# Conversations on Twitter

*Tatjana Scheffler, University of Potsdam*

### Abstract

In this paper, we analyse the linguistic structure of a corpus of German conversations on Twitter. Near real-time conversations conducted on social media are interesting from a linguistic viewpoint, because they show features of informal, spoken dialog while being transmitted asynchronously and in the written mode. The current study focuses on models of dialog structure developed for spoken conversations and their applicability to conversations on Twitter. We show that many well-known dialog phenomena can be observed in Twitter conversations, such as the use of particles, questions, turn-taking, informal lexical choice, corrections and fillers. At the same time, speakers on social media also frequently avail themselves of more formal, written-like options, and some spoken-like features take on new meanings in social media. Our approach allows for sub-dividing the conversations into three different types based on their structure, since a single medium such as Twitter combines several subgenres, such as chats among friends, surveys, customer-service dialogs, and so on. We distinguish broadcasts from linear conversations and group discussions.

**Keywords:** dialog, Twitter, social media, conversation structure, German

# 1 INTRODUCTION

In this paper we investigate German Twitter conversations. We identify properties of the structure of Twitter conversations and look specifically for phenomena typical of informal spoken conversations. We find that many features of spoken conversations are found equally in our Twitter corpus. However, there are also some differences that open interesting avenues for future work, such as a novel way of marking clarification requests, and idiosyncrasies in the use of discourse particles.

It is a defining feature of social media that they allow for interaction among their users. As opposed to traditional written (news) media, text is not only produced by a few and consumed by many, but instead linguistic data is produced and consumed near-simultaneously by many speakers.[1] Even though all "social" media enable conversations in this way, different channels can be distinguished by their interactive properties, as detailed in Table 1. Of the existing media with a mainly textual basis, Twitter is among the most conversational in nature. This paper studies the conversation structure of German Twitter data, in order to pin down the commonalities and differences of such computer-mediated conversations with spoken dialogs.

The paper makes three contributions. First, in Section 3, we detail our method for extracting conversations from Twitter and give an overview of the resulting corpus, a dataset of over 2.5 million threads (each between two and several hundred tweets). In Section 4, we analyse the dialog structure of the extracted Twitter threads and show structural measures to identify different types of conversations: broadcasts, group discussions, and linear conversations. In Section 5, we address several linguistic phenomena that are said to be typical of spoken conversations, in order to get a closer view of the linguistic properties of Twitter conversations. The careful comparison of "spoken" phenomena occurring in different social media allows us to tease apart the effects of the mode (spoken vs. written), interactional vs. informational style (Storrer 2013), informal vs. formal relations between speaker and hearer, binary interaction vs. multilog, etc. We find that some features of spontaneous interaction, for example questions, including clarification questions, occur frequently in the Twitter dialogs. On the other hand, while some modal particles are more frequent in the Twitter conversations than in monological text, this is not as pronounced overall. We argue that different social media with their specific configurations allow us to further study which property of a linguistic context licenses which types of expression.

[1] Though social media content is produced in writing, in this paper we use the terms 'speaker' and 'hearer' loosely to refer to the producers and addressees of utterances.

In order to enable comparison across different types of media, we focus here on linguistic phenomena that differentiate between spoken conversations and written text, and we exclude novel features specific to social media channels, such as emoticons, inflectives, across the board capitalization, etc. Though those social media innovations are important objects of linguistic study, we are more interested in the following research questions: Which characteristics typical of free spoken interactions carry over to social media conversations (on Twitter)? Which differences in frequency, use and meaning do we find between the modes, and how can this be explained?

# 2 BACKGROUND

In this paper, we study Twitter conversations from the perspective of the conceptual orality continuum (Koch and Oesterreicher 1985), comparing the medium to typical spoken or written data. In particular, we analyse to what extent the *dialog structure* of social media (Twitter) corresponds to what is known about spoken conversations. In this section, we address both lines of previous research in turn.

## 2.1 Characteristics of Spoken Dialogs

Herbert H. Clark and colleagues have established a view of conversations as a specific kind of linguistic communication in linguistics and psychology (Clark and Schaefer 1987, Clark and Schaefer 1989). From this perspective, conversations are not merely sentences uttered by different people in turn, but must be viewed as joint actions (like a hand-shake) of several participants (simultaneously speakers and hearers). Previous research shows how speakers and hearers coordinate across a conversation to achieve their common communicative goals. In prototypical face-to-face conversations, all participants are furthermore on equal footing (as opposed to, say, a radio interview, where one participant leads the conversation) with regard to access to and position in the dialog. Conversations are situated in a physical context and unfold in real-time, typically in spoken form. They are characterized by phenomena representative of spontaneous speech, such as clarification requests, corrections, fillers, pauses, and the like. This line of research is based on the analysis of natural conversations, either in person or over the telephone.

This work shows that contributions in dialog must be *grounded*, i.e. acknowledged and accepted by the conversation participants, in order to advance the discourse. Thus, unlike in written monolog, each contribution in spoken conversations consists of two phases, a *presentation* and an *acceptance* phase, where the presentation is done by the speaker and the acceptance must be taken over by the hearer (Clark and Schaefer 1989). If there are no problems, the acceptance of a dialog contribution is signalled by the hearer. When problems of understanding occur, these are signalled by one of the conversation participants and clarification requests and/or corrections may follow. In the easiest case, the hearer in a dialog signals understanding by choosing an appropriate, relevant following contribution. Since what is a "relevant next contribution" has been conventionalized in many cases, we find that dialog contributions can be well characterized by *adjacency pairs* (Clark and Schaefer 1989: 271), which are pairs of speech acts that often occur together in dialogs. The first part of the adjacency pair is the initiating act (for example, a question), while the second item in the pair provides the expected relevant reply (e.g., an answer).

Since the *kinds* of contributions made in a dialog are so important to characterize the conversation, dialog researchers have focused on the notion of *dialog acts*, an extension of the idea of speech acts (Austin 1975), but adapted to cover all possible linguistic contributions in dialog. The dialog act carried out by an utterance is the communicative function of that utterance, independent of the actual semantic content. Examples of dialog acts are Inform, Thank or Promise. The dialog acts that can be found in conversation depend on the type of conversations, and many different dialog act taxonomies exist, several of which have been used for extensive annotation studies of dialog acts in naturally occurring spoken conversation (Core and Allen 1997, Bunt et al. 2010).

Finally, it was noted early on in the literature that, because of the setting discussed above, spoken conversations typically contain specific linguistic features that are largely missing from written text, such as corrections, fillers and discourse particles. When contributions are not successful, this can be detected and rectified relatively quickly in conversation. Speakers use specialized markers to indicate the detection of communicative problems (mis- and non-understanding) and corrections of their own speech or the interlocutor's contributions. Fillers and particles are used to contribute non-truth conditional content in speech, in addition and in parallel to the at-issue meaning of the individual contributions. These items are said to be largely absent in written language, due to editing, planning, and genre restrictions (Rudolph 1991).

## 2.2 Spoken versus written media and CMC

It is clear that social media in general fall somewhere in between the prototypical poles of spontaneous spoken conversation and formal written text (Koch and

Oesterreicher 1985). But research points to the fact that conceptual orality cannot be captured as just one parameter on a continuous line, and that various linguistic phenomena reflect different aspects of speech-like linguistic contributions. For example, register studies following Biber (1993) distinguish several dimensions on which conversations and newspaper text differ: the informational/interactive dimension, the non-/narrative dimension, and so on. Each text type can then be situated along each of these dimensions, and the various forms of social media do not necessarily all group together. It is therefore interesting to study different types of social media, because it may allow us to distinguish which aspects of the context linguistic phenomena are facilitated or constrained by: e.g., informal style, interactive situation, real-world situatedness, synchronicity, etc.

German computer mediated communication has been the focus of several previous studies. Here, we only mention a few that touch upon the issues mentioned above. Beißwenger (2007) compares chats to spoken conversations, discussing the question of medial vs. conceptual orality, turn-taking, as well as the extra-linguistic action of deleting a drafted post. Chats closely resemble Twitter conversations, in that they are near real-time computer-mediated interactions (though some differences remain). In related work, Storrer (2013) investigates the conceptual orality continuum with regard to several computer mediated text types, and claims that the distinction between interactional and presentational writing is central in this context. This dimension distinguishes, for example, published Wikipedia articles (presentational) from the corresponding discussion pages (interactional). She points out that language adapts to the intended audience and topic and identifies differences in contribution lengths, and the use of computer mediated communication (CMC) specific items such as action words and emoticons. Similarly, (Storrer 2014) points out that there are large differences in language use within a medium based on the interactional style and the distance between speaker and hearer. A CMC medium cannot be considered a monolithic genre. Other studies identify linguistic phenomena that are specific to CMC (in German), or distinguish texts in these media from those in others (i.e. traditional newspaper texts) (Beißwenger 2013). Bartz et al. (2013) introduce a typology of such phenomena (across-the-board capitalization, emoticons, etc.) for use in the annotation of German CMC corpora. However, apart from colloquialisms, these items are not the focus of the current study. Here, we concentrate not on novel linguistic phenomena specific to social media, but on those features of spoken discourse that may also be found in the discourse carried out in Twitter conversations.

In this paper, we consider specifically the question of to what extent the dialog models that were developed for spoken conversations are applicable to written conversations on Twitter. We chose Twitter because its setting is most similar to spoken conversations among the major social media. Table 1 summarizes the main context properties of the linguistic contributions on the major social media platforms. All computer-mediated communications are available in written form. But while blogs are certainly written with a reader in mind, the production of blog posts does not in itself require a reader to be successful. Writing a blog is thus an individual action of a speaker, and while certainly informal, typically not interactional in nature. In contrast, forums, Facebook posts and tweets are more interactive in that they (at least in many cases) require an acceptance phase in Clark and Schaefer's (1989) use of the term, and thus constitute a joint action. These media also typically allow more than two participants in a conversation. There is a difference between blogs and Facebook on the one hand, and forums and Twitter on the other, in that the latter are common platforms where users interact, whereas in the former the platform (blog, Facebook page) belongs to one privileged user and the others are merely invited to "comment" on this page, yielding a power differential.

**Table 1: Interactive properties of a range of social media.**

| Property | Spoken | Blogs | Forums | Facebook | Twitter |
|---|---|---|---|---|---|
| **mode** | spoken | written | written | written | written |
| **action** | joint | individual | joint | joint | joint |
| **speakers** | 2+ | mainly one | many | many | many |
| **ownership** | common | single | common | single | common |
| **partic. status** | equal | unequal | equal | unequal | equal |
| **timing** | synchronous | asynch. | asynch. | near-synch. | near-synch. |
| **planning** | little | much | medium | little | little |
| **situatedness** | situated | online | online | online | online |

Further, the technical set-up and the way the media are consumed cause a difference in the timing of contributions and the amount of planning that goes into them. Spoken conversations happen in real time, speakers and hearers are synchronously active. As a result, there is very little time for planning utterances beforehand, and thus they are spontaneous in style. Even though writers on Facebook and Twitter are in principle able to access utterances later on, since they are written and remain on the platform, most conversations happen in near-real time. Individual utterances become unavailable quickly as they are "swamped out" of the timeline by new status updates from other users, especially on Twitter. In contrast, interactions on blogs and forums are centred around a topic of common interest, and span much longer time periods (as interlocutors return to the blog/forum to discuss topics of interest). It follows that these media allow more time for planning and editing contributions, with less pressure on timely responses. Finally, all social media differ from face-to-face conversations in that the latter are situated in a physical context that is the basis of grounding, and which can be referenced in the contributions. Instead, all social media are somewhat removed from any physical or often even previous social context of the interlocutors (the exception being private Facebook walls, where the conversation participants are usually known to each other). This can have effects on the linguistic means that must be chosen to make reference to people and events, and on the management of so-called common ground (Stalnaker 1978).

## 3 CONSTRUCTING A CORPUS OF TWITTER CONVERSATIONS

The overall communicative settings detailed in Table 1 show that, among the considered social media, Twitter is closest to conversational speech because it consists (at least in part) of conversations in near-real time, between two or more participants, who come together on an equal footing to jointly fulfil a communicative function. There are two main differences between spoken conversations and those on Twitter: the first is the spoken vs. written mode, and the second is the fact that face-to-face conversations are situated in a physical and social context, so that speakers can make reference to prior knowledge of the hearers or to objects and events that are easily inferable or apparent in the physical surroundings.

Twitter is a medium that allows users to post short “status messages”. Its contributors are private citizens, public institutions, and businesses, as well as bots that automatically post informational content, advertising, or jokes and memes. Since we are interested in the linguistic features exhibited on social media, with a focus on dialog, we would like to specifically extract tweets that are written by individuals (excluding for example press statements by organizations and companies as much as possible, as well as all tweets by bots), and that are part of larger conversations.

Unfortunately, Twitter’s API[2] does not make the extraction of entire conversations possible, and thus there has been limited computational linguistic research into Twitter conversations. In some cases, researchers have determined a set of users of interest and extracted all tweets by these, as well as by all their contacts (Ritter et al. 2010). This enables the reconstruction of conversations, including these seed users and some analyses. In this approach, the selection of users is crucial, and may restrict the general validity of any results. In contrast, we follow the approach proposed by Scheffler (2014) to construct a language-specific general Twitter corpus with a high recall, and then reconstruct all conversations contained in this general corpus. Since the Twitter API severely rate limits the

[2] https://dev.twitter.com/overview/api

number of tweets that can be extracted, this approach is only applicable to languages beyond the top five or so on Twitter: English, Spanish, Indonesian, Malay, and Japanese (Mocanu et al. 2013).

In the chosen approach, a stop word list of frequently occurring words in a language (in our case, German) is used to extract all tweets that contain these terms, using the Twitter API's *filter* keyword. The corpus examined in this work was created in April, 2013, using a precompiled stop word list for German with few manual corrections. The tweets are then filtered using the high-quality language identification module *langid*[3] (Lui and Baldwin 2012).[4]

The resulting dataset is estimated to contain > 90% of the German tweets sent during the time period. The conversation threads are reconstructed by following each tweet's *in-reply-to*-link in reverse (connecting a tweet to the one it was a reply to). This sorts all tweets into conversation threads. It must be noted that some threads may be incomplete for different reasons: (i) Tweets sent after the collection period are missing, even if they are in reply to existing conversations, because they were not included in the original dataset. (ii) A missing tweet somewhere within a conversation will lead to an erroneous split of the conversation into two subthreads. A tweet may be missing if it is not German, does not contain any of the stop words (e.g., is only a link), or was missed due to rate limiting by Twitter. In some cases, it is clear that a tweet is missing from the corpus because a subsequent tweet refers to it (by an *in-reply-to*-link). For those cases, we have attempted to re-fill the initial corpus by searching for these tweets specifically. This is a slow process due to rate limiting and not always successful, because users or tweets may have been deleted in the meantime.

The corpus was collected using the method described above from April 1–30, 2013, and is referred to as the "April13" corpus in the remainder of this work (Scheffler 2014). It contains 24,179,189 tweets from which we extracted 2,657,004 conversation threads (dialogs), consisting of 7,790,794 tweets, excluding the singletons. In this paper, we only consider conversations of at least length 2, i.e., that contain at least one reply in addition to the original tweet (we will call this the "TwitterDialogs," which is a new subcorpus studied for the first time in this paper). This restriction on conversations has the additional benefit of being a reliable filter for spam or automatic content. Typical bot tweets never receive any replies. To illustrate this effect, Table 2 shows the most frequent hashtags in

[3] https://github.com/saffsd/langid.py

[4] We have also created an improved stop word list for Twitter corpus extraction for German in collaboration with Nikolas Zoeller, FH Potsdam: We started with the 400 most frequent words in the large internet corpus deWaC , and manually removed a few obviously non-distinctively German words ('war', 'die'). We recorded all tweets retrieved using this list for two days (> 5 mio. tweets) and computed the ratio of German to non-German tweets using *langid* (confidence threshold: 0.85). A total of 27 words with a German/all-ratio < 0.2 were removed, to yield the final stop word list of 361 words. The list is available at https://github.com/TScheffler/TwitterCorpora.

the original April13 corpus compared with the most frequent hashtags in TwitterDialogs. The general corpus is dominated by automatic posts from mobile games (*#androidgames*, *#iphone*, etc.) and from other bots (*#pegelmv*, *#ostsee* originate with one bot posting water levels in the Baltic Sea). In contrast, the top ten hashtags used in dialogs reflect a few Twitter-specific items (*#ff* for "Follow Friday" recommendations, questions marked by *#followerpower*), but otherwise indicate important topics for discussions in the period and place when the data was collected: *#bvb* and *#fcb* denote popular soccer teams, *#piraten*, *#afd* and *#spd* are German political parties, *#tatort* is a popular TV crime show, and *#s21* and *#piratinnenkon* refer to prominent events during the collection time (a court investigation and a conference, respectively).

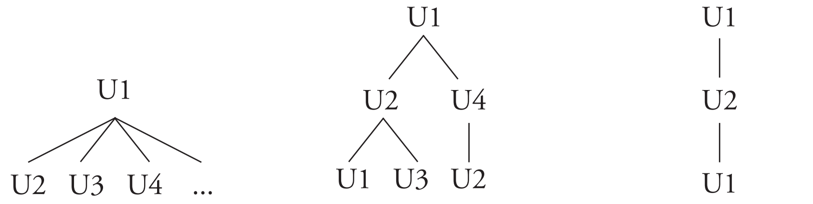
**Table 2: Most frequent hashtags in the April13 and TwitterDialogs corpora.**

| April13 | TwitterDialogs |
|---|---|
| #gameinsight | #ff |
| #android | #piraten |
| #androidgames | #bvb |
| #ipadgames | #afd |
| #ipad | #tatort |
| #pegelmv | #fcb |
| #ostsee | #spd |
| #iphone | #followerpower |
| #iphonegames | #s21 |
| #news | #piratinnenkon |

## 4 DIALOG STRUCTURE IN TWITTER

The resulting corpus includes (almost) all German Twitter threads during the sample month, but a closer look reveals that these are of different types. Visualizing the tree structure of these multilogs helps understand this. The tree structure of a conversation can be characterized by its size (the total number of tweets in the conversation), depth (defined as the length of the longest path from the root to a leaf, thus describing the longest conversation strand), and the number of users that take part in it. In some threads, one initial tweet receives hundreds of parallel answers, but no actual discussion ensues. This yields a conversation tree that is wide but whose depth is limited, possibly only to 2. We call those types of threads 'broadcasts,' since they often start with a statement by a (Twitter) celebrity which receives many responses from different people (see Figure 1(a)).

Note that this type of "conversation" cannot exist in face-to-face spoken dialog, since no contribution can receive hundreds of parallel replies. Linguistically, most broadcasts are very simple. An excerpt of a typical 'broadcast' thread is given in example (1). In this thread, 181 users reply to the 'Good morning, Germany' greeting by the actor Zach Braff, who has over 1.7 million followers.



*(a) Broadcast; depth=2. (b) Group discussion; d=3. (c) Linear conversation; d=3.*

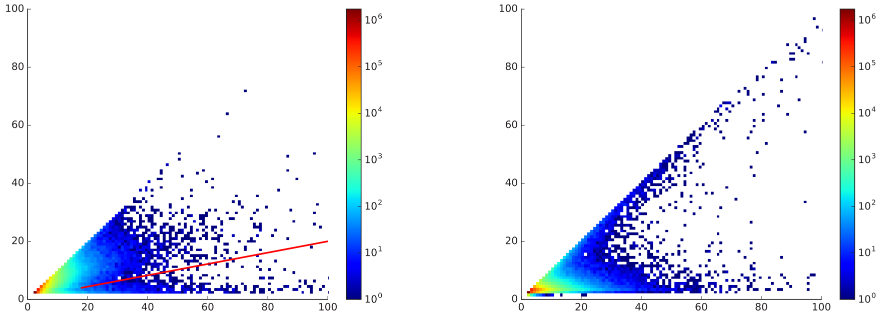
**Figure 1: Three different kinds of tree structure for threads.**

(1) Thread, size=182; maximum depth=2
@zachbraff: Guten Morgen Deutschland.
U2: @zachbraff oh ja, das ist gut!
U3: @zachbraff Guten Morgen, Zach Braff! Wie geht es Ihnen an diesem wunderschönen Tag?
U4: @zachbraff Guten Morgen mein süßes Schnitzel
U5: @zachbraff Guten Morgen Zach.
...[5]

Figure 2 shows 2D histograms of the size vs. depth and size vs. number of participants for all conversations in the corpus. In Figure 2, broadcast threads are along the x axis below the red line in plot (a), and along the diagonal in plot (b), which shows the number of distinct users that participated in each thread. Broadcast-type threads can have the properties of face-to-face conversations (such as question-answer pairs), but are unlike any spoken conversations in the number of participants (up to several hundred), and their short depth.

The second kind of threads on Twitter we call 'conversations.' If they are longer than 2 turns, their depth also increases, indicating that initial replies receive replies of their own, just like in spoken conversations. At the extreme (the diagonal in Figure 2(a)), the depth of the thread equals its size, so that the conversation consists entirely of a back-and-forth interchange between very few participants. In this case, the tree structure of the conversation is a linear chain, see Figure 1(c). Example (2) shows the start of an example linear conversation thread.

[5] @zachbraff: Good morning, Germany. U2: @zachbraff oh yeah, this is good! U3: @zachbraff Good morning, Zach Braff! How are you doing on this beautiful day? U4: @zachbraff Good morning my sweet dumpling. @zachbraff Good morning Zach.



*(a) Size vs. depth of conversations (b) Size vs. number of users in conversations.*

**Figure 2: Multilog structure in Twitter conversations (excluding a few longer threads).**

(2) Thread, size=103; maximum depth=28
U1: Kollers Klartext in den SN: “Es zahlt: Der Mittelstand”. http://t.co/Tpu3fGH4Wx schade, dass er nicht häufiger twittert @U2
U3: @U1 @U2 Die Abschaffung der Kapitalertragssteuer erscheint mir aber weder zweckmäßig noch den Mittelstand entlastend.
U1: @U3 nicht?
…[6]

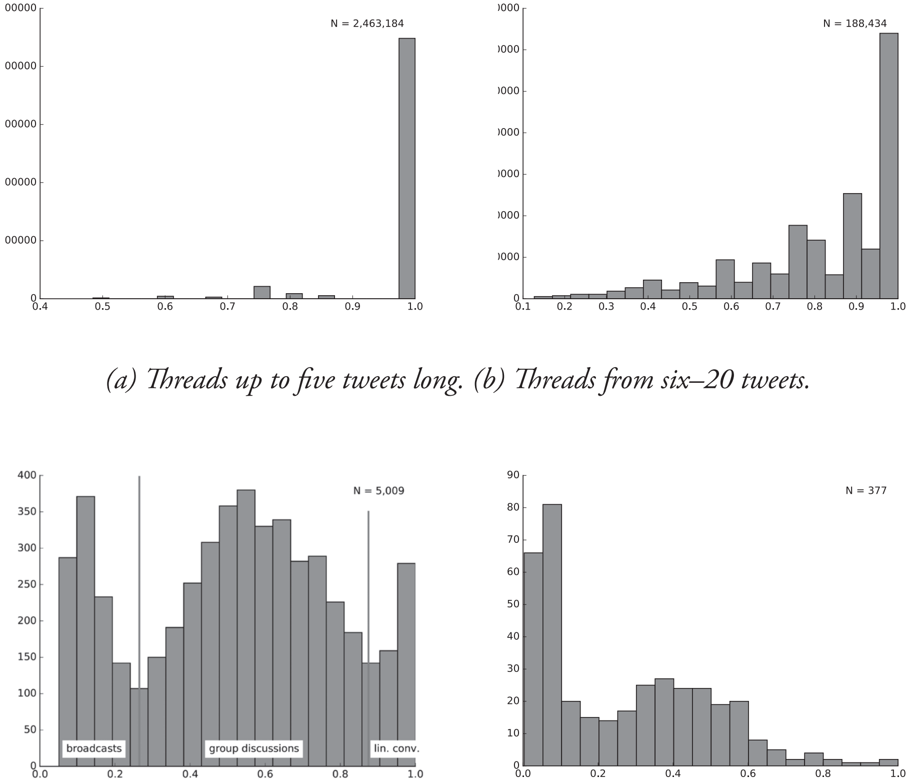
The diverse structure of threads becomes apparent when one analyses the angle of the vector pointing to the (x,y)-coordinates of each thread in the range of 0 to 1 from the size-axis to the diagonal. The equation is given in (3).

(3) $$z(x) = \frac{4}{\pi} \arctan \left( \frac{\mathrm{depth}(x)}{\mathrm{size}(x)} \right)$$

Figure 3 shows histograms of the factor z. It is clear from Subfigure (a) that shorter threads are overwhelmingly linear conversations. Very large threads are likely to be broadcasts with many replies but no depth (Subfigures (c) and (d)). Finally, threads with a medium angle (in the middle of the histograms) are likely to be group discussions, conversations with a relatively large size and medium depth, so they contain some branching structures (see Figure 1(b) for illustration). This diversity in the structure and nature of Twitter threads has implications for linguistic

[6] *U1: Koller says in SN: “The middle class has to pay” [link] Too bad that he doesn't tweet more @U2 — U3: @U1 @U2 Removing the capital gains tax doesn't seem useful or good for the middle class to me. — U1: @U3 it doesn't? — …*

analysis, since for example group discussions should be expected to be quite different from broadcasts in some respects. The red lines separating the broadcasts from the group discussions and linear conversations have been selected visually, but in future work the separation should be set algorithmically.



*(a) Threads up to five tweets long. (b) Threads from six–20 tweets.*

*(c) Threads from 21–50 tweets. (d) Threads over 50 tweets long.*

**Figure 3: Histograms of factor z relating size and depth for threads. N is the total number of threads pictured in each graph.**

## 5 LINGUISTIC PROPERTIES OF TWITTER DISCOURSES

In the following, we will consider some linguistic properties of Twitter conversations in turn, in order to determine their similarity and differences with spoken conversations.

## 5.1 Dialog Acts

In studying spoken conversations, dialog acts are often used to characterize their linguistic structure, topic composition, and type. For example, information exchanges contain many questions and answers, whereas argumentative exchanges include more agreements, disagreements, and so on. In earlier works (Zarisheva and Scheffler 2015, Scheffler and Zarisheva 2016) we annotated a set of 172 Twitter conversations (1,213 tweets) with 57 dialog acts from an adapted DIT++ schema (Bunt et al. 2010). The ten most frequent dialog acts found in Twitter conversations are shown in Table 3, along with the ten most frequent acts in the Switchboard telephone conversation corpus (Stolcke et al. 2000). The Twitter dialogs (we analysed a mix of long and short conversations) resemble spoken conversations in the way that declarative acts (Statement in the DAMSL schema, Inform and Information Providing in the Twitter schema) are by far the most frequent. Agreements and different types of questions also frequently occur in both kinds of conversations. However, spontaneous speech is characterized by Backchannels, Abandoned utterances and Non-Verbal material, which does not occur frequently in Twitter. Instead, the short length of most Twitter dialogs can be seen from the fact that Open[ing]s and TopicIntroductions can be found in the top ten dialog acts. In addition, the overall higher frequency of questions, agreements, and disagreements suggests a larger portion of informational and argumentative exchanges in the Twitter dialogs.

**Table 3: Dialog acts in the Switchboard telephone corpus and Twitter conversations.**

| Switchboard | | Twitter | |
|---|---|---|---|
| 36% | Statement | 25% | Inform |
| 19% | Backchannel | 11% | InformAnswer |
| 13% | Opinion | 9% | Agreement |
| 6% | Abandoned | 8% | SetQuestion |
| 5% | Agreement | 6% | Disagreement |
| 2% | Appreciation | 6% | PropQuestion |
| 2% | Yes-No-Question | 5% | Information-Providing |
| 2% | Non-Verbal | 3% | Correction |
| 1% | Yes-Answers | 3% | TopicIntroduction |
| 1% | Conventional-Closing | 3% | Open |

## 5.2 Questions

The dialog act analysis shows that questions are very common in Twitter conversations. Questions are an important marker of an interactional style (Storrer 2013), and are very rare in most written texts. All types of questions make up 18% of the utterances in the Twitter dialog act corpus. In contrast, the German newspaper commentary corpus PCC (Stede and Neumann 2014) contains only 75 questions in 2,900 sentences (2.6%).

There are a number of reasons for using questions on Twitter. While many questions are uttered to fill information gaps or ask for opinions, another typical use in conversation is for clarification, in order to initiate repair of communication problems. In German Twitter discussions, clarification questions are frequently marked by multiple question marks. (Purver et al. 2001) distinguish seven types of clarification questions. In an annotation study of 194 clarification questions from our corpus,[7] we found instances of all types except the rare gaps and gap fillers, which seem to depend on spoken interaction. Table 4 shows the prevalence of different types of clarification questions in Twitter conversations vs. the spoken conversations from the British National Corpus analysed in (Purver et al. 2001), with examples from our Twitter corpus. The linguistic means for marking clarification questions on Twitter resemble those used in spoken dialogs. Conventional phrases such as 'what?'/'really?' are frequently used, as are different types of reprise questions. Certain types of clarification questions that address a specific detail of the previous utterance (such as 'already?' as a reply to 'Should we pick you up?') do not fit any of the seven types of clarifications introduced in Purver et al. (2001). Finally, clarification questions on Twitter are sometimes marked solely with a range of question marks, without any further linguistic content. In speech, this may correspond to a confused facial expression and it could be seen as another (novel) conventional means of marking a clarification question on social media.

Even though the linguistic types of clarification questions found on Twitter resemble those in spoken conversation, their function is sometimes different. Since previous utterances are in the written mode and therefore persistent over time, clarification questions are not triggered by failure to hear/see what was said. Instead, questions like (6) are meant sarcastically or at a meta-level (= "Did you really mean to say what you just said?"). Many communication problems (and subsequent clarification questions) are due to the fact that it is hard to distinguish between sarcastic or ironic and literal utterances on Twitter. Many of the clarification questions thus tried to figure out whether the speaker meant what they said literally or was joking. Regular non-reprise clarification questions such as (7) can also be used for this purpose.

[7] Many thanks to Julia Gantzlin for annotating the data.

**Table 4: Types of clarification questions in Twitter and spoken conversation.**

| Type | BNC | Twitter | Example (Twitter) |
|---|---|---|---|
| Reprise fragments | 29.10% | 22.60% | (4) was ihr tun könnt??? Mich aus der insolvenz retten mir 150 tausend Euro überweisen!!!!<br>*what you can do??? Save me from bankruptcy wire me 150 thousand Euro* |
| Reprise sluices | 12.80% | 22.10% | (5) wieso heimlich??? Darf ruhig jeder wissen :D<br>*why secretly??? Anybody can know it :D* |
| Reprise sentences | 8.90% | 1.00% | (6) die Erde ist rund??? Oh Oh das musste schon mal jemand zurück nehmen!<br>*the Earth is round??? Uh oh someone had to take that back before!* |
| Non-reprise clarifications | 13.30% | 15.50% | (7) wie meinst du das?<br>*how do you mean?* |
| Gaps | 0.50% | 0% | |
| Gap fillers | 3.80% | 0% | |
| Conventional | 30.70% | 30.90% | (8) hä??? Eher overgedressed<br>*whaaa??? More like overdressed* |
| Question marks | – | 4.00% | (9) ?????????????? ich komm hier jetzt gar nicht mehr mit....<br>*?????????????? I can't keep up here....* |
| Others | – | 3.60% | (10) [sollen wir dich abholen? —] jetzt schon??<br>*[should we pick you up? —] already??* |

## 5.3 Particles

According to the literature, German modal particles are a phenomenon that is mainly found in spoken language (Bross 2012). Though the use of particles has a colloquial feel, it is not immediately clear whether the use of modal particles depends on the spoken medium, colloquial style, or interactional vs. informational types of conversation. Here, we compare the occurrence of modal particles in the Twitter conversations with the German newspaper corpus PCC and the spoken-like (though edited) OpenSubtitles[8] corpus (Lison and Tiedemann 2016). We study the 17 common modal particles listed in König (1997). In the newspaper commentaries, these particles make up 3.2% of (non-punctuation) tokens. In the Twitter conversations, they are more common, accounting for 4.4% of tokens. This is true despite the fact that these conversations contain many additional Twitter-specific tokens, such as user names and URLs, that inflate the token count. Particles make up 2.9% of tokens in the subtitles corpus.

[8] http://www.opensubtitles.org/

The distribution of particles among the three corpora is shown in Figure 4, which shows the occurrence frequency relative to the number of (non-punctuation) tokens in the corpora. It can be seen that the particle 'ja' in particular is much more frequent in Twitter and OpenSubtitle conversations. This is due to the fact that this item is used as the answer particle '*yes*' as well as a modal particle. In addition, 'aber' (*however*), 'auch' (*also*), 'halt' (*just*), and 'schon' (*already*) are also more frequent on Twitter. Other particles, such as 'doch' (*however*), 'wohl' (*possibly*), and especially 'nun' (*now*) may in fact be more typical of written language and/or informational style than conversations. It seems, therefore, that a blanket statement to the effect that modal particles are generally more frequent in speech (or spoken-like social media) is unsupported based on this data. Different particles show very different profiles depending on the context of the communicative situation.

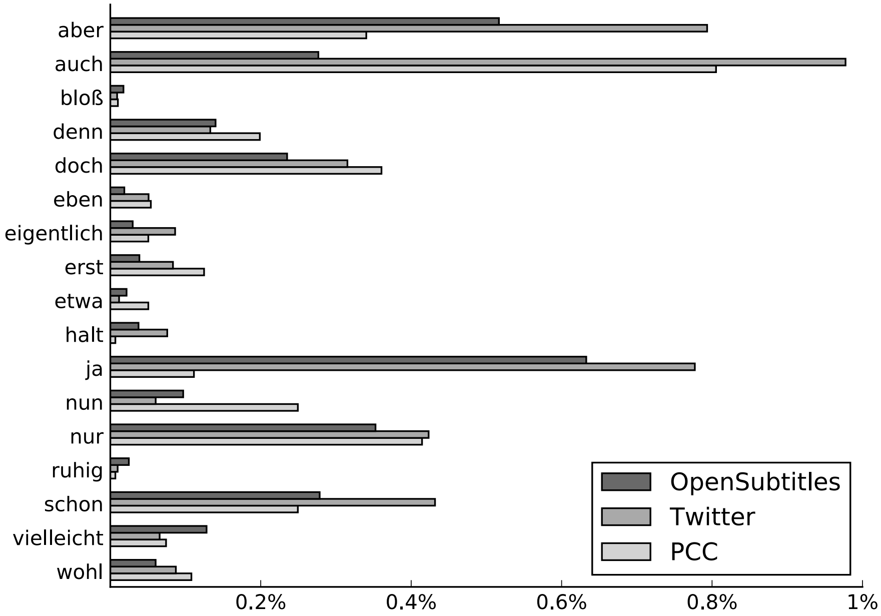


**Figure 4: Frequency of modal particles in Twitter, scripted speech (OpenSubtitles) and written newspaper text (PCC).**

## 5.4 Intensifiers

The use of intensifiers such as 'really' and 'very' is associated with informal and colloquial registers, in particular spoken conversations. Tagliamonte and Denis

(2008) analyse speech and IM text messages from Toronto teenagers and show that intensifiers also occur frequently in the text messages, though slightly less often than in speech. But they also note that the choice of intensifier depends on the medium. In text messaging, the teenagers prefer the innovative variant 'so' over formal 'very' and informal 'really,' whereas 'really' is the most frequent variant in speech.

Here, we look at the use of formal and informal intensifiers in the German Twitter conversations vs. newspaper texts. First, the expectation that intensifiers are more common in conversations carries over to the Twitter data. In the Twitter dialogs, 0.46% of all tokens are intensifiers. In the newspaper commentaries, intensifiers only amount to 0.14% of tokens. Next, we compare the use of formal vs. informal intensifiers given in (11) and (12), respectively. Formal intensifiers are relatively more frequent in the texts, accounting for 65% of all intensifiers. In Twitter conversations, the informal variants account for about the same number of intensifiers as the formal variants (50%; see Figure 5). But interestingly, the formal variants are still very common here as well. In future work this should be compared to spoken data, or that obtained from other social media.

(11) formal: wirklich ('really'), sehr ('very'), absolut ('absolutely')

(12) informal: echt ('really'), krass, extrem ('extremely'), ordentlich, total ('completely'), sau, voll, völlig ('completely')

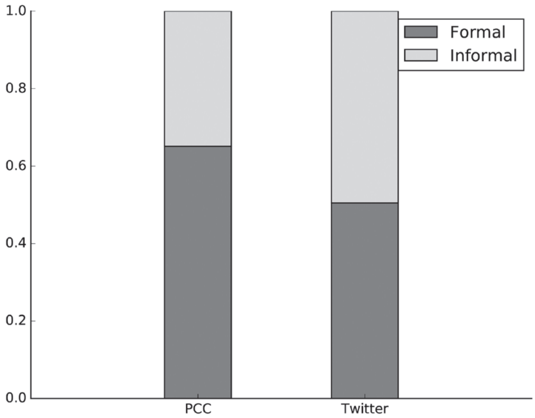


**Figure 5: Ratio of formal and informal intensifiers in newspaper text vs. Twitter conversations.**

## 6 CONCLUSION

In this paper, we have provided a view of one particular set of social media data, Twitter conversations. These conversations are computer-mediated and thus come in written form, but otherwise resemble spoken conversations in structural respects. The participants in Twitter conversations are not restricted in number and this can change throughout the conversation, just like in face-to-face interactions. The participants are furthermore relatively equal in standing, and make their utterances spontaneously and in a relatively short time span (though not synchronously, as in spoken conversations). Since successful communication is a joint action, speakers and hearers must coordinate to achieve their common communicative goals. This coordination process can be observed through adjacency pairs (or dialog act sequences) and other grounding phenomena, such as corrections and clarification questions.

The Twitter dialogs considered here exhibit all the linguistic markers typically attributed to face-to-face conversations, though some differences can be found. On the one hand, the most prominent dialog acts in Twitter conversations are informational, just like in speech. But due to the very short length of many Twitter threads, openings and topic introductions are also more frequent in the Twitter corpus. In addition, a subset of Twitter discussions is clearly argumentative, which leads to a slightly higher portion of agreements and disagreements. On the other hand, common phenomena of unplanned spontaneous speech, such as backchannels and fragments, are almost completely missing from Twitter conversations. Rehbein (2015) uses the example of filled pauses, and demonstrates that when such speech-specific phenomena are present on Twitter, they are used deliberately to carry extra-propositional meaning.

Based on the analyses shown here, computer mediated conversations can be interesting data sources for some linguistic phenomena that are specific to informal conversation, but difficult to study in spoken corpora. We have shown that, for example, questions are very frequent in the Twitter threads, but not in newspaper corpora. The case of clarification questions furthermore underlines the joint communicative action between speakers and hearers, as these instances highlight cases where communication breaks down because of mis- or non-understandings. Twitter users avail themselves of the same linguistic means to mark clarification questions, but they add an innovative variant thanks to the written mode, an indication of non-understanding with only a series of question marks.

Despite the similarities, it is not the case that Twitter conversations are just written versions of spoken dialogs. As expected, particles and intensifiers are found frequently in Twitter conversations as features of informal, colloquial language.

In this respect, the CMC conversations differ markedly from standard newspaper corpora in both the frequency and range of items that are used. But it is to be expected that the use of these linguistic items also differs from their use in speech corpora, as shown for English intensifiers by Tagliamonte and Denis (2008). Further work is thus needed to situate Twitter conversations (and other social media) on the 'conceptual orality' continuum and determine the mix of conservative and innovative features that can be observed.

Finally, we showed through an analysis of the dialog structure of Twitter conversations that even within this medium, different types of conversations must be distinguished. This distinction was made on structural grounds, not based on topic or linguistic features (which could make the definition circular). While most conversations are very short (typically, only one root plus a reply), longer conversations belong to three broad classes: 'Broadcasts' contain root tweets which get many replies (usually from different users) but do not lead to any further discussion; they are characterized by a short depth and are often linguistically less complex. 'Linear conversations' are private discussions among a very small number of users, which develop in a linear fashion, i.e. each answer is a reply to the last contribution. Finally, there is a number of conversations in between the two extremes, exhibiting some branching of the dialog tree. We called these 'group discussions'. All conversation data from Twitter is much less likely to contain bot generated content than a random set of tweets, which makes it very amenable to linguistic research.

In sum, Twitter conversations are made up of informal, interactive exchanges between speakers which allow us to tease apart the differences between highly edited, monological text and spontaneous, colloquial speech on several dimensions. This will enable more detailed studies of linguistic phenomena across different traditional and computer-mediated channels of communication.

## References

Austin, John L., 1975: *How to Do Things with Words*. Oxford University Press.

Bartz, Thomas, Michael Beißwenger and Angelika Storrer, 2013: Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL* 28/1/. 157–98.

Beißwenger, Michael, 2007: *Sprachhandlungskoordination in Der Chat-Kommunikation*. Berlin: De Gruyter.

Beißwenger, Michael, 2013: Das Dortmunder Chat-Korpus: Ein Annotiertes Korpus Zur Sprachverwendung Und Sprachlichen Variation in Der Deutschsprachigen Chat-Kommunikation. *LINSE-Linguistik Server Essen*. 1–13.

Biber, Douglas, 1993: The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26/5-6. Springer. 331–45.

Bross, Fabian, 2012: German Modal Particles and the Common Ground. *Helikon: a Multidisciplinary Online Journal* 2. 182–209.

Bunt, Harry, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria and David Traum, 2010: Towards an ISO Standard for Dialogue Act Annotation. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. 2548–2555.

Clark, Herbert H. and Edward F. Schaefer, 1987: Collaborating on Contributions to Conversations. *Language and Cognitive Processes* 2/1. Taylor & Francis. 19–41.

Clark, Herbert H. and Edward F. Schaefer, 1989: Contributing to Discourse. *Cognitive Science* 13/2. Wiley Online Library. 259–94.

Core, Mark and James Allen, 1997: Coding Dialogs with the DAMSL Annotation Scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*. 28–35.

Koch, Peter and Wulf Oesterreicher, 1985: Sprache Der Nähe–Sprache Der Distanz. *Romanistisches Jahrbuch* 36/85/. 15–43.

König, Ekkehard, 1997: Zur Bedeutung von Modalpartikeln im Deutschen: Ein Neuansatz im Rahmen der Relevanztheorie. *Germanistische Linguistik* 136/1997. 57–75.

Lison, Pierre and Jörg Tiedemann, 2016: Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. 923–929.

Lui, Marco and Timothy Baldwin, 2012: langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 25–30.

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang and Alessandro Vespignani, 2013: The Twitter of Babel: Mapping World Languages Through Microblogging Platforms. *PloS One* 8/4.

Purver, Matthew, Jonathan Ginzburg and Patrick Healey, 2001: On the Means for Clarification in Dialogue. *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*. 116–25.

Rehbein, Ines, 2015: Filled Pauses in User-Generated Content Are Words with Extra-Propositional Meaning. *Proceedings of ExProM*. 12–21.

Ritter, Alan, Colin Cherry and Bill Dolan, 2010: Unsupervised Modeling of Twitter Conversations. *Proceedings of NAACL*. 172–180.

Rudolph, Elisabeth, 1991: Relationships Between Particle Occurrence and Text Type. *Multilingua* 10. 203–23.

Scheffler, Tatjana, 2014: A German Twitter Snapshot. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. 2284–2289.

Scheffler, Tatjana and Elina Zarisheva, 2016: Dialog Act Recognition for Twitter Conversations. *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. 31–38.

Stalnaker, Robert, 1978: Assertion. Cole, Peter (ed.): *Syntax and Semantics 9: Pragmatics*. New York: Academic Press.

Stede, Manfred and Arne Neumann, 2014: Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. 925–929.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema and Marie Meteer, 2000: Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26/3. Cambridge, MA: MIT Press. 339–73.

Storrer, Angelika, 2013: Sprachstil Und Sprachvariation in Sozialen Netzwerken. Frank-Job, Barbara, Alexander Mehler and Tilmann Sutter (eds.): *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Storrer, Angelika, 2014: Sprachverfall Durch Internetbasierte Kommunikation: Linguistische Erklärungsansätze – Empirische Befunde. *Sprachverfall?: Dynamik–Wandel–Variation (Jahrbuch des IDS)*. 171–96.

Tagliamonte, Sali A. and Derek Denis, 2008: Linguistic Ruin? LOL! Instant Messaging and Teen Language. *American Speech* 83/1. 3–34.

Zarisheva, Elina and Tatjana Scheffler, 2015: Dialog Act Annotation for Twitter Conversations. *Proceedings of SIGDial16*. 114–23. Prague, Czech Republic: Association for Computational Linguistics. http://aclweb.org/anthology/W15-4614. (Last accessed 29 June 2017.)