Chapter 5

The role of information in modeling German intensifiers

J. Nathanael Philipp[a,b], Michael Richter[b], Tatjana Scheffler[c] & Roeland van Hout[d]

[a]Sächsische Akademie der Wissenschaften zu Leipzig [b]University of Leipzig
[c]Ruhr University Bochum [d]Radboud University

In this study, context-free and context-dependent information measures are applied to a new corpus of tweets and blog posts. The aim is to account for the expressive meaning and characterize the variability of available intensifying items. It comes to light that context-free and context-dependent information measures are highly correlated and account for the distribution of intensifiers in the data, giving credence to the notion that intensifiers form a common word class, even across syntactic and semantic differences.

Both information measures show that stacked intensifiers tend to be ordered from least to most expressive within a phrase, i.e., the information tends to increase. We explain this fact using the Uniform Information Density Hypothesis: The first, less expressive intensifier is used to introduce the phrase, ease the reader's processing load, and smooth the information flow.

1 Introduction

Intensifiers such as as *very* and *so* add little to the referential content of a sentence. Instead, these expressions serve to increase the expressive value of an utterance by indicating that some property applies to a higher degree (1), or even merely by signaling a heightened emotional state (2):

(1) *Der Sommer war so schön![1]*
    The summer was so pretty
    'The summer was so wonderful!'

(2) *Ich freu   mich so! :-)*
    I   be-happy me   so :-)
    'I am so happy! :-)'

Intensifying expressions are subject to immense variation in a language and can often be added to an utterance in informal discourse without adding much to its core denotation. Their use and choice are therefore governed by expressivity as well as information structure and general communicative needs. We argue that inserting an intensifier in an utterance can ease processing by evening out the information density in a sentence to avoid large heights or troughs.

In previous empirical studies (Richter & Van Hout 2020, Scheffler et al. 2023b), expressivity and stacking of intensifiers in Dutch and German Twitter data are modeled and operationalized by information theory (Shannon 1948) and information measures derived from it. Richter & Van Hout (2020) and Scheffler et al. (2023b) distinguish a paradigmatic information value that represents expressivity (and represents the strength of an intensifier) and a syntagmatic information value representing transitional information (i.e., its ability to combine with various targets). The authors observe that in stacks of intensifiers, expressivity increases from left to right, leading them to conjecture that blander, less expressive intensifiers precede the more expressive ones and thus smooth information flow in the sentence. In addition, they argue that intensifiers may lose their expressive value over time, while new and expressive intensifiers carrying high surprisal come into vogue, and that the expressivity of intensifiers can be boosted by capitalization, lengthening (*soooooo*), and repetitions (*very, very nice*). In addition, it has been shown that the choice of a longer (and thus typically more unusual) intensifier such as *atemberaubend* ('breathtakingly') over a short intensifier like *sehr* ('very') can further increase the expressivity of an utterance (Bennett & Goodman 2018).

In this paper, we follow the previous line of research by Richter & Van Hout (2020) and Scheffler et al. (2023b) and represent the expressivity of intensifiers through the amount of information encoded in an expression (Shannon 1948). That is, our model operationalizes the measurement of expressivity. We also draw on the work of Bennett & Goodman (2018) mentioned above, who related costs, i.e. information, to the strength and expressivity of intensifiers.

---

[1]All German examples in this paper are attested in our corpus, except where marked otherwise.

The idea of determining expressivity and strength of intensification is already found in the work of Zadeh (1972) couched in a fuzzy set theoretic model, and in work by Potts (2007), who proposes a separate expressive dimension of expressions. None of these works, however, include an empirically based determination of the strength or expressivity of particular expressions. Our proposed model does not give a definition of expressivity itself, but rather provides a methodology for precisely measuring the amount of expressivity. We use semantic surprisal and Shannon information content (abbreviated as IC in the remainder of this paper) as lexical features of an intensifier w. Surprisal and IC are negative log-probabilities of w that indicate the amount of information in bits. Surprisal is calculated in the sentential context of w, i.e., within the sentence in which w occurs and within its extra-sentential context (Levy 2008), that is, contexts in preceding and following sentences. IC, in contrast, is context-free, i.e., it is simply derived from the relative frequency of w. Context plays a key role in the calculation of surprisal, and in the present study, we define semantic surprisal in the context of the topics in the environment of w. This interpretation of the semantic context of a word is introduced by Kölbl et al. (2020, 2021), and previously used by Philipp et al. (2022, 2023). We will address our definition of surprisal and the topic contexts it is based on in more detail in Section 3 below.

We apply these information theoretic analyses to a new dataset of manually annotated intensifiers in two media, tweets and blog posts, in order to confirm the predictions established in the previous research. The data consists of texts from blogs and tweets from the same set of 44 authors (Scheffler et al. 2023a). Including blog post texts enables us to extend the empirical picture of intensification by observing their behavior in a new medium. Our results show that while the intensifier word class indeed shows high variability, as indicated by previous studies, the class conforms to the information theoretic predictions established in prior work. These results hold even though the data we report on here is no longer restricted to intensification of predicative adjectives. We compare the two media tweets and blogs to characterize the media related variability of intensification in our German corpus. We show that both context-free and context-dependent (semantic) information measures exhibit high correlation and can account for the distribution of intensifiers.

Finally, Scheffler et al. (2023b) argue that the order of stacked intensifiers (when multiple intensifiers follow each other in the same phrase) is determined by Uniform Information Density (Fenk & Fenk 1980, Aylett & Turk 2004, Levy & Jaeger 2007, Jaeger 2010): They observe that intensifiers tend to occur from least to most expressive and conjecture that the less expressive first intensifier serves to introduce the phrase and "bridge" the information flow from copula

verb to the expressive intensifier and adjective. In this paper, we test this conjecture on our entire corpus of sentences with intensifiers. Since we now employ contextual information values which we can compute for each word in a text, we can systematically compare sentences with stacked intensifiers with the communicatively equivalent alternatives (i) containing only the last, most expressive intensifier, or (ii) containing both intensifiers in the opposite order. We compute information profiles for both alternatives and compare them to the actually attested variants. Our results show that the attested variants exhibit a smoother information contour with a slightly higher uniformity of the wordwise information density. This supports the Uniform Information Density hypothesis, which states that smoother information flow is preferred keeping other factors equal. This analysis may serve to shed light on the puzzle why writers bother to add further, redundant intensifiers beyond the most expressive one: Their function is to introduce the phrase, alert the reader to the word class which follows, and smooth the information flow.

In the following, we first discuss the word class of intensifiers and why they are of particular interest for information theoretic approaches to language. We then discuss the empirical base of our annotation study for intensifiers in two written social media, blogs and tweets, before providing more details on the information measures we use for our analysis. Finally, we present our quantitative results.

2 Corpus study: Intensification in blogs and tweets

2.1 Intensifiers as a word class

Gradable properties such as the body size expressed by the adjective *fat* can be strengthened or toned down by certain fixed words or phrases. These words or phrases are called intensifiers because they increase (3) or reduce (4) the intensity to which the gradable property applies. In common usage, non-gradable properties are also often the target of intensification (5).

(3)   This seal is extremely fat.   (constructed)

(4)   This seal is pretty fat.   (constructed)

(5)   This seal is completely round.   (constructed)

Finally, it is well-known that although adjectives are the most common targets of intensification, other words such as verbs (2) or nouns (6) can also be intensified (see Bolinger 1972), as in the following two examples from our social media dataset (see Section 2.2):

(6)   a.   *Was   ganz   Neues!*
           Some completely new-thing
           'Something completely new!'

   b.   *Und warum zur Hölle habe ich damit   so ein Problem, mich von*
        And why   in hell have I   with-that so a   problem, me   from
        den Kindern lösen   zu können?
        the kids   separate to be-able-to
        'And why in hell do I have such a problem with separating from the
        kids?'

In this study, we analyze German intensifiers using measures from information theory. We assume that all intensifiers share a common semantic core, the basic notion of intensification, though they may differ in other connotations (e.g., level of formality, strength, or semantic domain restrictions). We focus on a semantically relatively uniform subset of the more broader set of "intensity particles" (Breindl 2007) sometimes discussed, namely only those that increase the extent to which a property applies, as exhibited in all previous examples (1–6). This notion includes both so-called maximizers such as *total* ('totally'), which indicate only the end point of a scale, as well as boosters such as *sehr* ('very'), which restrict a property to a higher section of a scale (Bolinger 1972).

This means that we expressly exclude downtoners, which lower the grade to which a property applies rather than increasing it (e.g., *ein bisschen* 'a little', *etwas* 'somewhat'), from further consideration. Accordingly, we do not consider intensifiers that can only occur with negated properties (e.g., *gar* 'at all'), since the negation implies a lowered intensity of the modified property. We further exclude words which look like intensifiers but which should better be analyzed as focus adverbs (*einfach* 'simply', *echt*, *wirklich* 'really'; see Beltrama 2022, Scheffler et al. 2023b). In the following, we reserve the term *intensifier* for our more narrow definition and call the broad category of gradability modifiers (including downtoners as well as some other expressions) *grade indicators*.

Given that all remaining intensifiers share a common semantic contribution, it remains a mystery why a language should provide so many different variants (for German, see e.g. Claudi 2006, Stratton 2020). In addition, more intensifiers within the same phrase do not add much semantic value. Prior work proposes that longer intensification phrases (i.e., longer intensifiers or additional intensifiers within one phrase) increase the strength and thus the expressivity of intensification (Fortin 2011, Bennett & Goodman 2018, Fortin & Rainer 2022). Note though that this only explains the presence of intensifier stacks, but not the

clear preferences for intensifier ordering that have been observed. In addition, establishing that the addition of an intensifier in an already intensified phrase increases expressivity is merely restating the facts, if no mechanism can be proposed why additional intensifiers are more expressive.

In the following, we propose, based on our previous work (Richter & Van Hout 2020, Scheffler et al. 2023b), that both the variance between intensifiers as well as their stacking behavior can be modeled using information theoretic measures, which explain the different contributions of different intensifying expressions. We argue that the class of intensifiers is relatively uniform with respect to their core semantic contribution (increasing the degree of a property), but that they differ in expressive value. While the notion of amount of information which a linguistic item contributes has been characterized as that item's core propositional semantics in the past (and operationalized rather simply by (contextual) frequency), we extend this notion to include the non-propositional, expressive content of the item as well, for example the strength of intensification or involvement of the speaker when using an intensifier. In our view, it is this non-propositional, expressive difference between intensifiers that can be seen in their differing information values.

2.2 Dataset of blogs and tweets

Intensifiers are a highly variable word class (Tagliamonte 2008, Stratton 2020), which is undergoing constant innovation and with which speakers/authors adapt flexibly to different registers and media (Scheffler et al. 2022). They are typical of speech, but also occur frequently in social media and even increase in frequency in newspaper corpora (Schmidt 2022). In this research, we use a novel multi-medium dataset to confirm previous results and advance the state of the art on the role of information measures in modeling intensifiers. Previous research has rarely addressed the way individuals change their use of intensifiers when switching media or text type. The data we study here consists of a corpus of tweets and blog posts from 44 parenting bloggers, collected in 2017. The corpus consists of 81,440 tweets (~1.2m tokens) and 468 blog posts (~360k tokens) in total, and is available for academic research by request.[2] Each author ($N = 44$) is represented with about 5–10 blog posts and up to 3200 tweets (for more details, see Scheffler et al. 2023a). Any differences between the subcorpora can be attributed to special properties of the two media formats, since the authors and topics stay the same.

---

[2]http://tiny.cc/twiblocop

The corpus was automatically split into sentences and tokenized, and manually pseudonymized. The tweet subcorpus consists of 137,914 sentences, while the blog posts contain 24,981 sentences in total. In all texts, all grade indicators have been manually marked by student research assistants with the help of the software WebAnno (Eckart de Castilho et al. 2016). To facilitate annotation, the set of frequent intensifiers identified by Scheffler et al. (2023b) was automatically pre-selected. Annotators were asked to both add additional intensifiers not included in the previously known list, as well as disambiguate all occurrences of the pre-annotated words regarding their status as an intensifier. The annotation was based on a short guideline specifying examples and counterexamples. Contrary to some previous work, the annotation did not restrict the class of grade indicators (i.e., downtoners were included along with boosters; and intensity indicators modifying verbs and nouns were included along with adjective intensifiers). We aimed directly for the semantic/pragmatic function of intensity modification of a gradable property or relation. The annotation thus covered a broader range of items than are analyzed in this paper. Difficult cases were discussed with the students and the annotators did not raise major problems in identifying the grade indicators. To prevent any misunderstanding, we call this class of words grade indicators in the remainder of this paper.

Three of the authors individually evaluated all grade indicators annotated in the corpus manually in order to identify true intensifiers in the more narrow sense which we defined above, i.e., expressions which indicate that a property is strengthened or increased. Of the three evaluations, between a more conservative and a more lenient evaluation, we ultimately selected the intermediate one as our data for the following analysis.

3 Modeling the information structure of intensification

We use three information theoretic measures to model the intensifiers in our data: first, local information content (IC_{local}), which captures the paradigmatic information of a word. This measure was proposed by Scheffler et al. (2023b) to capture the variability of adjectival intensifiers and, actually, is the common Shannon Information, i.e., the logarithm of the unigram frequency of a linguistic unit w. In our case, the frequency is calculated (i) separately for the blog posts and for the tweets corpus and (ii) for the combined blog posts and tweet corpus. The definition is given in (7):

(7)
$$IC_{local}(w) = -\log_2 P(w)$$

Second, we introduce here a semantic notion of information, that is, semantic surprisal based on the Topic Context Model (TCM) which incorporates the document context of a word. Third, we compare sentence variants using the average information change per word in a sentence, $\text{UID}_{\text{wordwise}}$.

3.1 Surprisal

Surprisal was introduced by Tribus (1961) in an engineering science context, and later adapted to communication theory and psycholinguistics (Hale 2001, Levy 2008). The surprisal of a word is its contextualized information and quantifies how unexpected it is, or among how many alternatives it was chosen. The occurrence of a linguistic unit causes an amount of surprisal for the language processor which is low when the linguistic unit is expected and high when it is unexpected. The amount of surprisal is proportional to the effort that is necessary to process it (Hale 2001, Levy 2008). One can say that surprisal expresses the discrepancy between what a language processor expects in a sentence and what actually occurs. For example, given the string *the old man*, a language processor might expect a verb as the next word. The surprisal is high if a determiner occurs, as in the garden path sentence *the old man the boat*.

Hale (2001) and Hale et al. (2015) define the surprisal of a word formally as its negative log-probability, which is equivalent to the definition of Shannon information IC. However, the calculation of surprisal requires conditional probabilities while the calculation of IC does not. Here, the probabilities $P(w_1, w_2...w_n)$ are calculated using the chain rule. Consider the example sentence from above *the old man the boat*. The second determiner is extremely unexpected in this context, and its surprisal is thus high.

Equation 8 gives the definition of surprisal as a conditional probability of a word w_i given the prior context $w_1 \ldots w_{i-1}$ (Levy 2008), where $w_1 \ldots w_{i-1}$ represent co-occurrences of any kind of the target word w_i within a sentence, and CONTEXT represents extra-sentential context of any kind, e.g., lexical, syntactic or semantic structures.

(8) $$\text{surprisal}(w_i) = -\log_2(P(w_i \mid w_1, \ldots, w_{i-1}, \text{CONTEXT})$$

3.2 Topic context model

The Topic Context Model (TCM)[3] (Kölbl et al. 2020, 2021, Philipp et al. 2022, 2023) outputs the semantic surprisal of a word given the topics in its environment. Simply put, in this operationalization, semantic surprisal expresses how surprising

---

[3]https://github.com/jnphilipp/tcm

a word is to the language processor in the context of the topics in the words' environment. For example, the word *chocolate* is likely to have a low probability in the context of topics such as *information* or *algorithms*, and would probably cause a high degree of semantic surprise.

TCM surprisal is thus specific to its environment, which can be the entire corpus, or documents within the corpus, paragraphs and even single sentences (in the remainder of the article, surprisal as information from the Topic Context Model is symbolized by TCM, and we use documents (blog posts, tweets) as the environments).

Like generative probabilistic topic models, TCM is based on the assumption that contexts/discourses are thematically structured, that documents are generated by topics, and that topics are characterized by words. In the present study, TCM employs Latent Dirichlet Allocation (LDA, Blei et al. 2003) for topic detection. LDA treats documents as probability distributions of topics and topics as probability distributions of words.

The definition of the average semantic surprisal (of a word w in a document d) from topics is given in Equation 9. $P(w \mid t_i)$ is the probability of a word w given a topic t_i, see Equation 10 where $c_d(w_d)$ is the frequency of a word w in a document d, $|d|$ is the total number of words in a document d, and WT is the normalized word topic distribution of the LDA.

$$(9) \qquad \overline{\text{surprisal}}(w_d) = -\frac{1}{n} \sum_{i=1}^{n} \log_2 P(w \mid t_i)$$

$$(10) \qquad P(w_d \mid t_i) = \frac{c_d(w_d)}{|d|} WT_{w_d,t_i} P(t_i \mid d)$$

For example, suppose a word w occurs in n documents in a corpus. LDA yields a topic distribution t for every document d. This distribution takes the form of a k-dimensional vector of probabilities, which must sum to 1. TCM calculates the surprisal of w by applying the formulas 9 and 10 (Philipp et al. 2022, 2023).

Suppose the word *mega* occurs two times in a document d with 42 words in total, further suppose we train a LDA with three topics. The three probabilities for $P(t_i \mid d)$ are 0.732, 0.183 and 0.085 and the three values for WT_{w_d,t_i} are 0.261, 0.101 and 0.096. This results in three probabilities for $P(w_d \mid t_i)$, that is, 0.00910, 0.00088 and 0.00039, respectively. Utilising Formula 9, we derive from these probabilities the average surprisal value of 9.41812.

In total, we trained three different LDAs and therefore have three different TCM structures. The first LDA was trained only on the blog posts, the second

one only on the tweets, and the last one on the complete corpus. All LDAs were trained with 32 topics.[4] The training corpora contained 30,111 words of blog data, 26,656 words from Twitter and 45,754 from the complete corpus. All three TCM values behave similarly in our analyses. However, since each tweet constitutes its own document, and the tweets are generally very short, we expect the tweet-based models to be less reliable than the models based on the blog data, where each document consists of an entire blog post.

3.3 Wordwise uniform information density

Following Collins (2014), we determine local Uniform Information[5] as a measure of the average squared information change per linguistic unit in sequences of linguistic units. For this measure, Scheffler et al. (2023b) use the term $UID_{wordwise}$. Its definition is given in (11), where id is the information value of a unit, n is the number of units in a sequence (for instance in a sentence or in a stack on intensifiers), and the index $i = 2$ indicates the second unit in the sequence (counting from left to right).

$$(11) \qquad UID_{wordwise} = -\frac{1}{n} \sum_{i=1}^{n} (id_i - id_{i-1})^2$$

Note that $UID_{wordwise}$ is negative by definition, and therefore a $UID_{wordwise}$ value close to zero indicates a high uniformity of the information density distribution, that is, smaller information differences from unit to unit. Information density is an important principle in linguistic communication, since information peaks and troughs must not be too extreme, so as not to make it too difficult for the recipient of a message to process it (Fenk & Fenk 1980, Aylett & Turk 2004, Levy & Jaeger 2007, Jaeger 2010).

4 Results

4.1 Grade indicators and intensifiers

We found 3446 grade indicators in the blog posts and 2034 in the tweets. Given the token size of these two subcorpora (blogs ~360k, tweets ~1.2m), grade indicators turn out to be more frequent in the blogs. This result is even more pronounced when looking at the sentence level: The blog subcorpus contains 24,981

---

[4]After some experiments, this appeared as a good number to us.
[5]https://github.com/jnphilipp/uid

sentences, yielding an average of 0.139 grade indicators per sentence. In contrast, the tweets, with 137,914 utterances, contain about ten times fewer grade indicators, only 0.015 per sentence. Sentence segmentation was carried out automatically in prior work (Scheffler et al. 2023a). Our database distinguishes single grade indicator tokens and stackings (sequences of more than one grade indicator). After lemmatization (combining variants), we obtained the top 10, which are presented in Table 1 (both single indicators and stacks are included here).

Table 1: Frequencies of the 10 most frequent grade indicators, per medium and in the total dataset

Lemma	Translation	N blogs	N tweets	N total
so	'so'	641	551	1192
ganz	'completely'	486	263	749
sehr	'very'	424	259	683
gar	'at all'	200	134	334
zu	'zu'	180	68	248
wirklich	'really'	159	61	220
echt	'really'	30	79	109
total	'totally'	55	49	104
ziemlich	'pretty'	80	20	100
etwas	'somewhat'	76	22	98

The list in Table 1 shows grade indicators that are not an intensifier in the strict sense, like *zu* ('too'), *ziemlich* ('pretty'), and *etwas* ('somewhat'). It also shows three highly frequent intensifiers, *so*, *ganz*, *sehr*, in both media. Compared to the extensive tweet database studied by Scheffler et al. (2023b), the high frequency of *ganz* ('completely') in the blogs is remarkable, but the most important observation is that a comparable set of frequent intensifiers is present in this new dataset, as well. In the following analyses, we exclude all grade indicators that are not intensifiers according to our definition, including downtoners (*ziemlich* 'pretty', *etwas* 'somewhat'), and the excessivity marker *zu* ('too').

New intensifiers that were not present in the extensive tweet database by Scheffler et al. (2023b) all occurred with a frequency of only 1, e.g., *affengeil* ('(monkey) awesome'), *bombenfest* ('bomb proof'), *entsetzlich* ('terribly'), *gewaltig* ('hugely'), *höllisch* ('hellish'), *riesen* ('giant'). The occurrence of new intensifiers in fact demonstrates the openness of the word class of intensifiers and its innova-

tive power, as pointed out earlier in comparing different inventories of German intensifiers (Scheffler et al. 2023b).

Our database of blogs and tweets contains single occurrences and stacks. We start with analyzing the single occurrences. Next, we present the results on the stacks.

4.2 Single intensifiers

We selected only those intensifiers that were also present in the large Twitter database studied previously (Scheffler et al. 2023b) for further analysis. In that study, 89,358 intensified predicative phrases of the form pronoun + *is/was* + … + adjective (e.g., *Frankfurt ist so arsch weit* 'Frankfurt is so damn far') from a corpus of over 6 million German tweets were extracted. These intensified phrases contained 124 different frequent intensifiers, 68 of which also occur in our new blog and tweet corpus. They include *echt* and *wirklich* ('really'), which can express an intensifying function, but are in fact focus adverbs. We proceed with these 68 items as they qualify as validated, frequent intensifiers in German.

For all 68 intensifiers, two information measures were computed, one for frequency (IC_{local}, see above) and one for the transitional probability between the intensifiers and the following adjective (IC_{trans}). In the extensive tweet database (Scheffler et al. 2023b) the anticorrelation between these two information measures was extremely high, −0.916. This still applies to the 68 shared intensifiers in our new database. Their anticorrelation is even a bit higher at −0.944, indicating that there is no selective bias in our new corpus compared to the previous large Twitter database. IC_{local} in the previously studied large Twitter database is correlated with the log frequencies in both blogs and tweets of the new corpus (respectively, $r = 0.735$, $r = 0.838$). Interestingly, the correlation with the Twitter part of our new corpus is significantly higher than with blogs ($t = 3.06$, $p < 0.001$), as is to be expected since the previous database also consisted of tweets. The register distinction between blogs and tweets is confirmed by data on capitalization and lengthening in the tokens of our 68 remaining intensifiers (respectively only 1 and 3 capitalizations, and 14 (0.6%) and 50 lengthenings (3.1%) in blogs vs. tweets); these numbers are much lower than in the extensive tweet database collected earlier. In particular, *so* was the most popular intensifier for lengthening, similar to the findings in Scheffler et al. (2023b).

In this paper, we, in addition to IC, introduced the contextualized information measure TCM. This measure computes the information value of a word depending on its context (in our case, a document). These two information values are

related to each other, insofar as the simple IC value captures some of the information content of a word, while TCM modulates this value to a specific context. The scattergrams of these two measures are given in Figure 1 and Figure 2. In the blog posts, the TCM value of single intensifiers correlates linearly with the IC values (Figure 1), whereas the relation is more fuzzy in tweets (Figure 2) in the higher regions of IC and TCM.
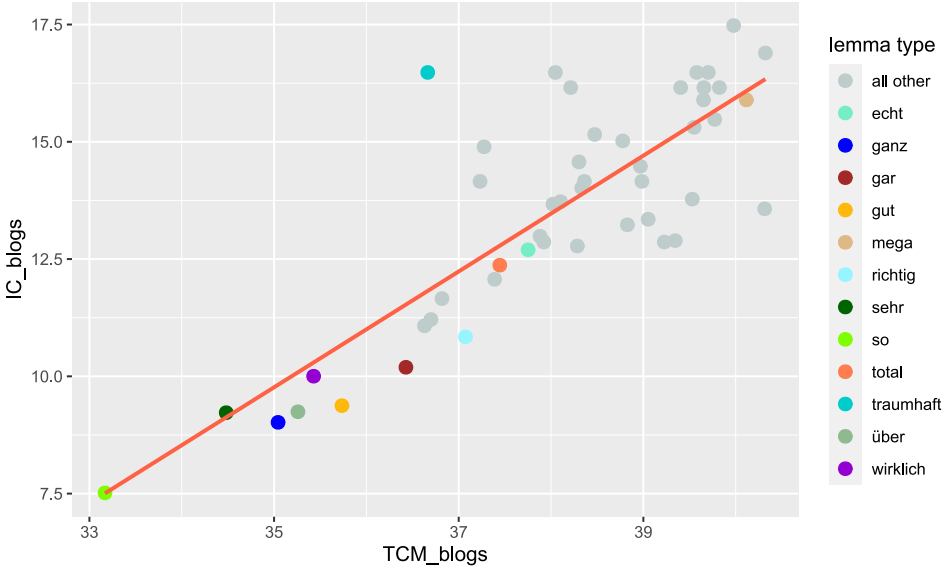

Figure 1: Information (TCM by IC) of single intensifiers in blogs

The TCM values are considerably higher in general for blogs than for tweets. This is due to the different context sizes. While the context in the tweets data set is restricted to only a single tweet in each case, the context for the blogs consists of several sentences. This means that the topic model on which TCM is based has more information available for training and can thus be more specific.

Despite these differences in TCM values, the two patterns in Figure 1 and Figure 2 resemble each other. The same frequent intensifiers can be found in their left parts. The patterns get more diffuse moving to the less frequent intensifiers in the second position in the stack. We also included *wirklich* ('really') that we considered a focus particle. This status as a focus particle seems to be supported by the different relative position it has in the two scattergrams, although they fit the linear pattern.

The scatter graphs indicate that the two measures are less closely related in the tweets subcorpus. It might be that the more diffuse pattern in tweets is (partly)
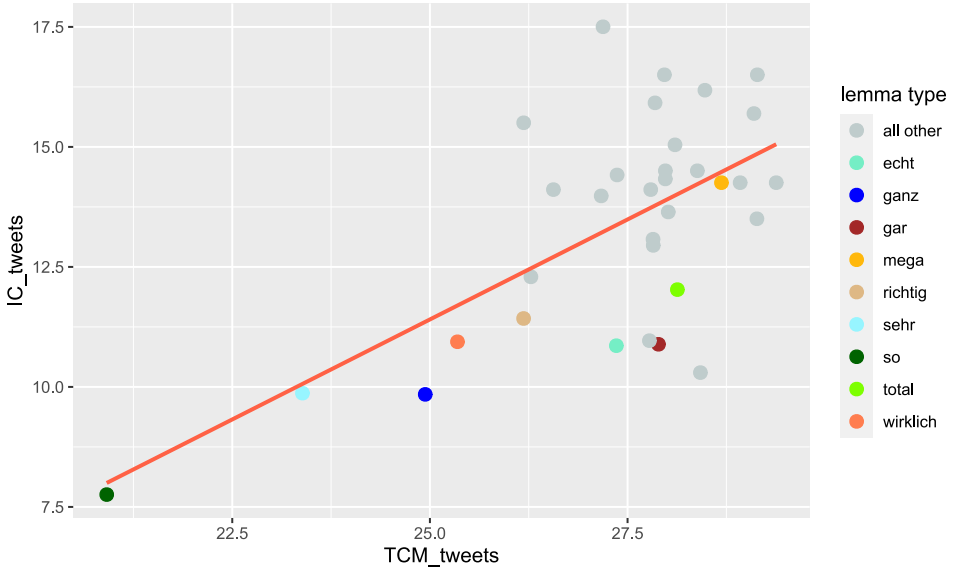
Figure 2: Information (IC by TCM) of single intensifiers in tweets

related to the size of the database, a smaller database means more noise, in particular since the number of intensifiers is also lower in the tweets. To investigate the relation between IC and TCM more closely, we not only need more data, we also have to estimate the additional value of the TCM index more precisely. We start doing so in the next subsection by looking at the stacks.

4.3 Information values in intensifier stacks

Both IC and TCM capture the information, and in our understanding, the expressiveness, of an intensifier. The first question we asked ourselves was whether the two types of information values TCM and IC correlate in the intensifiers occurring in the stacks. Since stacks with three intensifiers are extremely rare in our data sets (only three occurrences), we focused solely on stacks with two intensifiers.

We determined the correlation between the two types of information values of the first intensifier in a stack (TCM1–IC1) and the values of the second intensifier in the stack (TCM2–IC2). Table 2 shows that the correlations between the information values for both intensifier positions are positive; the correlations (Pearson) in the Twitter data set are higher than in the blogs.

In addition, the correlation is much lower across the entire data set, indicating a clear register difference between the blogs and tweets. This is due to higher

Table 2: Pearson correlation between TCM and IC information values

		Pearson correlation
Data	TCM1–IC1	TMC2–IC2
all	0.46	0.52
blogs	0.66	0.72
tweets	0.83	0.71

TCM values in the blog subcorpus, since the TCM values depend on the number of different contexts and instances. This becomes clear in the scatter plots in Figures 3 and 4, which show the data points of TCM compared to IC for the first intensifier in blogs (Figure 3) and the data points of TCM and IC for the second intensifier (Figure 4).
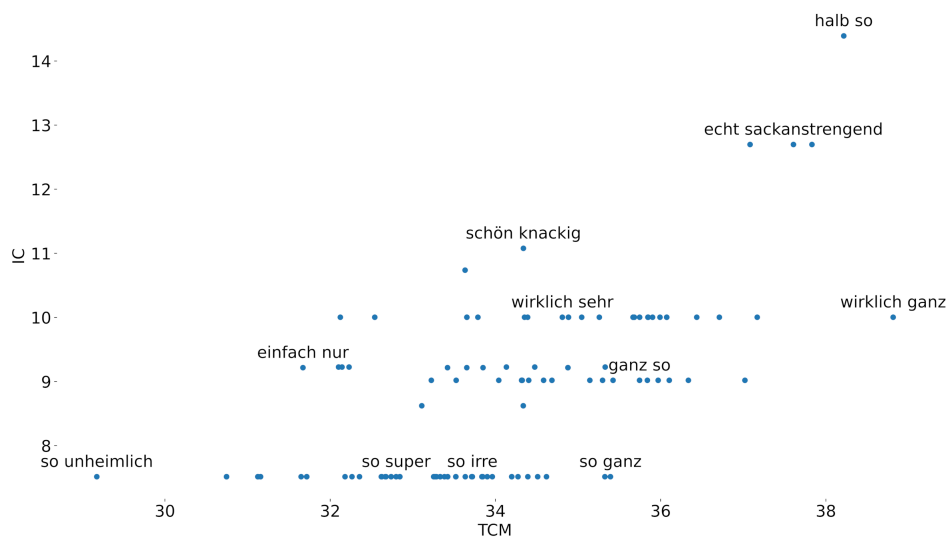
Figure 3: Scatterplot of the blog data set, using the *IC/TCM* values from the first intensifier in the stack

Figures 3 and 4 clearly demonstrate the differences in the calculation of the actual IC and TCM values. Figure 3 shows the IC and TCM values for the first intensifier in each stack in the blog subcorpus, Figure 4 shows the values for the second intensifier. All occurrences of *so* have the same IC value, seen in the bottom row of dots in both figures, as this value depends only on the overall
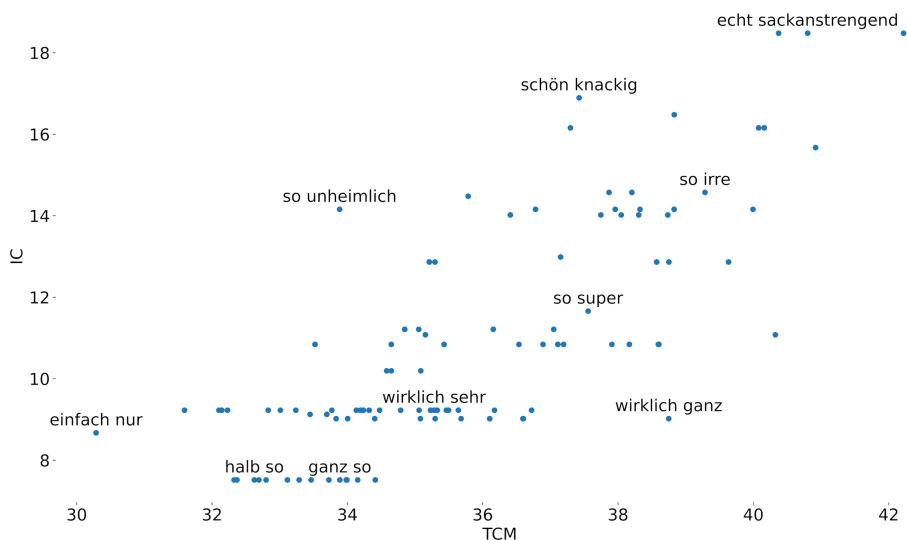
Figure 4: Scatter plot of the blog data set, using the *IC/TCM* values from the second intensifier in the stack

frequency of *so*, independent of its context. In contrast, the TCM surprisal values differ greatly between the instances, see Figure 3, with *so unheimlich* having the lowest and *so ganz* having the highest value. This is due to the fact that the topic modeling part of the TCM allows for a more fine-grained approach, enabling us to calculate a surprisal value that is text specific. The different TCM values for the same token indicate that it contributes different amounts of information to the context. It is thus not surprising that the correlation between the two information values IC and TCM is not perfect, as shown in Table 2, and that the correlation is lower for blogs, which have a greater context, when calculating the surprisal values.

Figures 3 and 4 also make it possible to compare the spread of TCM for an intensifier like *so* as the first in a stack and *so* as the second intensifier in a stack. This shows that the spread, when *so* occurs as the first intensifier in a stack, is much greater, compared to when it is only in the second position. One reason for this is that the word, being a low information intensifier, can be the start of many different intensifying phrases. This bridging function of the first intensifier towards another intensifier is discussed in Section 4.5.

4.4 Stacking intensifiers

Confirming previous studies using information theory to model intensifiers in Dutch and German (Richter & Van Hout 2020, Scheffler et al. 2023b), we observed the tendency that in stacks of intensifiers, the information values increase from left to right. Table 3 depicts the number of stacks that behave according to the predictions, i.e., with information values increasing from the first to the second intensifier for both information types TCM and IC, and of stacks with violations of these predictions (mere repetitions such as *so so* are excluded). Table 3 illustrates the tendency towards increasing information in the blog, Twitter and the combined data sets. The proportions are almost identical when using TCM and IC.

Table 3: Counts of intensifier stacks following the predicted, increasing order of TCM and IC information values, compared to violations of this predicted order

Data		Predicted	Violations	% Pred.	% Viol.
all	TCM	92	51	0.64	0.36
	IC	92	51	0.64	0.36
blogs	TCM	78	32	0.71	0.29
	IC	69	41	0.63	0.37
tweets	TCM	26	7	0.79	0.21
	IC	25	8	0.76	0.24

We follow the argumentation by Scheffler et al. (2023b) that some number of violations of the predicted stacking order are expected due to the creative choices of the authors, but also due to the inherent uncertainty of our estimates of the information values. Both information values are computed within our relatively small corpus and thus may not fully reflect the true underlying information values that drive authors' choices when selecting intensifiers. Scheffler et al. (2023b) propose that this view can be confirmed by comparing the difference in information values between predicted stacks and violations: Violations of the stacking order should tend to exhibit small differences in information values between the two intensifiers, indicating that the involved intensifiers are similar to each other, and variation in their order may be due to chance or to our estimation error. On the other hand, when there are large differences in information value in an intensifier pair, authors should be much more likely to choose the predicted order.

In the current study, we again find this picture confirmed, as Figure 5 illustrates. The figure depicts the density distributions of the $UID_{wordwise}$-values, i.e., the distribution of the average information change between the two intensifiers in a stack. The x-axis shows the $UID_{wordwise}$-values, the y-axis the normalized relative frequency of each value. Let *min* be the minimum of $UID_{wordwise}$ in a data set and the maximum = 0. The area under the curve is determined by the integral $\int_{min}^{0} x \, \mathrm{d}x = 1$.

(a) Information Density of IC in the blog data

(b) Information Density of TCM in the blog data

(c) Information Density of IC in the Twitter data

(d) Information Density of TCM in the Twitter data
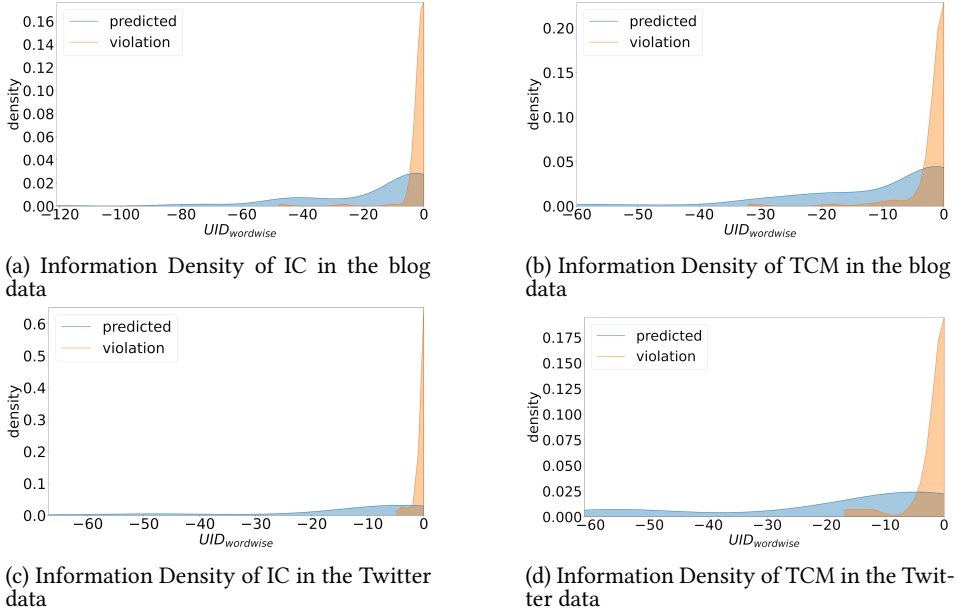
Figure 5: $UID_{wordwise}$ density plot of predicted information structures (TCM values and IC values) (blue) and of violations (orange)

In the stacks with violations, the change in information values as captured by $UID_{wordwise}$ tend to be closer to zero than in stacks with the predicted order. This indicates closer information values within violations than in the stacks with predicted increasing information. The differing information density between predictions and violations holds both for the surprisal values TCM and for the IC values. This means that violations of the predicted increase in information values are more likely when the word-wise information changes are rather small. This replicates the previous finding by Scheffler et al. (2023b) on our new dataset in this study, for both information values. When the difference between the information values of two intensifiers in a stack is very small, the preference for their ordering is not very strong and they may thus occur in reverse order.

4.5 Can UID explain intensifier choice and order?

Scheffler et al. (2023b: 14) proposed "to interpret established intensifiers as the 'glue' that holds the whole set together: in stacks, established intensifiers prepare the language processor for more expressive, innovative intensifiers to follow". The conjecture was that adding a frequent intensifier with little own information contribution smooths the information flow within a sentence by decreasing the overall change in $UID_{wordwise}$. In this paper, we compute information values for all words in a sentence and can thus systematically test this conjecture, by comparing the attested version of a sentence with two intensifiers with two communicatively equivalent variants: (i) the shortened sentence where the first
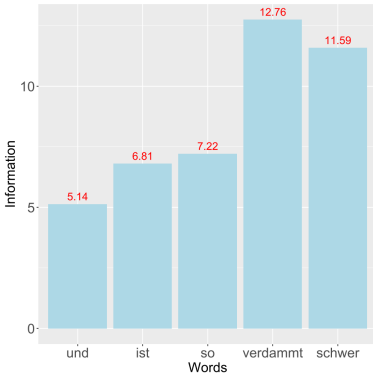
(a) Flow of IC in the phrase *und ist so verdammt schwer*
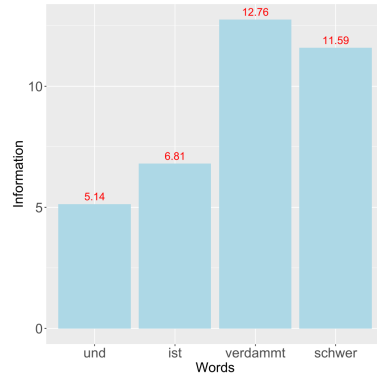
(b) Flow of IC when *so* is omitted
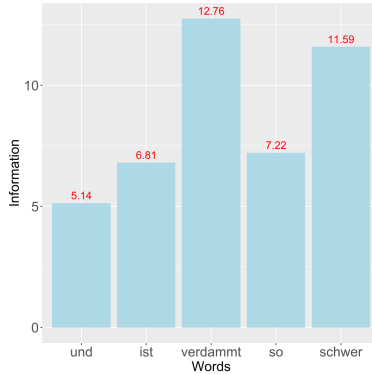
(c) Flow of IC when the order of *so* and *verdammt* is reversed

Figure 6: Example flow of IC in (a) complete, (b) shortened, and (c) reversed stacks, for the phrase *und ist so verdammt schwer* ('and is so damn hard')

intensifier is omitted (shortened), and (ii) the sentence where the two intensifiers are reversed in order (reversed).

To give an example, in Figure 6, the flow of information in a stack with the extremely common intensifier *so* is depicted in Figure 6a. Figure 6b shows how the flow of information changes when *so* is omitted: the gap between the copula verb and the remaining second intensifier is somewhat larger, but the effect is not very pronounced. Finally, Figure 6c shows the flow of information when both intensifiers are reversed, which is much more bumpy and thus less optimal with respect to the UID hypothesis.

We computed the wordwise information density for all original sentences in the corpus, as well as in the shortened and reversed variants. Figure 7 shows the information density distribution over all sentences in the blog subcorpus, compared to the shortened and reversed variants, for both the IC (top) and TCM (bottom) models. We use the blog data because it provides fully coherent texts which allow for better estimates of the information values. However, the Twitter data shows similar behavior and the corresponding graphs for the Twitter subset and the complete data set are shown in the Appendix.

We applied an ANOVA to compare the three types of outcomes. For IC we obtained a significant result $(F(2, 158) = 13.037, p < 0.001)$ as Figure 7a illustrates. Posthoc comparisons (Bonferroni) showed a significant difference between regular and reversed (higher scores) versus shortened (lower scores). For TCM we obtained a significant result as well $(F(2, 158) = 34.712, p < 0.001)$, see Figure 7b. Moreover, all posthoc comparisons (Bonferroni) were significant $(p < 0.01)$, going from regular (higher scores) via reversed to shortened (lower scores). The density plots for the Twitter data and the Twitter and blogs together (complete data) can be found as Figure 9 and Figure 10, respectively, in the Appendix. The number of observations in the Twitter data is rather low $(n = 24)$. The result is thus not significant for IC $(F(2, 46) = 2.752, p = 0.076)$. The result for TCM is significant $(F(2, 46) = 5.312, p = 0.013)$. The only posthoc significant difference is between regular (higher scores) and shortened stacks (lower scores). The results for the complete data are similar to the results of the blog data. For IC we obtained a significant result $(F(2, 206) = 16.074, p < 0.001)$, with posthoc outcomes similar to the blog data. For TCM we obtained a significant result $(F(2, 206) = 32.0293, p < 0.001)$, as well, with the posthoc outcomes similar to the blog data.

It can be seen that the average change in density is very small, since most words stay the same within all variants of each sentence, and the change by removing or reordering the frequent intensifier is divided by the number of words
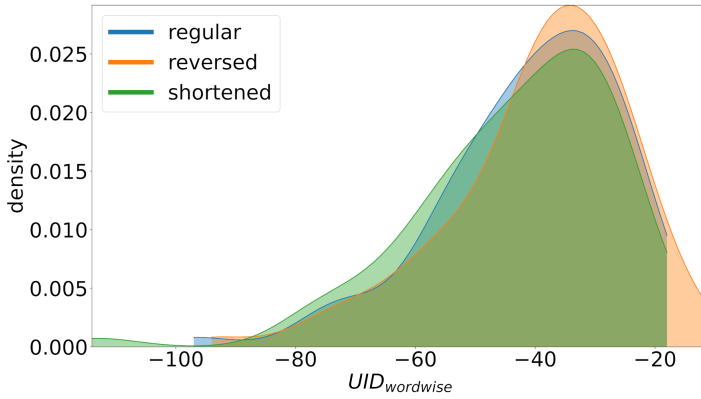
(a) Information Density of IC in the blog data, compared to shortened and reversed stacks
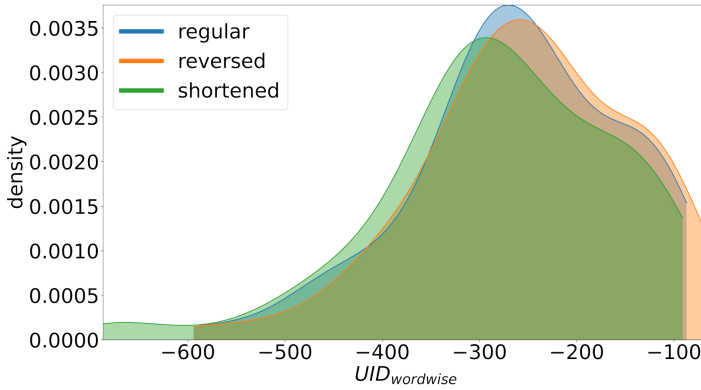
(b) Information Density of TCM in the blog data, compared to shortened and reversed stacks

Figure 7: Average per-word information density in complete sentences compared to their shortened and reversed variants, in the blog subcorpus

in the sentence. However, the observable small change is in the expected direction, at least for the shortened version: The alternative utterances show a decrease in the smoothness of the information flow compared to the actually attested utterances. Note that the information values used here are based on the relatively small Twitter and blog corpus. This means that the context for semantic surprisal TCM is very limited. We expect thus a more precise estimate of the information values based on a larger corpus to lead to a clearer result.

4.6 A quick note on ad-DP intensifiers

An interesting observation about German intensifiers is that they can occur in a DP-external position when modifying an attributive adjective or a noun, as shown in the constructed examples in (12), and discussed by Gutzmann & Turgay (2014), Willich & Politt (2023)[6]. Note that the English translations exhibit the intensifiers in the expected, pre-adjectival DP-internal position.

(12)   a. *Das war voll das schöne   Wochenende!*
           That was INT the beautiful weekend
           'That was a super nice weekend!'

       b. *Ich habe voll das schöne   Bild   gemalt.*
           I   have INT the beautiful picture drawn
           'I have drawn such a nice picture.'

In this construction, the intensifier *voll* ('fully') precedes the determiner of the DP; this determiner is typically definite (but indefinites are also possible). In terms of meaning, the sentences in (12) are denotationally equivalent to their canonical variants in (13) (Gutzmann & Turgay 2014). Gutzmann & Turgay (2014) make the important observation that the noun is interpreted as indefinite, even if a definite determiner is present.

(13)   a. *Das war ein voll schönes   Wochenende!*
           That was a   INT beautiful weekend
           'That was a super nice weekend!'

       b. *Ich habe ein voll schönes   Bild   gemalt.*
           I   have a   INT beautiful picture drawn
           'I have drawn such a nice picture.'

The DP-external intensifier variant in (12) is understood as more expressive and stronger than the canonical, DP-internal variant (13). We believe that the information theoretic measures introduced in this paper can explain this difference. Comparing the TCM-surprisals, that is, the information values derived from topic contexts, for the sentences in (12) with their corresponding (13) variants, we observe that the canonical versions have a much smoother information profile (see Figure 8). Thus, the speaker seems to be encoding additional surprisal by using the marked ad-DP intensifier construction. Our blog/tweets corpus contains no instances of this construction, but it would be very worthwhile to carry out a larger corpus study to corroborate this observation.

---

[6]We thank an anonymous reviewer for encouraging us to discuss this phenomenon.
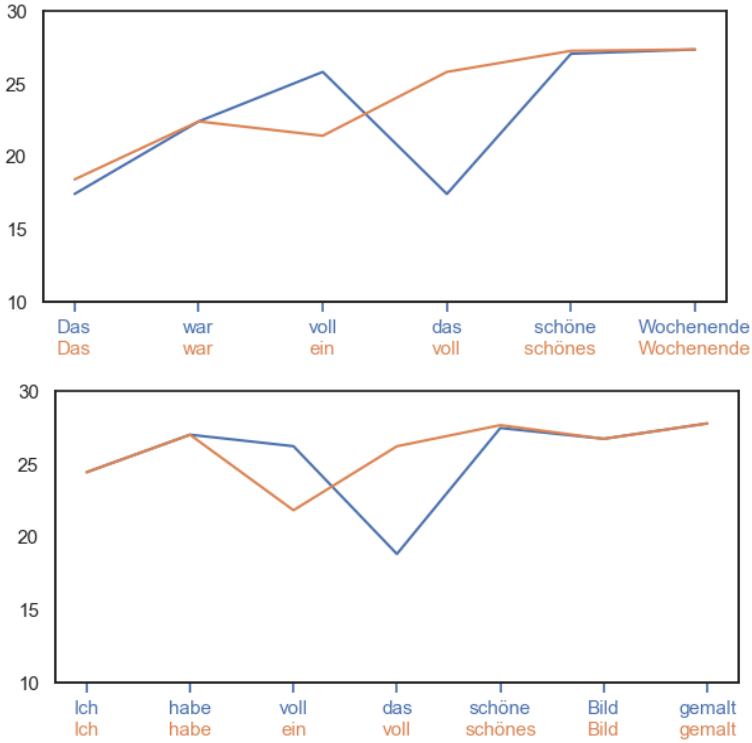
Figure 8: TCM information profiles for ad-DP and canonical sentences (12–13)

5 Discussion

In this paper, we apply information theoretic analyses to a new dataset of manually annotated intensifiers in two media, tweets and blog posts, to account for their expressive meaning and characterize the variability of available intensifying items. We confirm previous information theoretic predictions about the distribution of intensifiers in a new medium, blog posts. These results hold even though the data we report on here is no longer restricted to intensification of predicative adjectives, but includes intensifiers of adjectives in other positions, as well as nominal and verbal intensification. We show that the context-free information value IC and the context-dependent information measure TCM are highly correlated for the entire dataset. Both measures can account for the distribution of intensifiers in our data, as well as their tendency to occur in order of increasing information in a phrase, giving credence to the notion that intensifiers form a common word class, even across syntactic and semantic differences.

Since both values measure the surprisal of a word, it is expected that they are related. However, they are not identical, since IC is a static value computed based on overall word frequency, while TCM is dependent on the context of a specific instance (here, dependent on the document and its topic distribution). The TCM values are more reliable for the blog posts than the tweets. This is most likely due to the size of the contexts, which for the tweets only include the single tweet, which means an extremely small context for topic detection. We assume that the relationship between TCM and IC will change further when the context is increased again, i.e., larger than for the blogs and, for example, includes longer documents or encompasses an entire subcorpus. The prediction is that for large coherent contexts, TCM will yield a "better", that is more accurate, information value for words, which could manifest itself in a lower number of violations, i.e., deviations from the expected increase in information values in stacks of intensifiers.

We explicitly tested the conjecture from previous work, that stacked intensifiers tend to be ordered from least to most expressive within a phrase, on our entire corpus of sentences with intensifiers. We systematically compared sentences with stacked intensifiers with the communicatively equivalent alternatives (i) containing only the last, most expressive intensifier, or (ii) containing both intensifiers in reverse order. We computed information profiles for both variants and compared them to the actually attested variant. Our results show that the attested sentences exhibit a smoother information contour (higher uniformity of information density). This supports the Uniform Information Density hypothesis, which states that smoother information flow is preferred in communication, when keeping other factors equal. This analysis may serve to shed light on the puzzle why writers bother to add further, redundant intensifiers beyond the most expressive one: Their function is to introduce the phrase, alert the reader to the word class which follows, and smooth the information flow.

Acknowledgements

References

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56. DOI: 10.1177/00238309040470010201.

Beltrama, Andrea. 2022. Just perfect, simply the best: An analysis of emphatic exclusion. *Linguistics and Philosophy* 45. 321–364. DOI: 10.1007/s10988-021-09326-x.

Bennett, Erin D. & Noah D. Goodman. 2018. Extremely costly intensifiers are stronger than quite costly ones. *Cognition* 178. 147–161. DOI: 10.1016/j.cognition.2018.05.011.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bolinger, Dwight. 1972. *Degree words*. De Gruyter Mouton. DOI: 10.1515/9783110877786.

Breindl, Eva. 2007. Intensitätspartikeln. In Ludger Hoffmann (ed.), *Handbuch der deutschen Wortarten*, 397–422. De Gruyter. DOI: 10.1515/9783110217087.397.

Claudi, Ulrike. 2006. Intensifiers of adjectives in German. *Language Typology and Universals* 59(4). 350–369. DOI: 10.1524/stuf.2006.59.4.350.

Collins, Michael Xavier. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research* 43(5). 651–681. DOI: 10.1007/s10936-013-9273-3.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank & Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH)*, 76–84. Osaka: The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/W16-4011.

Fenk, August & Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis – Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie* 27(3). 400–414.

Fortin, Antonio. 2011. *The morphology and semantics of expressive affixes*. Oxford University. (Doctoral dissertation).

Fortin, Antonio & Franz Rainer. 2022. Evaluative suffixes. In Adam Ledgeway & Martin Maiden (eds.), *The Cambridge handbook of Romance linguistics*, 434–458. Cambridge: Cambridge University Press. DOI: 10.1017/9781108580410.017.

Gutzmann, Daniel & Katharina Turgay. 2014. Expressive intensifiers and external degree modification. *Journal of Comparative Germanic Linguistics* 17(3). 185–228. DOI: 10.1007/s10828-014-9069-3.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*. https://aclanthology.org/N01-1021.

Hale, John, David Lutz, Wen-Ming Luh & Jonathan Brennan. 2015. Modeling fMRI time courses with linguistic structure at various grain sizes. In Tim O'Donnell & Marten van Schijndel (eds.), *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, 89–97. Denver: Association for Computational Linguistics. DOI: 10.3115/v1/W15-1110.

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

Kölbl, Max, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf & Tariq Yousef. 2020. Keyword extraction in German: Information-theory vs. deep learning. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence – Volume 1: NLPinAI.* 459–464. DOI: 10.5220/0009374704590464.

Kölbl, Max, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf & Tariq Yousef. 2021. The semantic level of Shannon information: Are highly informative words good keywords? A study on German. In Roussanka Loukanova (ed.), *Natural language processing in artificial intelligence – NLPinAI 2020*, vol. 939 (Studies in Computational Intelligence (SCI)), 139–161. Springer International Publishing. DOI: 10.1007/978-3-030-63787-3_5.

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

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt & T. Hoffman (eds.), *Advances in neural information processing systems 19 (NeurIPS 2006)*.

Philipp, J. Nathanael, Max Kölbl, Yuki Kyogoku, Tariq Yousef & Michael Richter. 2022. One step beyond: Keyword extraction in German utilising surprisal from topic contexts. In Kohei Arai (ed.), *Intelligent computing*, 774–786. Cham: Springer International Publishing. DOI: 10.1007/978-3-031-10464-0_53.

Philipp, J. Nathanael, Michael Richter, Erik Daas & Max Kölbl. 2023. Are idioms surprising? *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*.

Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2). 165–198. DOI: 10.1515/TL.2007.011.

Richter, Michael & Roeland Van Hout. 2020. Ranking Dutch intensifiers: A usage-based approach. *Language and Cognition* 12(2). 343–359. DOI: 10.1017/langcog.2019.50.

Scheffler, Tatjana, Lesley-Ann Kern & Hannah Seemann. 2022. The medium is not the message: Individual level register variation in blogs vs. tweets. *Register Studies* 4(2). 171–201. DOI: 10.1075/rs.22009.sch.

Scheffler, Tatjana, Lesley-Ann Kern & Hannah Seemann. 2023a. Individuelle linguistische Variabilität in sozialen Medien. In M. Kupietz & T. Schmidt (eds.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022* (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr.

Scheffler, Tatjana, Michael Richter & Roeland Van Hout. 2023b. Tracing and classifying German intensifiers via information theory. *Language Sciences* 96. DOI: 10.1016/j.langsci.2022.101535.

Schmidt, Jessica. 2022. Do intensifiers lose their expressive force over time? In Remus Gergel, Ingo Reich & Augustin Speyer (eds.), *Particles in German, English, and beyond.* Amsterdam: John Benjamins. DOI: 10.1075/slcs.224.03sch.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Stratton, James M. 2020. Adjective intensifiers in German. *Journal of Germanic Linguistics* 32(2). 183–215. DOI: 10.1017/S1470542719000163.

Tagliamonte, Sali A. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12(2). 361–394. DOI: 10.1017/S1360674308002669.

Tribus, Myron. 1961. Information theory as the basis for thermostatics and thermodynamics. *Journal of Applied Mechanics* 28(1). 1–8. DOI: 10.1115/1.3640461.

Willich, Alexander & Katja Politt. 2023. *"Du hast eh voll den guten Musikgeschmack!" – Underdetermination and grammaticalization of NP-external intensifying constructions in German.* Talk presented at ICCG12. http://tiny.cc/18v7vz (30 November, 2023).

Zadeh, Lotfi A. 1972. A fuzzy-set-theoretic interpretation of linguistic hedges. *Journal of Cybernetics* 2(3). 4–34. DOI: 10.1080/01969727208542910.

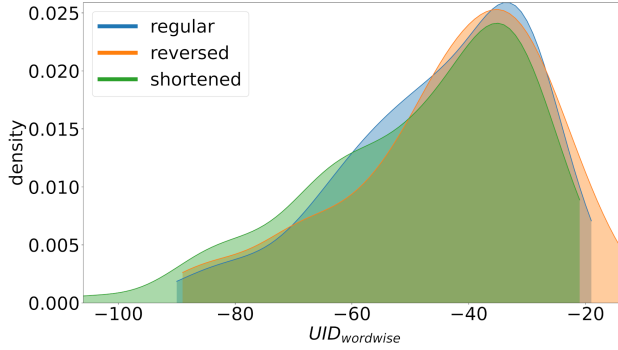# Appendix A Information density in shortened and reversed stacks
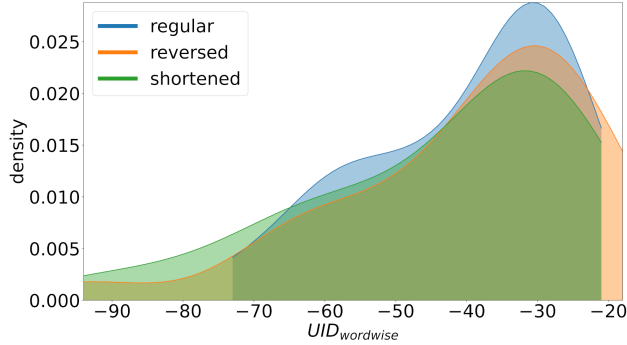
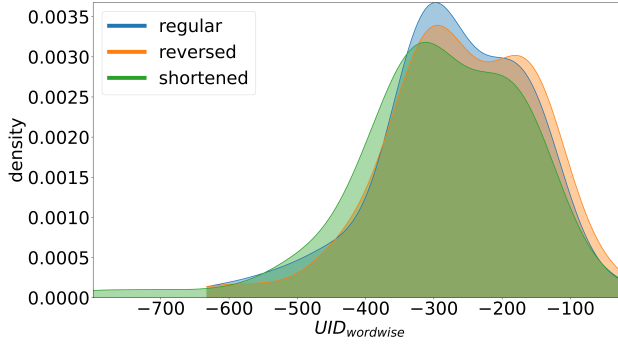(a) Information Density of IC in the Twitter data, compared to shortened and reversed stacks
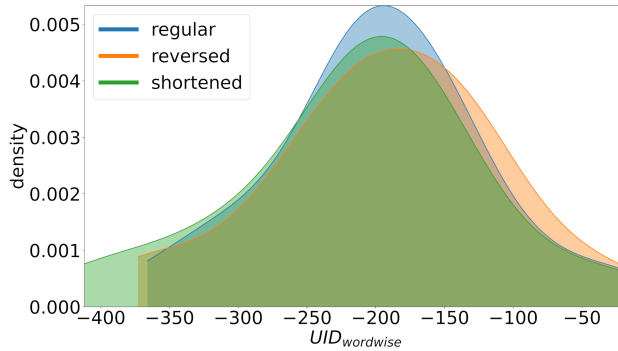
(b) Information Density of TCM in the Twitter data, compared to shortened and reversed stacks

Figure 9: Average per-word information density in complete sentences compared to their shortened and reversed variants, in the Twitter sub-corpus.

(a) Information Density of IC in the complete data, compared to shortened and reversed stacks

(b) Information Density of TCM in the complete data, compared to shortened and reversed stacks

Figure 10: Average per-word information density in complete sentences compared to their shortened and reversed variants, in the complete corpus (tweets and blogs)