring, Dieter Stein, Tuija Virtanen & Wolfram Bublitz (eds.), *Pragmatics of computer-mediated communication* (Handbooks of Pragmatics 9), 163–190. Berlin & New York: De Gruyter Mouton. 10.1515/9783110214468.163.

Trutkowski, Ewa. 2016. *Topic Drop and Null Subjects in German* (Linguistics & Philosophy 6). Berlin & Boston: De Gruyter. 10.1515/9783110446173.

Walker, Marilyn, Aravind Joshi & Ellen Prince. 1998. Centering in naturally-occurring discourse: An overview. In Marilyn Walker, Aravind Joshi & Ellen Prince (eds.), *Centering in discourse*, 1–28. Oxford: Oxford University Press.

Wickham, Hadley. 2016. *Ggplot2: Elegant graphics for data analysis*. New York: Springer.

Zdorenko, Tatiana. 2010. Subject omission in Russian. A study of the Russian National Corpus. *Language and Computers* 71(1). 119–133. 10.1163/9789042028012_009.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache* 1 (Schriften des Instituts für Deutsche Sprache 7). Berlin & New York: De Gruyter. 10.1515/9783110872163.