The medium is not the message

Tatjana Scheffler, Lesley-Ann Kern, Hannah Seemann

The medium is not the message

Individual level register variation in blogs vs. tweets

Tatjana Scheffler[1], Lesley-Ann Kern[2] & Hannah Seemann[1]
[1] Ruhr-Universität Bochum | [2] Philipps-Universität Marburg

Linguistic expressions in social media vary along many axes, including author style, the specific medium and its affordances, and others. In this paper, we argue that different registers must be distinguished within social media and that register should be included as an important factor independent of (social) medium in analyses of variable linguistic phenomena. We introduce a new German cross-media corpus, consisting of blog posts and tweets from the same 44 authors. We define the registers as 'Informative', 'Narrative', and 'Persuasive', based on situational characteristics of the texts. We then correlate the registers with two variable linguistic phenomena: German modal and intensifying particles. In each case, we document considerable inter- and intraindividual variation in the expressions used and their frequency across texts. The statistical analysis shows that the register grouping corresponds more closely to linguistic similarities between texts than the grouping by medium does.

Keywords: register, social media, modal particles, intensifiers, German, corpus linguistics

1. Introduction

Texts in social media differ from one another in idiosyncratic ways. In addition, linguistic style differs in a systematic way between texts along several axes, including medium, register, but also genre or topic, as well as individual author preferences (Wolfram 2006; MacKenzie 2019; Schleef 2021). In this paper, we demonstrate the effect of systematically distinguishing between the effects of personal style, medium, and register in the analysis of social media data in two case studies of German particles. We follow Biber and Conrad (2019) in distinguishing 'register' from 'genre' and 'style' by highlighting different kinds of lin-

https://doi.org/10.1075/rs.22009.sch | Published online: 27 October 2022
Register Studies 4:2 (2022), pp. 171–201. ISSN 2542-9477 | E-ISSN 2542-9485

guistic variation: Whereas 'register' often groups different texts just by observing non-linguistic factors in the situational context they are used in and attributing functional characteristics to certain linguistic features, 'genre' describes a set of linguistic features that is conventionally attributed to a variety or a type of text independent from the context of use. Biber and Conrad define 'style' similarly to register – "analyzing the use of linguistic features that are common in texts" (Biber & Conrad 2019: 2) – with the difference that the choice of words here expresses individual stylistic or aesthetic preferences and not functional characteristics of the chosen linguistic features. To a certain extent, the choice of medium does have an influence on 'register', 'genre' and 'style'. We define 'medium' in the context of this paper as a specific communication channel via which the user can express themselves, which is implemented in a certain way (e.g., in a technological system), and which comes with certain affordances. In our case, the two media we discuss are weblogs ('blogs') and Twitter posts.

We achieve our proposed differentiation between the effects of medium and register by creating a new corpus that is controlled for genre and stratified by authors: As genre, we choose personal and family life in the context of parent blogging. We collect both blog posts and tweets in the German language from the same 44 authors who are active on both social media platforms. This enables us to study the interplay of stable individual characteristics of language use and variable adaptation to the medium (blog vs. tweet), using corpus pragmatic means.

Our main research question is how to decide whether specific types of discourse level linguistic variability can be attributed to variability within the individual (following the orderly heterogeneity assumption: Honeybone 2011; Wolfram 2006), or to differences between groups of authors, registers, or medium conventions. Our unique dataset opens up new ways to approach this question, as each type of variability makes different predictions about the distribution of variable linguistic expressions in our corpus. For this paper, we select two main phenomena that are known to be variable and have been studied as register markers and across media before: German intensifying particles and modal particles. Both types of particles are assumed to encode conceptual orality in the sense of Koch & Oesterreicher (1985), but occur relatively frequently in digital writing (Scheffler 2017). This shows that particles are rather a phenomenon of conceptual orality than of spoken language. Following this idea, particles should be more frequent in tweets than in blog posts, since Twitter is more conversational and less conceptualized than blog posts (Scheffler 2017; Storrer 2013). But there might be more factors in play than conceptual orality. For example, different communicative contexts could lead to variation of particle use within one medium. We thus assume that these particles are of special interest in the context of our social media corpus, in order to establish which specific properties of a certain text promote or

suppress the use of these features. Therefore, we analyze how these two linguistic variables are connected to register and medium variation.

First, we investigate the use of German modal particles (*ja, doch, wohl*, etc.) These are frequent discourse markers typical of informal language which express features of the discourse context, such as author's prior knowledge, bias, or certainty (Döring 2016). We track the highly individual use of these particles by authors across two media and three registers in order to pinpoint the effect of the register and the medium on their use. Our results show that individual authors show highly idiosyncratic usage patterns of particles, but that the overall frequency of particle usage also differs significantly between media and between registers: particles occur more often in blog posts, and most frequently in narrative texts where authors relate personal stories. We will present examples of modal particles in translation and their meaning in detail in Section 2.4.1. Second, we look at intensifying particles (for short, 'intensifiers') like *so* (*so*), *sehr* (*very*) – a type of item typical of spoken language, and common in informal interaction on social media. More information concerning the intensifiers we investigated can be found in Section 2.4.2. We show that the frequency with which the intensifier *so* is used differs between registers. *So* is used more often in narrative and persuasive texts, which exhibit more emotional involvement of the author. This result is mirrored for all intensifiers in a small subcorpus consisting of the texts from six authors. Our analysis indicates that register is an orthogonal category cutting across (social) media to characterize text types. We define three registers present in our social media corpus ('Informative', 'Narrative', and 'Persuasive'), which we are able to reliably annotate based on situational characteristics of the texts. We show that the registers correlate in turn with variation in the linguistic expressions studied (modal particles and intensifiers), and that the linguistic features we analyzed reflect register distinctions more closely than medium distinctions in our corpus. Thus, we argue that the medium is not the (whole) message (cf. McLuhan 1964): We must distinguish between different registers within and across social media, and take register into account when analyzing variable linguistic phenomena in digital texts.

All tables with annotation results and frequencies as well as the R script for analysis can be accessed via the Open Science Framework.[1]

1. https://osf.io/3j2d6/

2. Register and social media

2.1 Defining register

The term 'register' defines the situational context of communication (Biber & Conrad 2005). Registers are thus primarily differentiated in non-linguistic terms, although linguistic variability can be observed when comparing different registers with each other (Argamon 2019; Biber & Conrad 2005). Contributing factors to register variation (amongst others) are situational characteristics such as the setting and channel of communication, discourse participants and their relationships, as well as the purpose of communication (Biber & Conrad 2019), which might require certain linguistic features in order to be successful. Register variation in social media has been analyzed punctually (e.g. Clarke and Grieve's (2019) analysis of stylistic variation in former US-president Donald Trump's Twitter posts) or within a very broad analysis of online speech, namely as part of a multi-dimensional analysis on content of the searchable web in general (Biber & Egbert 2016). Bildhauer, Pankratz and Schäfer (2021: 18) argue that it is impossible to determine universal registers in German and instead propose an exploratory method for defining registers in a data driven way according to the occurrences of linguistic features in the corpus. For this study, we will not follow this approach since it may lead to circularity in argumentation: We want to investigate how linguistic phenomena such as intensifiers and particles vary across media and registers, which means that we cannot use these linguistic features to differentiate between registers in the first place. Instead, we must rely on the extra-linguistic conditions within which utterances are made to define registers in our social media corpus.

2.1.1 *Describing the felicity conditions of register variation*

The felicity conditions of utterances differ according to the communication situation (Argamon 2019); therefore, register variation can be observed when communicative situations change. This might be the case when the channel of communication changes (e.g., when switching from written to oral communication), when new participants join the conversation, or when the function of the message changes (e.g., from informational to persuasive in promotional communication). These changes in register also closely reflect the affordances a medium offers its user. 'Affordances', a term originally coined by Gibson (2014), can in this context be described as properties an artifact (e.g., a social network such as Twitter) offers to an actor (the user) and which contribute to the way actor and artifact can interact (Gibson 2014; Zhao, Liu, Tang & Zhu 2013). Zhao et al. (2013) discuss in particular perceived affordances which in the context of social media

primarily concern aspects such as usability or user experience design. In the case of our social media corpus, perceived affordances can be characterized as the ways a user assumes they can communicate via a social medium (Twitter vs. personal blog). Affordances are an important factor when analyzing register in social media contexts since aspects such as interactivity are often given by the artifact, i.e., the social network in question, and profoundly influence the way users interact not only with the network itself but also with each other. In addition to the affordances of the medium, multiple other factors come into play when it comes to describing register variation. Thus, the approach to describing different registers must be multidimensional and hence consider various aspects which taken together produce a register. Following Biber (1993), we use a multidimensional approach to register which assumes variation of linguistic factors, whose combination enables the development of new registers. To determine the register used in a text, Biber and Conrad (2019) propose to first identify and analyze the situational characteristics of the communicative context by looking at aspects such as participants, relations among participants, channel, processing circumstances, setting, communicative purposes and topic. After finding and labeling these categories, an analysis of the typical linguistic features of the register follows, defining the register in linguistic parameters such as "[…] words or grammatical characteristics that are (1) pervasive […] and (2) frequent […]" (Biber & Conrad 2019: 54). Linguistic features that do not occur in other registers are classified as 'register markers' (Biber & Conrad 2019: 54). Since size and variety of a corpus have a strong influence on the outcome of corpus linguistic research (Stefanowitsch 2020), we are aware that the occurrences of linguistic features alone do not define our corpus in terms of register specification. We will focus instead on the "underlying dimensions of variation" (Biber 1993: 228), because linguistic features appear in different registers where they might have different meanings according to the communicative context, purpose or participants. In addition, one must avoid defining a register by the linguistic phenomena that occur within it when the goal is to later analyze how certain linguistic features vary between registers. Thus, we base the register distinctions developed below mainly on extralinguistic properties of the communicative context rather than a combination of linguistic variables (see also Argamon 2019).

2.2 Register variation across media

The choice of linguistic features within computer mediated communication (CMC) and the resulting development of 'new' registers in digital communication has been studied in sociolinguistic contexts, yet there has not been a linguistic agreement concerning the interaction of medium and register variation. While

various variationist linguists consider the medium irrelevant for systemic language change (cf. Labov (2010) on mass media), Androutsopoulos (2014: 3) describes this perspective as 'unsatisfactory'. We generally agree with the need of taking the medium into account when analyzing linguistic variation, yet the medium itself must not be equated with the register and therefore is not the main factor when analyzing (systemic) language variation and change. This is for example the case in Tagliamonte's (2016: 2–3) analysis of private CMC of Canadian students where CMC in general is described as including a "diverse range of different registers", but eventually register and medium are defined as synonyms and 'instant messaging', 'texting on phones' and 'email' are identified as "different CMC registers". Biber and Egbert (2016: 96) choose a similar perspective when characterizing internet language in accordance with "new internet registers, such as blogs, Facebook/Twitter posts, and email messages". In these cases, medium and register are treated as equivalent and there is no further distinction for registers within a medium.

Following the brief definition of 'medium' we give in Section 2.1, we distinguish between 'medium' (a channel of communication, e.g., Twitter or a personal weblog) and 'register' (a communicative situation requiring certain linguistic characteristics).

2.3 Developing registers for personal narratives on social media

While perceived affordances may be used to distinguish between different social media, it is not clear whether each medium or platform constitutes its own (unique) register. When considering the framework by Biber and Conrad (2019) mentioned above, defining register too closely in accordance with the medium is insufficient: The same social medium can be used in different communicative contexts, which may result in using different registers within the same medium, e. g. using Twitter posts to talk to close friends vs. using Twitter to spread news about an upcoming election. Clarke (2022) shows several additional communicative contexts of communication on Twitter, e. g. promoting one's own or opposing other user's content/persona and linguistic variation between these contexts.

Therefore, we argue that register variation within social media (i) exists and (ii) can be detected by observing the interplay of different linguistic features that usually belong to a certain type of communication, as well as situational characteristics that include extra-linguistic factors as discussed above. Following these considerations, we have identified three register dimensions ('Informative', 'Narrative', and 'Persuasive') based on several situational characteristics (purpose, topic, interactivity, and involvement) within our social media corpus. These three registers have been developed following Biber and Conrad (2019) and the model

of communicative closeness and distance by Koch and Oesterreicher (1985). We chose these three dimensions because they cover the entire corpus (see Section 3.3) and also match the content/topic of our corpus. We are aware that for different corpora, these dimensions might be insufficient and would need to be expanded.

We define the register dimensions ('Informative', 'Narrative', 'Persuasive') in our corpus with the help of the non-linguistic factors specified in Table 1.

Table 1. Register dimensions 'Informative', 'Narrative' and 'Persuasive' defined in detail

	Informative	**Narrative**	**Persuasive**
Purpose	passing on information, showing expertise, drawing attention to a topic, notifications, transfer of knowledge, announcements	reporting everyday life and experiences, authenticity, relatability, sharing experiences, diary	influencing readers, activism, politics, change, awareness, positioning, promotion
Topics	milestones in personal life, competitions, reviews, recipes, nutrition, crafting	everyday life, holidays, career, health, nutrition, personal life	parenting, nutrition, product placement, politics, gender, education, sustainability
Interactivity, reader involvement	low, rarely addresses reader, no engagement, no call to action, no reaction needed / expected	optional, can be medium to high, depending on format/level of community	high, expected, might be provocative, direct address, high level of emotionalization
Author involvement	low, rare use of *I/me*, personal position not expressed	medium to high depending on content and aspired level of relatability	high, speaking from own point of view, position visible and distinctly expressed, use of *I/me*, arguments on a personal level

The non-linguistic factors purpose, topic, interactivity, and involvement provide an analytical framework for classifying the texts in our corpus. We do not use all aspects proposed by Biber and Conrad (2019) since not every subcategory they propose is useful for analyzing our corpus. Instead, we focused on aspects that are crucial when differentiating between texts from our tweet and blog corpus since they provide the most interesting findings and also facilitated the development of our three register dimensions 'Informative', 'Narrative', and 'Persuasive'.

The communicative purpose of a text can be defined as the goal of communication, i.e., the aim the writer pursues in composing and publishing the text.

Frequent communicative goals in our parenting corpus are the transmission of information or knowledge ('Informative'), entertainment and storytelling ('Narrative') or the presentation of a certain attitude the author wants the reader to accept or even adopt ('Persuasive'). Communicative purpose and topic are closely connected as some topics tend towards certain purposes and vice versa. In our case this means that topics such as milestones in personal life are things that are typically presented (and thus informative) rather than discussed in an argumentative way (which would be persuasive). Other topics, such as recipes, are also rarely persuasive and mainly categorized as informative. When analyzing our corpus, we have on the other hand identified a range of 'controversial' topics which are more often involved in persuasive communication, such as gender, care work, nutrition or politics, and rarely just presented in an informative manner. The register dimension 'Narrative' presents a variety of topics which usually lack the controversial potential of topics in persuasive texts but are more personal or intimate than informative topics as they come from the personal life of the author.

Finally, we consider the level of involvement of both author and reader: While informative texts usually lack emotional involvement of the author and expressing the author's personal position is not necessary, e.g., in recipes or reports, narrative texts sometimes exhibit personal opinions of the author because of the personal topics that might be addressed. Persuasive communication on the other hand shows a high level of author's involvement since in order to convince the reader of a certain position, the author usually expresses their own position or uses personal experiences or examples as arguments. Author's involvement can thus be ranked from low ('Informative') via medium ('Narrative') to high ('Persuasive'). We see a similar spectrum when it comes to reader's involvement or interactivity. Informative texts can achieve their goals without the explicit interaction of readers with the text. They usually lack interactive elements such as calls to action or questions. Narrative communication can be categorized as sometimes interactive, since authors sometimes ask for opinions or feedback in the form of comments, or they encourage their readers to share experiences which are similar to or differ from the issue presented in the text. The felicity of persuasive communication, however, relies on the interaction of the readers with the content, ideally by accepting and adopting the opinions that are presented by the author.

Figure 1 shows the register dimensions visualized following Koch & Oesterreicher (1985) and their model of communicative closeness and distance. Informative texts are more distanced and factual, persuasive texts are more involved and emotional and narrative texts are in between. We apply the resulting register categories – 'Informative', 'Narrative', 'Persuasive' – to our social media corpus. In the following section, we introduce the data and the analysis of two linguistic case studies in relation to these register dimensions.

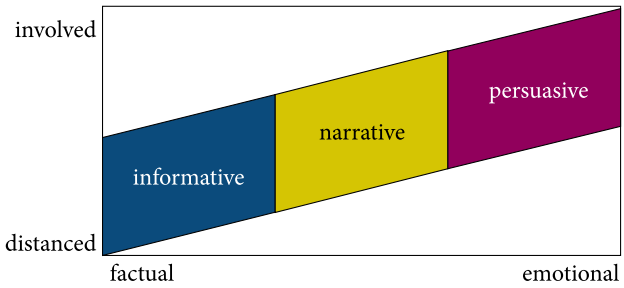

Figure 1. Distribution of register dimensions in terms of involvement and emotionality

2.4 Exemplary linguistic markers for register variation: German modal particles and intensifiers

As has been established before, register variation is due to non-linguistic factors within communicative situations but manifests itself in the use of certain linguistic features. We analyze register variation by researching the use of German modal and intensifying particles within our corpus since they are used differently in different communicative situations (spoken vs. written and formal vs. informal communication) and are thus represented in different quantities within our social media corpus.

2.4.1 *Modal particles*

In the first case study, we analyzed the use of German modal particles and its interaction with medium, register, and individual variation. We define 'modal particles' as noninflected sentence modifiers that do not affect the truth conditions of a sentence, but are used to express the author's attitude towards the denoted proposition (Bross 2012; Zimmermann 2011), or make assumptions concerning shared knowledge (the so-called 'common ground') of author and reader (Zimmermann 2011). The meaning of modal particles is very difficult to describe and varies in different contexts. Most German modal particles do not have an exact match in other languages[2] (Degand, Cornillie & Pietrandrea 2013: 7). Modal particles are generally assumed to be typical of speech and their use varies greatly between individuals and linguistic modes (Thurmair 1989). See Zimmermann (2011) or Diewald (2009) for an extensive overview of research on German modal particles. We therefore expect that the use of modal particles in our corpus will depend on the individual author, but also the medium and register.

2. We chose not to translate all modal particles since the result would not have improved the understanding of the examples we present.

The most frequent modal particle in our corpus is *ja*. We try to translate *ja* in the following examples, keeping in mind that modal particles are in fact very hard to translate (Bross 2012). The particle *ja* is often used to indicate known information, as in Example (1).

(1) Ich bin ja ein Einzelkind.
'As you know, I am an only child.'

Here, *ja* acts as an expression of the assumed shared knowledge of speaker and hearer (the speaker is assuming the hearer knows that the speaker is an only child). But *ja* can also be used to express surprise:

(2) Du hast ja ein Loch im Ärmel!
'Look, you have a hole in your sleeve!' (Kratzer 1999: 1)

In Example (2), *ja* is used to express a change in common ground, because a new piece of information (the hearer's sleeve having a hole) is introduced to the common ground by the speaker.

2.4.2 *Intensifiers*

For the second case study, we analyzed the use of intensifiers in our corpus. Intensifiers are words or expressions that can be used to boost or tone down the intensity of a gradable expression or utterance (Os 1989). Like modal particles, intensifiers are typically used in speech rather than in written language, and in informal rather than in formal language (Tagliamonte & Denis 2008). Intensifiers show great inter-individual variation and are subject to rapid change (Ito & Tagliamonte 2003). It has been shown that intensifiers can be frequent in written social media data as well (Scheffler 2017). German intensifiers have not been studied in as much detail as English ones. Claudi (2006) provides an overview of German intensifiers by their source semantics, and Breindl (2007) discusses (issues with) the categorization of German intensifiers. Stratton (2020) conducts a large-scale corpus study of intensifiers in spoken German, showing that they are used frequently (37% of all intensifiable adjectives were in fact intensified) and that the use of intensifiers varies by gender and age, confirming results from other languages.

Based on the previous sociolinguistic results, we expect that the use of intensifiers in social media varies by individual demographic factors of the authors (as shown for speech), but may also vary by medium and register.

3. A multi-register corpus of individual variation in social media

Existing German social media corpora are collections of blog posts (Barbaresi & Würzner 2014), tweets (Scheffler 2014; Barbaresi 2016), chat histories (Beißwenger 2013) or other types of data from a single social medium. Beißwenger, Lemnitzer and Müller-Spitzer (2022) list additional social media corpora in German and other languages. None of these German corpora contain parallel data from the same authors in different media. This means that direct comparison of linguistic features between corpora would always face a potential confound of which types of users choose to write about which types of topics in a particular medium (a difference in language use could be due to personal differences or topic differences, not the medium or situation). We further argued above that register must be studied in addition to (and separate from) the social medium in our corpus. Therefore, it is necessary that we look at texts from the same author in different media and registers and study whether and how the use of linguistic features changes. Theoretical considerations led us to distinguish three registers which cut across the medium distinction of blogs vs. tweets.

3.1 Data collection

In order to document intra-author linguistic variation across several social media and registers, a method was devised to identify individuals with linked blogs and Twitter accounts. As listed above, blogs and tweets are relatively commonly studied media in German, however they have not been addressed in direct comparison. Blogs are long-form text with limited interactivity, while tweets are short posts (all our tweets are still under 140 characters) which allow direct responses; both media are public. Thus, the two media offer different types of communicative situations, but they are both available for all three registers introduced above, depending on the individual usage. In order to keep the subcorpora comparable, we chose a community that is relatively coherent and active in both media, and which writes about a similar range of topics in both media: parenting bloggers.

We started from a Twitter list[3] of German parenting bloggers, which was accessed automatically via the Twitter API. We saved each listed user's Twitter name and the URL given in their user profile (if any), yielding 170 Twitter users with linked URLs. The provided URL for this group of parenting bloggers in most cases pointed to their blog. We tried to automatically extract the most recent blog posts via the blog's RSS feed, if available. An RSS feed provides standardized access to updates on a website, such as the list of recent blog posts. URLs

3. https://help.twitter.com/en/using-twitter/twitter-lists

that did not point to blogs or blogs without RSS feeds were ignored and the corresponding users were removed at this stage. We also collected all available tweets for each user via the Twitter API (a maximum of around 3200 tweets per user can be accessed). For 70 users, both blog posts (in most cases, the most recent 5 or 10) and tweets could be collected. We manually removed persons who posted primarily not in German or whose data was otherwise not usable (e.g., when multiple people tweeted from one account and authorship was not clearly identifiable). The initial data collection was carried out in February, 2017. A more detailed description of the corpus and data collection can be found in (Scheffler, Kern & Seemann forthcoming).

3.2 Preprocessing

German copyright law allows the use of web-scraped corpora for text and data mining purposes (§ 60d UrhG; BGBl. I, 2021: 1204). However, since the corpus consists of personal blogs and tweets, in 2021 we informed the authors via email, describing the purpose and range of our data collection and including an opt-out possibility. All authors who could not be contacted at this point as well as those who explicitly objected to the inclusion of their data (4 persons) were removed from the data set. In a final filtering step, we excluded blog posts not in German or authored by guests but kept the rest of the posts of these users. The resulting corpus consists of data from 44 authors/parent bloggers, comprising 390 blog posts (~350k words) and 81,440 tweets (~1.2 million words). Information about how to access the corpus can be found on its webpage.[4]

Since many of the blog posts and tweets contain information about the author's personal life and their children, all texts were anonymized semi-automatically: user mentions in tweets (@user) and URLs were replaced with a placeholder automatically. Person names, places and contact information were anonymized manually. Person names and places were left in the corpus if they refer to public entities, such as conference venues or politicians' names. All data is retained separately from the author metadata, and authors were assigned random numerical labels to keep the association between blogs and tweets corresponding to the same author. Examples (3) and (4) show such anonymized sentences.

(3) Um 5 Uhr ist die Nacht vorbei, [NAME] hat Hunger.
'At 5 am the night is over, [NAME] is hungry.' [blog-2995-1]

4. http://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/

(4) Seit ich im [ORT] bin, kann ich [ORT] gar nicht mehr verstehen.
[ORT]-Bashing aufm Blog: [URL]
'Since I am in [PLACE] I can't understand [PLACE] at all. [PLACE]-Bashing on the blog: [URL]' [tweets-7621]

In all documents, sentences were split automatically using the SoMaJo tokenizer,[5] which was developed specifically for German social media data.

3.3 Register annotation

Before analyzing register variation for individual authors as well as register variation in combination with the medium, we systematically annotated the register of each blog post and each collection of tweets per user. In doing so, we provided a starting point for our research in register variation and simultaneously tested our proposal for identifying register in social media corpora.

We assigned one of the three registers ('Informative', 'Narrative', 'Persuasive') to each individual blog post in the corpus. Excerpts of blog posts annotated as 'Informative' and 'Persuasive', translated into English, are given in Text Samples 1 and 2, respectively.

Text Sample 1. 'Informative'
The pregnancy provided the impetus to create non-alcoholic beverages. The idea came to [NAME], he says, actually together with his wife [NAME] was pregnant and both were looking for an appealing alternative to alcoholic drinks. […] The main idea finally came from a colleague who never drinks alcohol and had made a discovery in Denmark: a basic juice – in Horvath this is a mix of celery, apple and carrot, refined with appropriate additives. The basic juice is heated to 80 degrees and only the essence is used; the heavy particles settle to the surface and are skimmed off. [blog-1639-8]

Text Sample 2. 'Persuasive'
Children are our mirrors. If you want to change your child, change YOUR behavior, not the child's. My son has these tantrums all the time. Regularly. Then it is very difficult to get him out of it. And that is exactly what I would like to do. Get him out of it. So that he is calm as quickly as possible. So that I don't have to walk through the village or stand in the department store with this angry child. Because anger, yes, anger is an emotion that is not welcome in our society. Hm. At some point I asked myself why these fits upset me so much. And that's when I saw it: this angry little girl inside me, who would also like to let off a little steam. So now

5. https://github.com/tsproisl/SoMaJo

I regularly start to give space to my anger, my dark side. Since then I feel better. My migraines are also getting less frequent. And my child has fewer tantrums, and when he has one, I give him space. And I give space to my anger, too. So he doesn't have to be angry for so long anymore. [blog-4421-10]

Two of the authors independently annotated all blog posts and tweet collections in the corpus for register in order to check for inter-annotator agreement and ensure the quality of our annotation. The measured overall inter-annotator agreement was rated substantial (Cohen's $\kappa = 0.76$), partially validating our approach. Where the annotators disagreed, we discussed the choice and were able to agree on a final set of register annotations of our corpus. We annotated 434 items, of which 390 are blog posts and 44 are tweet collections (tweet collections consist of 154 up to 3197 tweets). We annotated registers by individually reading the texts and marking the texts with the previously introduced registers. As noted, tweet collections were assigned one register in total. To do this, we read the subcorpus of all tweets and blog posts for each user and applied our framework by characterizing the entire collection in terms of purpose, topic, interactivity, and involvement as defined above. This showed us that our register dimensions were in fact relevant and fitting for our corpus because we were able to fit each of the texts into one of the three register dimensions (a mix between two registers was additionally allowed for tweet collections, which resulted in the annotation of five registers in total). When we were unsure about the most fitting register, we investigated more features of the text such as the communicative goal of the posts.

For the tweet collections, we also permit the hybrid dimensions – 'Narrative-persuasive' and 'Narrative-informative', reflecting Biber and Egbert's (2016) findings for intermediate registers in their multi-dimensional analysis of web registers, which suggests that registers fade into each other rather than being independent from each other, and that some registers are closer to one another than others.

The annotation of register dimensions was in some parts difficult, because registers sometimes varied within singular texts, making it challenging to attribute one specific register to each blog post or collection of tweets. Annotating the tweet collections was especially complex, since the individual tweets are diverse and a wide range of propositions, topics and communicative goals was found. In these cases, we used the overall tenor of a user's tweets in order to attribute an appropriate register. We are aware that combining tweets by one user into a tweet collection simplifies the register diversity that is present within the tweets of said person. It has been shown in previous work (e. g. Clarke & Grieve 2019; Clarke 2022) that Twitter users vary the register they use on Twitter clearly, since different communicative situations and interactions take place on the platform. We have nevertheless chosen to combine tweets to maintain a manageable workload,

because individually annotating ca. 22,000 tweets would not have been feasible for us. In addition, it is often difficult to assess the register dimensions (purpose, topic, author and reader involvement) based on a single tweet, some of which are only a few words long. We therefore annotate the dominant register(s) within each user's tweet collections in aggregate. We maintain that this is still useful and note that even long texts such as blog posts may contain a few sentences which match a different register, such as when a primarily persuasive text contains an argument presented as a narrative. We have identified this fine-grained register annotation as an issue that we will work on in future research.

In Table 2, we summarize the results of our annotation.

Table 2. Register distribution within our annotated social media corpus

	Blog posts		Tweet collections	
Register	**Total**	**%**	**Total**	**%**
Informative	132	33.85%	9	20.45%
Narrative-informative	0	0.00%	8	18.18%
Narrative	205	52.56%	20	45.45%
Narrative-persuasive	0	0.00%	4	9.09%
Persuasive	53	13.59%	3	6.82%
Total	**390**	**100.00%**	**44**	**100.00%**

The most frequently used register dimension within the corpus was 'Narrative', which applies to about half of all annotated items (225), followed by 'Informative' (141 items, 32.49%) and 'Persuasive' (56 items, 12.9%). The distribution of registers is almost the same within the two media blogs and tweets, indicating that register, as defined above, really constitutes an orthogonal category not subsumed under medium distinctions. The dominance of the narrative register is perhaps reflective of the chosen topic domain of parenting blogs, which focus on personal stories and community building.

4. Case studies

Since we want to study whether discourse level linguistic variability can be attributed to variability within the individual and authors' choices of linguistic features, we conduct case studies of two selected variables: Modal particles and intensifiers. Before we test whether there are statistically significant differences in modal particle and intensifier use between different registers and social media, we describe

the quantitative distribution of our chosen linguistic features in the corpus. The frequencies of the features are given as absolute counts, as well as relative frequencies, including both instances per million words (pmw) and the fraction of all sentences containing the feature. We choose to include 'sentences with feature' as a relative measure, because both are discourse markers operating at the propositional (sentence) level semantically. Our main research interest is not how frequent modal particles and intensifiers are (both are rather infrequent), but rather how many sentences in a corpus are modified with such a feature. This removes differing sentence length in the different registers as a possible confound of the frequency of our target features.

4.1 Case study 1: Modal particles

A definitive list of German modal particles does not exist. In order to identify the particle usage within our corpus, we defined our own list of 39 particles based on previous work (König 1997; Thurmair 1989; Weydt 1979). We created detailed annotation guidelines, containing corpus examples for each particle, as well as a set of criteria for inclusion and exclusion for each instance. This was crucial since most modal particles have homophones in other lexical classes (Bross 2012) that can be mistaken easily for modal particles.[6]

For annotating, we used the web-based tool 'brat' (Stenetorp, Pyysalo, Topić, Ohta, Ananiadou & Tsujii 2012).[7] Within 'brat', each document (blog post or tweet) was read by the annotator and particles were identified manually according to our guidelines. Problems or limiting cases were discussed within our research group. We manually annotated all particles in all blog posts and in 500 tweets from each user (fewer if the user had not authored 500 tweets).

In total we annotated 3,611 modal particle instances in our corpus. The five most frequently used modal particles (identical within blog posts and tweets) were *ja* (1021 occurrences in total), *doch* (382), *einfach* (347), *eigentlich* (301) and *denn* (214). We calculated that on average, 6.09% of all sentences across the corpus contain at least one modal particle. A table containing all annotated modal particles and their frequencies can be found in Appendix A.

Comparing the two media, 8.82% of all sentences in the blog posts contain a modal particle and 5.14% of all sentences in tweets. Because sentences in blog

6. We will publish these annotation guidelines alongside our corpus on the website of our corpus. Due to the difficulty of translating all our modal particles, the guidelines are only published in German.

7. https://brat.nlplab.org/

posts are substantively longer than in tweets, the frequency of modal particles per million words is actually lower in the blog posts than in the tweets.

We also looked into the individual distribution of modal particle use between the two media, which showed that 77.3% of all users used more modal particles in their blog posts than in their tweets, whereas 22.7% of all users used modal particles relatively more frequently in their tweets than their blog posts.

We were interested in intra-author variation in terms of modal particle use to examine whether users tend to use different modal particles in tweets and blog posts or whether they have 'preferred' particles they use no matter which medium they choose. Out of 44 users, 19 users had a different most used modal particle in their blog posts and their tweets, while 25 users used the same particle most frequently within their blog and tweets. For 22 out of these 25 users, the most used modal particle was *ja*, two users preferred *doch*, and one person used *mal* in both their blog posts and tweets most frequently.

We also looked at the specific particles used by each user in our corpus. Most users clearly preferred the same particle in all their texts, regardless of medium. Interestingly, this kind of consistency of modal particle use goes hand in hand with register consistency across texts: It seems that people who stay in the same register in blog posts and tweets show less variation in their preferred modal particles. We marked 27 users as register-consistent, which means their overall Twitter and blog corpus were annotated with the same register.

As for the interaction of particle use and register, narrative texts contain the most particles, followed by informative blog posts. Table 3 lists the relative frequencies of particles by register and medium. It can be seen that modal particles are relatively frequent in German texts, occurring in about 8% of sentences in blogs (0.46% of words) and 5.4% of sentences on Twitter (0.61% of words). The fact that the word-based frequency of particles is higher on Twitter than in blogs, whereas the opposite is true for sentence-based frequency, is due to the fact that the sentences in the blog posts are longer, on average, than the Twitter sentences. The overall frequencies of particles are similar in the two media, however.

A bigger difference can be seen with respect to the registers, where narrative texts contain significantly more particles than other registers. We conducted a χ^2-test to test whether there is a statistically significant dependency between register and particle count in each medium. The significance level is set to be $\alpha = 0.05$.

H_0: There is no dependency between register and particle count.
H_A: There is a dependency between register and particle count.

With $\chi^2 = 618.5$, $df = 4$, $p < .0001$, the null hypothesis can be rejected. The statistical analysis shows that there is indeed a dependency between register and particle count. This indicates that modal particles are among the linguistic features that

Table 3. Absolute and relative counts of modal particles in each register in all blog posts and tweet collections

	Blog posts			Tweet collections			Entire corpus		
	Count	Sentences with modal particles	Pmw	Count	Sentences with modal particles	Pmw	Count	Sentences with modal particles	Pmw
Informative	284	5.44%	810	212	3.44%	708	496	4.36%	763
Narrative-informative	0	0.00%	0	370	5.65%	1236	370	5.65%	569
Narrative	1224	10.54%	3489	924	5.67%	3088	2148	7.70%	3304
Narrative-persuasive	0	0.00%	0	204	7.59%	682	204	7.59%	314
Persuasive	97	2.88%	276	133	5.17%	444	230	3.87%	354
Total	**1605**	**7.94%**	**4575**	**1843**	**5.38%**	**6159**	**3448**	**6.33%**	**5304**

Note. Counts are relative to the total number of sentences in this register and medium.

vary by register. We believe that the high usage of modal particles in narratives is due to their meaning as markers that establish common ground or indicate speaker stance (functions that can be useful in narratives).

4.2 Case study 2: Intensifiers

We could not annotate all the intensifiers in our data, but we wanted to study intensifiers which are commonly used in our corpus. Therefore, we created a smaller subcorpus with the data from all blog posts and tweet collections of 6 (out of the 44) users – two chosen randomly for each register: 'Informative', 'Narrative' and 'Persuasive'. We used the list of over 200 German intensifiers compiled in (Scheffler, Richter & van Hout in review) to search for intensifiers with the data analysis tool MAXQDA 2020 (VERBI Software 2019). All matches of all intensifiers were disambiguated manually. This subcorpus consists of ~148k words, ~103k in tweets and ~45k in blog posts.

We found a total of 1024 uses of 43 different intensifiers in the 6-user subcorpus; the five most frequent ones are *so* (*so*), *sehr* (*very*), *ganz* (*totally*), *gar* (*at all*), and *wirklich* (*really*), shown in Table 4. A table containing all annotated intensifiers and their frequencies in the 6-user subcorpus can be found in Appendix B. For these 6 users, 6.77% of all sentences contain at least one of the 43 different intensifiers that were found in the data.

Table 4. Distribution of the five most frequent intensifiers in the 6-user subcorpus

	Blog posts			Tweet collections			Entire subcorpus		
Intensifier	Count	Sentences with this intensifier	Pmw	Count	Sentences with this intensifier	Pmw	Count	Sentences with this intensifier	Pmw
so	92	3.03%	2067	193	1.52%	1863	285	1.81%	1924
sehr	54	1.78%	1213	150	1.18%	1448	204	1.30%	1337
ganz	66	2.17%	1483	127	1.00%	1226	193	1.23%	1303
gar	36	1.18%	809	74	0.58%	714	110	0.70%	743
wirklich	16	0.53%	359	24	0.19%	232	40	0.25%	270

The intensifier which is used by far the most in the 6-user subcorpus is *so*. To study the use of intensifiers in the full 44-user corpus, we annotated all instances of *so* in all blog posts and tweet collections in the full corpus. *So* was searched for with regular expressions to match all cases of capitalization, lengthening (*sooo*), reduplication (*sososo*), etc. Schumann (2021) discusses all the different uses of the German particle *so*; her work was used to disambiguate the intensifying function of *so* from others.

So is used 3,552 times in the 44-user corpus. Of all 3,552 instances of *so*, it is lengthened, capitalized or reduplicated 284 times (7.53% of uses). Except for one user, who does not use *so* in their tweets, *so* is used in tweet collections and blog posts by all users. With regard to register, *so* is used the most in narrative blog posts and tweet collections and least in informative documents. Table 5 shows the absolute and relative counts of *so* as an intensifier in all blog posts and tweet collections. The table further shows that *so* is far less frequent in the informative register. This is not surprising since strong emotionality is on the one hand tied to the use of intensifiers (Scheffler et al. in review), and on the other hand reflected in our register characteristics as high involvement of the author: Informative texts are less emotional than narrative or persuasive texts, which can be quite personal.

To test whether there is a statistically significant dependency between register and intensifier (*so*) count in each medium, we conduct a χ^2-test. Again, the significance level is set to be $\alpha = 0.05$.

H_0: There is no dependency between register and intensifier count.
H_A: There is a dependency between register and intensifier count.

With $\chi^2 = 320.59$, $df = 4$, $p < .0001$, the null hypothesis can be rejected. The statistical analysis shows that there is a dependency between register and intensifier

Table 5. Absolute and relative counts of intensifying 'so' in each register in all blog posts and tweet collections

	Blog posts			Tweet collections			Entire corpus		
Intensifier		Sentences			Sentences			Sentences	
so	Count	with *so*	Pmw	Count	with *so*	Pmw	Count	with *so*	Pmw
Informative	96	1.76%	374	125	0.93%	107	221	1.17%	145
Narrative-informative	0	0.00%	0	439	1.78%	375	439	2.04%	289
Narrative	397	3.80%	1132	2107	2.60%	1800	2504	2.70%	1646
Narrative-persuasive	0	0.00%	0	101	1.24%	86	101	1.24%	66
Persuasive	112	3.30%	319	175	1.64%	149	287	2.04%	189
Total	**605**	**2.94%**	**1725**	**2947**	**2.14%**	**2517**	**3552**	**2.24%**	**2334**

Note. Counts are relative to the total number of sentences in this register and medium.

count. This indicates that intensifiers, as well as modal particles, are fit to be investigated as linguistic features that vary by register.

5. Analysis: Linguistic influence of medium and register

Our corpus annotation at the level of medium, register, and variable linguistic phenomena allows us to tease apart the interactions between these different levels. We want to investigate which effects medium and register have on specific linguistic expressions in German social media data. In addition, we hope to show that register must be considered as a separate category of analysis from medium in order to achieve explanatory adequacy.

To document the effect of medium and register on the linguistic variables, we want to test if there is a statistically significant difference between the group's means if we use modal particles and the intensifier *so* as dependent variables. We use the data described in Tables 3 and 5: the relative frequencies (sentence level) of all modal particles and intensifying *so* in all 44 user's blog posts and tweet collections.

For this data, assumption of independence is met, but Shapiro-Wilk-test and Fligner-Killeen-test show that the assumptions of normality and homogeneity of variances of the residuals are not met. For this reason, we use the Kruskal-Wallis-test for non-parametric data and Wilcoxon-test with Bonferroni correction as a post hoc analysis to test whether any groups differ significantly from each other. The null hypothesis of each Kruskal-Wallis-test is:

H_0: There is no difference between the groups' (= each register's/medium's) mean rank.
H_A: There is a difference between the groups' (= each register's/medium's) mean rank.

Table 6 shows that at $\alpha = 0.05$ we can reject the null hypothesis for particle use by register, particle use by medium and intensifying *so* by register. In these three samples, there are significant differences between the groups' mean ranks. That indicates that there is indeed a statistically significant difference in the use of modal particles in different registers and social media, and a difference in the use of the intensifier *so* in different registers.

Table 6. Kruskal-Wallis-test for particles/'so' by register/medium

Data	Kruskal-Wallis chi-squared	*df*	*p*-value
Particles by register	27.711	4	< 0.0001
Particles by medium	50.765	1	< 0.0001
Intensifier *so* by register	44.754	4	< 0.0001
Intensifier *so* by medium	0.479	1	0.4891

We use the Wilcoxon-test with Bonferroni correction to test between which groups (= registers/media) of these three samples there are differences. At $\alpha = 0.05$, there are, in both particles and *so*, statistically significant differences between the mean ranks of the registers 'Informative' and 'Narrative' ($p = 0.0029$ / 1.4e-08) and between the mean ranks of 'Informative' and 'Persuasive' ($p = 0.0481$ / 5.7e-06). In the sample particles by register, there is also a significant difference between 'Narrative-informative' and 'Persuasive' ($p = 0.004$) and between 'Narrative-informative' and 'Narrative' ($p = 0.01$). In the sample particle by medium, there is a statistically significant difference between the mean rank of blog posts and tweets ($p = 1e\text{-}12$).

The statistical analysis supports our hypothesis that the social medium, as well as register, independently influences linguistic choice. There is a difference in the choice of linguistic features between the registers 'Informative' and 'Narrative' as well as 'Persuasive' for both modal particles and the intensifier *so*. Furthermore, there is a difference in modal particle choice between the registers 'Narrative-informative' and 'Persuasive' as well as 'Narrative-informative' and 'Narrative'. It is not surprising to see this in only modal particle choice and not intensifier choice as well, since we used a lot more modal particles for this test than intensifiers. In all cases, there seems to be no statistically significant difference in the use of these linguistic features between the registers 'Narrative' and 'Persuasive'.

To track variability between media, we take a look at which particles are used more or less often in which media. In our corpus, blog posts and tweets differ markedly from each other in terms of particle use. This applies to the overall frequency of modal particles: they are more frequent per sentence in blog posts (7.94% of sentences) than in tweets (5.38%). The medium also can be observed in the choice of specific particles chosen by authors. For example, the inquisitive particle *denn* (restricted to questions) is found much more frequently in the tweets than in the blogs. This can be explained by the medium's affordances (see Section 2.1.1): The affordances of Twitter allow for a more interactive question-answer pattern in communication, whereas interaction on blogs is limited to the comments section. See Table 7 for a list of the most frequent particles in the two media.

Table 7. Most frequent modal particles in blogs and tweets (absolute counts in brackets)

Blogs	ja (428), einfach (227), doch (170), eigentlich (138), denn (80), mal (79), eben (69), sogar (69), wirklich (63), wohl (57)
Tweets	ja (539), doch (212), eigentlich (163), denn (134), einfach (120), wohl (101), mal (65), halt (64), schon (38), sogar (36)

There is no clear medium difference in the use of *so* as an intensifier between blogs and tweets (see Tables 5 & 6), since *so* is the most frequently used intensifier in colloquial German (Stratton 2020; Scheffler et al. in review) and makes up an equal share of all intensifiers in both media in the 6-user subcorpus (about 27% of all intensifiers). But the overall distribution of intensifiers is different between the media. For example, more traditional, formal intensifiers such as *wirklich* (*really*) and *sehr* (*very*) are used less often in Twitter (see also Scheffler 2017).

The register, on the other hand, shows a clear effect on the frequency and types of particles and intensifiers chosen by authors. As demonstrated by the statistical analysis in Table 6, both the frequency of particles and of the intensifier *so* correlates with the register of the text: narrative texts make use of the most modal particles, while informative texts employ the least. We conjecture that particles are used in narrative texts to show author and reader involvement (which also applies to persuasive texts), but mainly to establish authenticity and a personal, 'close' (in the sense of Koch & Oesterreicher 1985) style of interaction. The intensifier *so* is used most in persuasive and narrative texts, reflecting the high level of author involvement (emotionality) in these registers. As described above, this preference is independent of the medium (blog/Twitter) of the text, but merely determined by the register.

6. Discussion and conclusion

Our study starts from the observation that linguistic expressions in social media vary along many axes, including individual style, the specific medium and its affordances, and the register employed. We define register as the situational characteristics describing an instance of language use. Registers are thus differentiated primarily by extralinguistic properties of the communication situation. This view allows us to then investigate the influence of the register on specific variable linguistic phenomena similarly to comparing linguistic expressions in different media or from different demographic groups. Here, we define three registers relevant to the parenting blogger corpus that constitutes the foundation of our study. The registers are distinguished based on four situational factors adapted from prior work, and are defined as 'Informative' (with a main purpose of providing information to readers in a neutral manner), 'Narrative' (retelling personal stories, a primary function in our domain), and 'Persuasive' (calling to action or making arguments).

We introduce a new German cross-media corpus, consisting of blog posts and tweets from the same 44 parenting bloggers. We assign registers to each of the blog posts and summarily to each author's entire tweet collection. About half of the texts in the corpus are narrative, followed by informative and persuasive texts. The corpus is special in that it revolves around a common topic, but includes texts from multiple authors in two different media (blogs/tweets) and three registers ('Narrative', 'Informative', 'Persuasive') in a fully crossed fashion. The distribution of register in the two media is similar, with most texts being narrative, followed by informative and finally persuasive.

Given the register annotation, we focus on two variable linguistic phenomena: German modal and intensifying particles. In each case, we document considerable inter- and intra-individual variation in the types of expressions used and their frequency across texts. The statistical analysis shows that the particle usage differs significantly across media, but also (independently) between the different registers. For the usage frequency of the most common intensifier *so*, only register had a significant effect. These results confirm that modal particles are typical of narrative texts and are used there to establish a relation between the author and the reader. Intensifiers such as *so* are used in persuasive and narrative contexts that show high emotional involvement of the author. Both case studies demonstrate a clear relevance of distinguishing media-independent registers within social media texts when analyzing linguistic features.

Further research could study lexical and grammatical features that are frequently used in informative contexts. Conjunctions might be such a feature, as there might be a difference in the conjunction used depending on the register.

Media differences in the use of (causal) conjunctions have for example been documented in persuasive news texts and Twitter posts (Scheffler & Stede 2016), but without reference to other registers.

As for the chosen media and the similarities between blog posts and Twitter posts, the fact that both are a form of computer mediated communication and can be either used for formal or informal communication reinforces our hypothesis: Language use does not only depend on the medium a text is written in, but also on the register. To test how well our register dimensions are suited for other kinds of CMC, further research could use our non-linguistic factors specified for register annotation to annotate register dimensions in voice messages or Instagram posts.

To summarize, we studied a new cross-media corpus of German social media posts. It is known that the medium influences the linguistic choices of language users, but we argue that the medium is not the (whole) message (cf. McLuhan 1964): We introduce three broad registers, which cut across the two media in our corpus (blogs and tweets), and have a greater influence than the medium on the particular linguistic variables we investigate. We therefore propose to (i) acknowledge that there are different registers represented within social media (and not just one social media register dimension or one register for each medium) and to (ii) include these registers as an important factor in future linguistic analyses of social media texts.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287.

Acknowledgements

We would like to thank the anonymous reviewers for their detailed and valuable comments.

References

Androutsopoulos, J. (2014). Mediatization and sociolinguistic change. Key concepts, research traditions, open issues. In J. Androutsopoulos (Ed.), *Mediatization and sociolinguistic change* (pp. 3–48). Berlin/ Boston: De Gruyter. https://doi.org/10.1515/9783110346831.3

Argamon, S. E. (2019). Computational register analysis and synthesis. *ArXiv:1901.02543 [Cs], to appear in Register Studies*. Retrieved from http://arxiv.org/abs/1901.02543. https://doi.org/10.31235/osf.io/t64sy

Barbaresi, A. (2016). Collection and indexing of tweets with a geographical focus. *Tenth International Conference on Language Resources and Evaluation* (pp. 24–27). Portorož, Slovenia.

Barbaresi, A., & Würzner, K. M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. *Proceedings of NLP4CMC Workshop* (pp. 2–10). Hildesheim: Hildesheim University Press.

Beißwenger, M. (2013). *Das Dortmunder Chat-Korpus: ein annotiertes Korpus zur Sprachverwendung und sprachlichen Variation in der deutschsprachigen Chat-Kommunikation*. Online-Publikation auf dem Linguistik Server Essen (LINSE). https://doi.org/10.1515/zgl-2013-0009

Beißwenger, M., Lemnitzer, L., & Müller-Spitzer, C. (Eds.). (2022). *Forschen in der Linguistik: eine Methodeneinführung für das Germanistik-Studium*. Paderborn: Brill / Fink. https://doi.org/10.36198/9783838557113

BGBl. I (2021). Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz). *§ 60d Text und Data Mining für Zwecke der wissenschaftlichen Forschung*.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, *19*(2), 219–241.

Biber, D., & Conrad, S. (2005). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (1st ed., pp. 175–196). Hoboken: Wiley. https://doi.org/10.1002/9780470753460.ch10

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, *44*(2), 95–137. https://doi.org/10.1177/0075424216628955

Bildhauer, F., Pankratz, E., & Schäfer, R. (2021). *Corpus, inference, and models of register distribution*. Talk presented at the CCMLMA at DGfS.

Breindl, E. (2007). Intensitätspartikeln. In L. Hoffmann (Ed.), *Handbuch der deutschen Wortarten* (pp. 397–422). Berlin/ New York: De Gruyter.

Bross, F. (2012). German modal particles and the common ground. *Helikon. A Multidisciplinary Online Journal*, *2*, 182–209.

Clarke, I. (2022). A Multi-dimensional analysis of English tweets. *Language and Literature: International Journal of Stylistics*, 1–26. https://doi.org/10.1177/09639470221090369

Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLOS ONE*, *14*(9), e0222062. https://doi.org/10.1371/journal.pone.0222062

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

Degand, L., Cornillie, B., & Pietrandrea, P. (Eds.). (2013). *Discourse markers and modal particles: categorization and description*. Amsterdam/ Philadelphia: John Benjamins. https://doi.org/10.1075/pbns.234

Diewald, G. (2009). Abtönungspartikel. In L. Hoffmann (Ed.), *Handbuch der deutschen Wortarten* (pp. 117–141). Berlin/ New York: de Gruyter.

Döring, S. (2016). Modal Particles, Discourse Structure and Common Ground Management. (Dissertation). Humboldt-Universität zu Berlin.

Gibson, J. J. (2014). *The ecological approach to visual perception*. New York/ London: Taylor & Francis Group. https://doi.org/10.4324/9781315740218

Honeybone, P. (2011). Variation and linguistic theory. In W. Maguire & A. McMahon (Eds.), *Analysing variation in English* (pp. 151–177). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511976360.008

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, *32*(2), 257–279. https://doi.org/10.1017/S0047404503322055

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, *36*, 15–43. https://doi.org/10.1515/9783110244922.15

König, E. (1997). Zur Bedeutung von Modalpartikeln im Deutschen: Ein Neuansatz im Rahmen der Relevanztheorie. In F. Debus (Ed.), *Studien zu Deutsch als Fremdsprache III: Aspekte der Modalität im deutschen-auch in kontrastiver Sicht*. Hildesheim/ New York: G. Olms.

Kratzer, A. (1999). *Beyond ouch and oops: How descriptive and expressive meaning interact*. https://semanticsarchive.net/Archive/WEwNGUyO/

Labov, W. (2010). *Principles of linguistic change. 2: Social factors*. Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781444327496

MacKenzie, L. (2019). Perturbing the community grammar: Individual differences and community-level constraints on sociolinguistic variation. *Glossa: A Journal of General Linguistics*, *4*(1). https://doi.org/10.5334/gjgl.622

McLuhan, M. (1964). *Understanding media: The extensions of man*. Boston: MIT Press.

Os, C. van. (1989). *Aspekte der Intensivierung im Deutschen*. Tübingen: Narr.

Scheffler, T. (2014). A German Twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Presented at the LREC'14 (pp. 2284–2289), Reykjavik, Iceland. European Language Resources Association

Scheffler, T. (2017). Conversations on Twitter. In D. Fišer & M. Beisswenger (Eds.), *Investigating computer-mediated communication: corpus-based approaches to language in the digital world* (1st ed.). Ljubljana: Ljubljana University Press.

Scheffler, T., Kern, L.-A., & Seemann, H. (forthcoming). Individuelle linguistische Variabilität in sozialen Medien. In M. Kupietz & T. Schmidt (Ed.), *Korpora in der germanistischen Sprachwissenschaft – mündlich, schriftlich, multimedial*. Tübingen: Narr.

Scheffler, T., Richter, M., & van Hout, R. (in review). Tracing and classifying German intensifiers via information theory. *Language Sciences*.

Scheffler, T., & Stede, M. (2016). Realizing argumentative coherence relations in German: A contrastive study of newspaper editorials and Twitter posts. In P. Saint-Dizier & M. Stede (Eds.), *Proceedings of the COMMA Workshop 'Foundations of the Language of Argumentation'* (pp. 73–80). https://www.ling.uni-potsdam.de/comma2016/pdf/FLA16-proceedings.pdf

Schleef, E. (2021). Individual differences in intra-speaker variation: t-glottalling in England and Scotland. *Linguistics Vanguard*, *7*(2). https://doi.org/10.1515/lingvan-2020-0033. https://doi.org/10.1515/lingvan-2020-0033

Schumann, K. (2021). *Der Fokusmarker ‚so': Empirische Perspektiven auf Gebrauch und Verarbeitung eines Ausnahmeelements*. Boston: de Gruyter. https://doi.org/10.1515/9783110731149

Stefanowitsch, A. (2020). *Corpus linguistics: a guide to the methodology*. Berlin: Language Science Press.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: a Web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107). Avignon, France: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E12-2021

Storrer, A. (2013). Sprachstil und Sprachvariation in sozialen Netzwerken. In *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW* (pp. 331–366). Wiesbaden: VS Verlaf für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93336-8_15

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

Tagliamonte, S.A. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, *45*(1), 1–32. https://doi.org/10.1017/S0047404515000780

Tagliamonte, S.A., & Denis, D. (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech*, *83*(1), 3–34. https://doi.org/10.1215/00031283-2008-001

Thurmair, M. (1989). *Modalpartikeln und ihre Kombinationen*. Tübingen: M. Niemeyer. https://doi.org/10.1515/9783111354569

VERBI Software. (2019). *MAXQDA 2020*. Berlin: VERBI Software. Retrieved from https://www.maxqda.com/

Weydt, H. (Ed.). (1979). *Die Partikeln der deutschen Sprache*. Berlin/ New York: de Gruyter. https://doi.org/10.1515/9783110863574

Wolfram, W. (2006). Variation and language: Overview. In R. Asher (Ed.), *Encyclopedia of Language & Linguistics* (pp. 333–341). Amsterdam: Elsevier. https://doi.org/10.1016/B0-08-044854-2/04256-5

Zhao, Y., Liu, J., Tang, J., & Zhu, Q. (2013). Conceptualizing perceived affordances in social media interaction design. *Aslib Proceedings*, *65*(3), 289–303. https://doi.org/10.1108/00012531311330656

Zimmermann, M. (2011). Discourse particles. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics* (pp. 2011–2038). Berlin: Mouton de Gruyter.

Appendix A. Annotated modal particles

Table 8. Absolute and relative counts of modal particles in all blog posts and tweet collections

	Blog posts			Tweet collections		
Modal particle	Count	Sentences with this modal particle	Per million words	Count	Sentences with this modal particle	Per million words
aber	15	0.07%	43	25	0.07%	84
allenfalls	0	0.00%	0	0	0.00%	0
allerdings	12	0.06%	34	2	0.01%	7
annähernd	0	0.00%	0	0	0.00%	0
auch	42	0.2%	120	21	0.06%	70
besonders	2	0.01%	6	0	0.00%	0
bestenfalls	0	0.00%	0	0	0.00%	0
bloß	6	0.03%	17	7	0.02%	23
denn	80	0.38%	228	134	0.4%	448
doch	170	0.81%	485	212	0.63%	708
eben	69	0.33%	197	24	0.07%	80
eh	31	0.15%	88	34	0.1%	114
eigentlich	138	0.65%	393	163	0.48%	545
einfach	227	1.08%	647	120	0.36%	401
fei	0	0.00%	0	0	0.00%	0
freilich	0	0.00%	0	0	0.00%	0
gleich	12	0.06%	34	6	0.02%	20
halt	25	0.12%	71	64	0.19%	214
hoffentlich	5	0.02%	14	3	0.01%	10
irgendwie	33	0.16%	94	25	0.07%	84
ja	428	2.03%	1220	593	1.76%	1982
jetzt	45	0.21%	128	27	0.08%	90
leider	20	0.09%	57	11	0.03%	37
mal	79	0.37%	225	65	0.19%	217
nicht	11	0.05%	31	14	0.04%	47
noch	9	0.04%	26	1	0.00%	3
offenbar	4	0.02%	11	2	0.01%	7
ruhig	2	0.01%	6	2	0.01%	7
schon	48	0.23%	137	38	0.11%	127
selbst	5	0.02%	14	3	0.01%	10
sicher	11	0.05%	31	11	0.03%	37
sogar	69	0.33%	197	36	0.11%	120

Modal particle	Blog posts			Tweet collections		
	Count	Sentences with this modal particle	Per million words	Count	Sentences with this modal particle	Per million words
tatsächlich	23	0.11%	66	19	0.06%	63
vielleicht	45	0.21%	128	15	0.04%	50
vor allem	4	0.02%	11	1	0.00%	3
wahrscheinlich	12	0.06%	34	5	0.01%	17
wenigstens	3	0.01%	9	1	0.00%	3
wirklich	63	0.30%	180	21	0.06%	70
wohl	57	0.27%	214	101	0.30%	338
Total	**1805**	**8.82%**	**5145**	**1806**	**5.14%**	**6035**

Note. Counts are relative to the total number of sentences in this medium.

Appendix B. Annotated intensifiers in the 6-user subcorpus

Table 9. Absolute and relative counts of intensifiers in blog posts and tweet collections in the 6-user subcorpus

Intensifier	Blog posts			Tweet collections		
	Count	Sentences with this intensifier	Per million words	Count	Sentences with this intensifier	Per million words
absolut	1	0.03%	22	2	0.02%	19
allzu	1	0.03%	22	0	0.00%	0
arg	1	0.03%	22	2	0.02%	19
äußerst	1	0.03%	22	0	0.00%	0
besonders	5	0.16%	112	3	0.02%	29
dermaßen	0	0.00%	0	2	0.02%	19
doll	0	0.00%	0	2	0.02%	19
echt	0	0.00%	0	21	0.17%	203
einfach	9	0.30%	202	8	0.06%	77
enorm	1	0.03%	22	0	0.00%	0
extra	0	0.00%	0	5	0.04%	48
furchtbar	3	0.10%	67	1	0.01%	10
ganz	65	2.14%	1460	127	1.00%	1226

	Blog posts			Tweet collections		
Intensifier	Count	Sentences with this intensifier	Per million words	Count	Sentences with this intensifier	Per million words
gar	36	1.18%	809	74	0.58%	714
gewaltig	1	0.03%	22	0	0.00%	0
höllisch	0	0.00%	0	1	0.01%	10
komplett	1	0.03%	22	0	0.00%	0
mega	0	0.00%	0	1	0.01%	10
richtig	5	0.16%	112	21	0.17%	203
schön	4	0.13%	90	4	0.03%	39
schwer	0	0.00%	0	1	0.01%	10
sehr	54	1.78%	1213	150	1.18%	1448
so	92	3.03%	2067	193	1.52%	1863
super	1	0.03%	22	5	0.04%	48
tief	1	0.03%	22	1	0.01%	10
total	**5**	**0.16%**	**112**	**14**	**0.11%**	**135**
überhaupt	1	0.03%	22	0	0.00%	0
ultra	0	0.00%	0	1	0.01%	10
unendlich	1	0.03%	22	1	0.01%	10
unerträglich	0	0.00%	0	1	0.01%	10
unfassbar	0	0.00%	0	1	0.01%	10
ungemein	1	0.03%	22	0	0.00%	0
unglaublich	0	0.00%	0	1	0.01%	10
unheimlich	1	0.03%	22	0	0.00%	0
verdammt	5	0.16%	112	4	0.03%	39
voll	0	0.00%	0	10	0.08%	97
vollkommen	1	0.03%	22	3	0.02%	29
völlig	3	0.10%	67	12	0.09%	116
wahnsinnig	0	0.00%	0	4	0.03%	39
wirklich	16	0.53%	359	24	0.19%	232
wunderbar	2	0.07%	45	1	0.01%	10
zu	0	0.00%	0	2	0.02%	19
zutiefst	0	0.00%	0	3	0.02%	29
Total	**318**	**10.47%**	**3100**	**706**	**5.56%**	**6815**

Note. Counts are relative to the total number of sentences in this medium.

Address for correspondence

Tatjana Scheffler
Fakultät für Philologie, Germanistik
Ruhr-Universität Bochum
Universitätsstraße 150
44780 Bochum
Germany
tatjana.scheffler@rub.de
https://orcid.org/0000-0001-7498-6202

Co-author information

Lesley-Ann Kern
Germanistische Sprachwissenschaft
Philipps-Universität Marburg
lesley-ann.kern@uni-marburg.de
https://orcid.org/0000-0002-2818-3692

Hannah Seemann
Fakultät für Philologie, Germanistik
Ruhr-Universität Bochum
hannah.seemann@rub.de
https://orcid.org/0000-0001-8568-2124

Publication history

Date received: 7 January 2022
Date accepted: 15 August 2022
Published online: 27 October 2022