Dialog Act Annotation for Twitter Conversations

Elina Zarisheva
Hasso-Plattner-Institut
Potsdam, Germany
elina.zarisheva@
student.hpi.uni-potsdam.de

Tatjana Scheffler
Department of Linguistics
University of Potsdam, Germany
tatjana.scheffler@
uni-potsdam.de

Abstract

We present a dialog act annotation for German Twitter conversations. In this paper, we describe our annotation effort of a corpus of German Twitter conversations using a full schema of 57 dialog acts, with a moderate inter-annotator agreement of multi-π $= 0.56$ for three untrained annotators. This translates to an agreement of 0.76 for a minimal set of 10 broad dialog acts, comparable to previous work. Based on multiple annotations, we construct a merged gold standard, backing off to broader categories when needed. We draw conclusions wrt. the structure of Twitter conversations and the problems they pose for dialog act characterization.

1 Introduction

Social media and particularly Twitter have become a central data source for natural language processing methods and applications in recent years. One issue that has not received much attention yet, is the *social* or *interactive* nature of many posts. Often, only individual tweets are analyzed in isolation, ignoring the links between posts.[1] However, it is known that up to 40% of all Twitter messages are part of conversations—(Scheffler, 2014) report that 21.2% of all tweets in their German corpus are replies. In this paper, we view tweets in their original dialog context and apply a dialog annotation scheme to analyze the function of Twitter utterances. To our knowledge, this is the first attempt to apply a detailed dialog act annotation to Twitter dialogs[2].

We view our work as a first step in studying the make-up of Twitter conversations. So far, not much is known about the types of conversations that occur there, since the focus has been on analyzing single tweets. Our guiding question is in which way Twitter dialogs differ from the relatively well-studied genres of human-human and human-machine spoken dialogs. In this paper, we apply dialog act annotation because it captures the functional relevance of an utterance in context. This will enable us to answer questions about the nature of discourse on social media, such as whether individuals from different opinion "camps" talk with each other, whether Twitter dialogs are just exchanges of opinions and emotions, or whether true argumentation is taking place, etc. In addition, dialog act annotations are useful for further research on Twitter dialogs, as well as for applications dealing with this kind of data, e.g., automatic analyses of conversations on different types of topics, or simulated conversation participants (Twitter bots). We address both practical issues related to applying dialog act annotation to tweets as well as theoretical implications about the nature of (German) Twitter conversations that can be gleaned from our annotated data.

2 Related Work

In the following, we briefly summarize the relevant previous literature on dialog act annotation for other media, and existing research on Twitter dialogs in general.

Dialog act annotation One of the first steps towards analyzing the structure of dialogs is dialog act (DA) annotation. Dialog acts, a notion based on Austin's speech acts (Austin, 1975), characterize the dialog function of an utterance in broad terms, independent of its individual semantic content. There is a large number of DA schemata for conversational and task-based interactions (Core and Allen, 1997; Bunt et al., 2010; Traum, 2000, among many others), and these taxonomies have

[1] Usually, this is done by necessity, as Twitter data is most commonly accessed through an API stream that provides a random 1% of public statuses.

[2] really, multilogs, but we use the term broadly here

Proceedings of the SIGDIAL 2015 Conference, pages 114–123,
Prague, Czech Republic, 2-4 September 2015.

been applied to the construction of annotated corpora of human-human dialogs such as the Map-Task corpus (Carletta et al., 1997), Verbmobil corpus (Jekat et al., 1995), or the AMI meeting corpus (McCowan et al., 2005). DA taxonomies and annotated resources have also been used in automatic DA recognition efforts (Stolcke et al., 2000, and many others). Dialog act annotation has also been carried out for some types of social media. (Forsyth and Martell, 2007) annotated chat messages with a custom-made schema of 15 dialog acts, and built a dialog act recognizer. They consider each turn to correspond to only one DA, even though they note that several acts can appear within one turn in their data. However, Twitter conversations have only recently become of interest to researchers.

Twitter conversations Twitter data is a mix of different genres and styles. But users are generally able to reply to existing messages, producing either personal discussions or interactions with strangers. Up to a quarter of tweets are replies to other messages (Scheffler, 2014; Honey and Herring, 2009), and due to the log-scale length distribution of conversations (most are just one tweet + its answer (Ritter et al., 2010)), around 40% of tweets thus are a part of conversations.

There are few studies that analyze Twitter dialogs, most likely because connected conversational data cannot easily be obtained through the Twitter API. Studies concentrate on samples based on individual, random users (Ritter et al., 2010) or based on frequently-updated snapshots over a short time-scale (Honey and Herring, 2009). We know of only two previous studies that address dialog acts in Twitter conversations. (Ritter et al., 2010) train an unsupervised model of dialog acts from Twitter data. Their system learns 8 dialog acts that were manually inspected and received labels such as STATUS, QUESTION, REACTION, COMMENT, etc. They also obtain an informative transition model between DAs from their data.

In contrast, (Zhang et al., 2011) build a supervised system that can classify between 5 broad speech acts (STATEMENT, QUESTION, SUGGESTION, COMMENT, MISC), using 8613 hand-annotated tweets to train their model. However, this work uses disconnected tweets in isolation (disregarding the underlying dialog structure). They do not report on inter-annotator agreement. Further, both this work and (Ritter et al., 2010) also assume that each tweet can be characterized by exactly one dialog act. We will show that this is not borne out in our data.

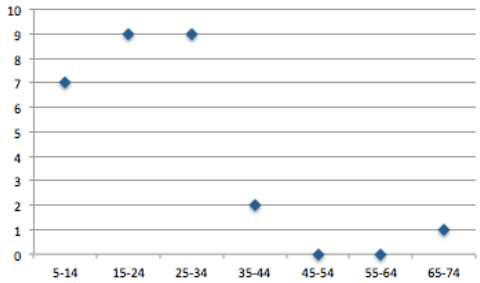


Figure 1: Distribution of depth in long conversations. X axis shows binned depth, values = number of conversations in the corpus.

3 Dialog Act Annotation

3.1 Corpus

For our work we use Twitter data that was collected within the BMBF project *Analysis of Discourses in Social Media*[3]. In the scope of this project, social media data concerning the topic *Energiewende* (energy turnaround) from Twitter and other sources was collected during the months of Aug-Nov, 2013. During November 11-30, Twitter conversations were automatically completed by re-crawling. Each conversation (= thread) can be represented as a tree with the first tweet as root node, and the edges between tweets drawn according to the `in_reply_to_status_id` field. The thread's *length* or size is the total number of tweets in the thread, its *depth* is the maximum level of embedding of a tweet (= the tree depth). Since we assume that the dialog structure of long Twitter discussions might differ from short interactions (which comprise the bulk of Twitter conversations), we extracted our corpus from the available data according to the two following criteria:

1. all *long* conversations of more than 20 tweets and minimum depth 5;
2. a random selection of *short* conversations of 4-5 tweets and arbitrary depth.

The total number of tweets is 1566, grouped in 172 dialogs. Figure 1 shows the depth distribution of long conversations.

[3]http://www.social-media-analytics.org/

For 18 tweets the text is missing: either they were deleted or they originate from a private account. To filter out non-German tweets we used the langid (Lui and Baldwin, 2012) and Compact Language Detection[4] libraries for Python 2.7, with some manual correction. 1271 tweets were recognized as German by both packages. Further problems with the raw and annotated data and our cleaning steps are described in Section 4.

3.2 Schema

We based our DA annotation schema on the general-purpose DIT++ taxonomy for dialog acts (Bunt et al., 2010)[5]. Twitter conversations are a type of human-human, non-task-oriented dialog. Many existing DA taxonomies are more suitable for task-oriented dialogs (even DIT++ has a very limited range of non-task-oriented acts) or for human-machine dialog. In order to reflect the type of interactions we expected in our data, and to reduce the difficulty of the annotation task, we changed the DIT++ schema according to our needs. Our adapted DA schema is shown in Figure 3 in the Appendix. In many places, the DA hierarchy was simplified by removing the finest distinctions, which are either hard to judge for novice annotators (e.g., subtypes of directives), or can be recovered from other properties of the data (e.g., types of check questions). We only included DAs from the dimensions Information Transfer, Action Discussion, and Social, as well as selected items from Discourse Structure Management and Communication Management. Even though the dimensions are in principle often independent of each other, we instructed the annotators to assign only the most relevant DA label to each segment.

3.3 Annotation task, annotators, tool

In recent years, crowdsourcing annotations has become ever more popular in linguistics. This approach is useful for quickly creating new resources based on newly available data (like the Twitter conversations we use). However, dialog act segmentation and labelling is a relatively complex task that is not easily done by untrained volunteers. For example, the taxonomy needs to be explained and internalized, and native knowledge of German is required. For this reason we used minimally trained undergraduate linguistics students as annotators for this study. The 36 students were participants of a Fall 2014 seminar on *Dialogs on Twitter* at the University of Potsdam, and received instruction on dialog acts as well as an overview of the DIT++ and other annotation schemes.

The students viewed entire conversations and were asked to segment each tweet (if necessary) into individual dialog acts and assign a DA label from the presented taxonomy. We used the WebAnno framework (Yimam et al., 2013), a free, web-based application that is especially easy to use for novice annotators. Although there were some technical problems with the tool (difficulty deleting annotations, the ability of annotators to add new labels), it was generally well-suited to the basic span-labelling annotation we required.

Each conversation in the corpus was assigned to three annotators, but no two annotators worked on the exact same set of conversations. For each annotator, WebAnno provides a token-based B-I label format as output, which is the basis of further analysis in this paper.

4 Annotation Validation

In this section we discuss initial steps to cleaning the raw annotation data and an evaluation of the quality of annotations.

4.1 Pre-processing

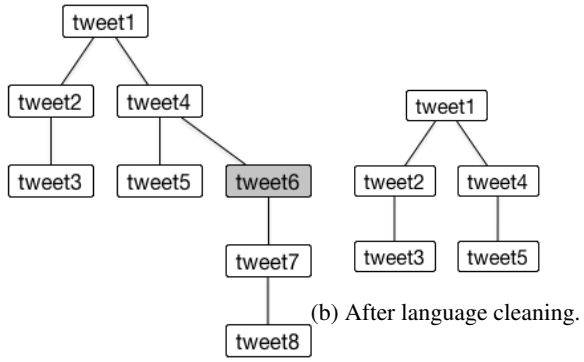
Before further analysis steps are possible, some cleaning steps were necessary. Although we designed the schema in a such way that tags are unambiguous, some tokens were assigned several tags by the same annotator. There are 122 tweets with ambiguous annotations. Unless one annotation was removed for another reason (see below), these additional annotations were retained during the construction of the gold standard.

In Section 3 we discussed that 1271 tweets of 1566 were classified as German. The other tweets were checked manually, so that only 106 tweets were deemed non-German and had to be excluded. We rebuilt the conversations by deleting non-German tweets, as well as all their replies (see Figure 2). After rebuilding, 1213 German tweets remain in the corpus.

As a second step, we standardized the annotations of @-tagged user names at the start of tweets, which mark the tweet as a reply to that user's tweet. Some annotators have included these @-tags in the following dialog act, others have not

[4] https://code.google.com/p/cld2/

[5] http://dit.uvt.nl



(a) Before language cleaning.

(b) After language cleaning.

Figure 2: Twitter conversation with non-German tweets (in gray) before and after cleaning.

tagged these at all. We decided to delete all tags for all user names at the start of the tweet. For this case we introduced a new label 0, indicating that there is no DA tag for this particular token.

The third step was to delete faulty annotations. In the annotations we found four "dialog act" labels that are not included in our DA schema and had been introduced by annotators: IRONIE (irony), NEIN (no),WURST (sausage) tags and the O- label (Table 1).

Tags and labels	Number of tweets
O-	51
IRONIE	72
NEIN	1
WURST	3

Table 1: Odd tags

We deleted these odd tags. In some cases (e.g., irony), an annotator also assigned a proper label to the token, which then remains as the sole annotation. In other cases, the token becomes untagged (marked with 0) for this annotator, resulting in missing annotations.

4.2 Segmentation

In order to evaluate the quality of the annotation and the chosen schema, we have separately determined the inter-annotator agreement for the segmentation and dialog act labelling steps.

Several of the proposed methods for determining the validity of annotations are based on comparing two annotations with each other (i.e., one candidate annotation with a gold standard). Even when more annotators can be included, it is often assumed that those annotators have worked on the same data, as for example with the popular Cohen's κ-statistic (Carletta, 1996). Instead, we chose *Fleiss' multi-*π, which measures how consistent the assigned labels are for each item, without regard to which annotator gave the label (Artstein and Poesio, 2008). In order to be able to use this metric, which nevertheless assumes a fixed number of annotations per item, we include in our validation only those tweets for which we have three annotations after the cleaning steps described above (1004 tweets). We exclude tweets with missing annotations and those where removal of spurious labels resulted in missing annotations for some tokens.

The overall observed agreement is the mean of the individual agreement values for each item:

$$agr_i = \frac{1}{\binom{c}{2}} \sum_{k \in K} \binom{n_{ik}}{2} \tag{1}$$

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i \tag{2}$$

where agr_i is the relative frequency of agreeing judgment pairs among all pairs of judgments, I the number of taggable items in the corpus, k the number of tags in the schema, and $c = 3$ the number of annotators (Artstein and Poesio, 2008, p. 563).

The overall expected agreement is calculated as the random chance event that two annotators assign an item to the same category/DA k (4). Each annotator's chance of assigning an item to k is based on the overall proportion $\hat{P}(k)$ of items assigned to k, n_k, over all assignments.

$$\hat{P}(k) = \frac{n_k}{ic} \tag{3}$$

$$A_e^\pi = \sum_{k \in K} (\hat{P}(k))^2 \tag{4}$$

We calculate the amount of agreement beyond chance by the standard formula:

$$S_\pi = \frac{A_o - A_e}{1 - A_e} \tag{5}$$

For the segmentation task, we used the simplest approach by taking each token to be a taggable item which can be labelled either a BOUNDARY or NON-BOUNDARY. As discussed in (Fournier

and Inkpen, 2012), such measures are too strict by punishing even small disagreements over the exact location of a segment boundary (e.g., if annotators disagree by one token). In addition, since most judgments fall into the majority class (NON-BOUNDARY), the expected agreement will be high, making it harder to improve upon it. However, we show in Section 5.3 that the DA segments in our Twitter data are relatively short on average, possibly partially relieving this problem. Consequently, the agreement determined this way can be seen as a lower limit that underestimates the actual agreement between annotators.

We observe a segmentation agreement of 0.88 between three annotators, which indicates very good agreement. Disagreements are due to additional segments that some annotators posited (= Does an explanation after a question constitute its own speech act?) or were triggered by special Twitter vocabulary such as emoticons, to which some annotators assigned their own DA labels (see example (6) on page 8). Some of these disagreements can be solved by more comprehensive annotation guidelines.

	Segment.	**DA labelling**
A_o	0.966	0.658
A_e^π	0.716	0.224
Fleiss' multi-π	**0.883**	**0.559**

Table 2: Chance-corrected coefficient between three annotators for segmentation and DA labelling tasks.

4.3 DA labelling

We then computed the inter-annotator agreement for DA labels on the raw annotation data, using the same procedure. For this measure, we only included those tweets where all three annotators agreed on the segmentation. The results for the full DA schema of 57 dialog acts are shown in Table 2. As such, the agreement on DA labels is at most moderate, but the measure does not take the DA taxonomy into account. For example, disagreements on a subtype of QUESTION are counted as one error, just like a mix-up between top-level DA labels would be. Other annotation efforts report even worse IAA values with novice annotators, even using a weighted agreement score (Geertzen et al., 2008). In order to better compare our annotation effort to other work, we also computed agreement scores for two reduced DA schemas by merging similar DAs. With a reduced set of 14 DAs, three annotators achieve multi-$\pi = 0.65$, whereas a minimal DA set of 10 basic DAs yields multi-$\pi = 0.76$, a good agreement.

To better evaluate the chosen DA schema we built a confusion matrix, recording the DA labels that caused the most disagreements. The great majority of disagreements occurred within the different subtypes of INFORMATION PROVIDING functions. In addition, there were 36 cases of confusion between INFORM and the discourse structuring functions OPEN, TOPICINTRODUCTION and TOPICSHIFT. These errors indicate a limited applicability of the chosen schema to conversational Twitter data. The INFORM category is too broad for conversational statements, and annotators thus had two kinds of problems: First, clearly delineating plain INFORMs from other dialog moves that may be carried out simultaneously (like the discourse structuring moves or social moves), and second, deciding whether a statement can be classified as INFORM at all—in cases of doubt, annotators may have chosen the higher level label INFORMATION PROVIDING but not INFORM. We discuss this issue further in Section 6.

Another source of multiple disagreements is the distinction between different types of questions. These confusions are true errors than can be corrected with better training of annotators.

In contrast, there were no systematic cases of confusion between between the ACTION DISCUSSION, INFORMATION TRANSFER, and SOCIAL functions. Tables 8 and 9 in the Appendix show the frequencies of confusion between DA labels.

5 Analysis

The evaluation in the previous section has shown that (i) about two-thirds of judgment pairs on individual items are in agreement (i.e., on average, two out of the three annotators agree), and (ii) most disagreements between annotators exist in the lower tiers of the annotation schema, whereas the agreement on broader categories is better. Based on these observations, we devised an algorithm to automatically merge the annotations into a gold standard.

5.1 Merging annotations

As was mentioned in Section 3, each tweet should be annotated by three students, in principle provid-

ing a possibility to use majority voting (the most common decision tactic in crowdsourced lay annotations (Sabou et al., 2014)) to decide on the 'correct' annotation. However, since the annotators carry out two tasks simultaneously (segmenting and labelling), merging became less trivial. If we first merge the segmentations we would lose DA information. Instead we observe tag variations for a particular word token and determine the true tag based on the results.

In the raw data there were 1004 tweets annotated by three students, 180 tweets – by two, 29 – only by one. Moreover, some tokens have received more than one label even by the same annotator (contrary to the guidelines). Therefore we adapted our algorithm to differing numbers of annotations.

The merging process is composed of three steps. For this phase, we disregard segmentation boundaries because there are no tweets with several successive segments with the same tag. We can recognize segment boundaries by simply observing the tag change.

First step: Perfect agreement We find all tweets that have exactly the same segmentation for all their annotators (405 unique tweets). Among these, 82 tweets have the same annotation as well. Since there is already perfect agreement for these tweets, no further work is required.

Second step: Majority vote In this step we pick one tag from several for a particular token. For each occurrence of a tag we assign weight 1. Tags whose weight is higher than the sum of weights for other tags are deemed 'correct' and assigned to that token.

For example, the word *Erde* has been assigned INFORM once, tag DIRECTIVE once, QUESTION three times. Since $3 > 2$, we keep QUESTION and the other tags are deleted. After this step, another 421 tweets have no ambiguous tokens left and can be added to the 'done' tweets from the first step.

Third step: DA generalization Our DA taxonomy has a tree structure, viz., some DA labels have the same ancestor, or one tag is a child of another. In this phase we compare tags for a particular token based on their relationship in the DA hierarchy. In the DIT++ taxonomy, it is assumed that parent DAs subsume the function of all children (they indicate more general dialog functions). In case of inapplicability of all the leaf-level labels, or in case the annotator isn't sure, a higher-level DA label can be chosen from the hierarchy. In this step, we use this structure of the DA taxonomy in order to capture some of the information that annotators agreed upon when labelling tweets.

If DA tags for a token are in a direct inheritance (parent-child) relationship or siblings, we choose the parent tag for this token. The other tags that take part in this relationship are deleted (they are replaced by the higher-level option). Below is an example of the two scenarios.

Patent-child relationship:

Tag IT_IP_INFORM_AGREEMENT and parent tag IT_IP_INFORM. Parent tag IT_IP_INFORM is kept and child is deleted.

Siblings:

Tag IT_IP_INFORM_AGREEMENT and tag IT_IP_INFORM_DISAGREEMENT both have the parent tag IT_IP_INFORM. We assign tag IT_IP_INFORM and delete the siblings.

This step results in another 66 'done' tweets. To account for the changes in the voting pattern after the third step, we apply the second (majority vote) merging step once again. After each merge the segments are recalculated. As a result we have 816 'done' tweets and 397 tweets that still need to be reviewed because disagreements on at least one segment could not be resolved automatically. This happened particularly for tweets with only two annotators, where majority voting did not help to resolve problems. Two students among the annotators adjudicated the remaining problem tweets manually. Further analysis in this paper is based on this merged 'gold standard' dialog act annotation for German conversations, in part in comparison with the original raw annotations.

5.2 DA n-grams

First, we examine DA unigrams to see which kind of acts/functions are common in our data. Both the original and merged data lack the same two tags: PCM and INTRODUCE_RETURN. In the merged data the root tag of the annotation schema, DIT++ TAXONOMY appears additionally. This is the result of a merging error, unifying two top level dimension tags. These mistakes will be manually corrected in the future.

Table 3 shows the top 5 and bottom 5 tags that are used in the original and merged data. As we can observe, the top 5 tags stay the same after merging but some rare tags appear by merging (IS, the main question label), and some of the

Original annotation	Merged annotation
0	0
INFORM	INFORM
ANSWER	ANSWER
AGREEMENT	AGREEMENT
SETQUESTION	SETQUESTION
...	...
APOLOGIZE	OCM
BYE_RETURN	BYE_RETURN
INTRODUCE	INTRODUCE
OCM	IS
DSM	INTRODUCE_ INITIAL

Table 3: Unigrams in the original and merged data.

rarest tags in the raw data move higher up after the merging process. We have also extracted the unigram frequencies for long and short conversations (see above) separately, but the frequency of certain DAs is generally very similar in these different types of conversations. By far the most frequent DA (26% or 22%, respectively) is INFORM. This is in line with data from spoken human-human dialogs, where STATEMENTs are sometimes even more frequent, at 36% (Stolcke et al., 2000). However, about twice as many dialog acts (8.7%) are characterized as SOCIAL in the long conversations as in the short conversations (4.4%), showing that short conversations are more aligned with the task.

To get a first glimpse of the structure of Twitter conversations, we calculated DA label bigrams as well. Twitter dialogs differ from more conventional dialog types in their branching structure: one turn can have several replies, each of which can be the basis of additional answers (see Figure 2b). In Twitter, in contrast to spoken conversations, this does not necessarily indicate a split of the conversation (and participants) into two separate strands. Instead, speakers can monitor both parts of the conversation and potentially contribute. Still, since replies mostly refer to the linked previous tweet, we can observe DA bigrams either within one tweet or across a tweet and its reply. Thus the last tag from the previous tweet and the first tag of the reply tweet are registered as a bigram. To distinguish the conversation start, we add another additional tag <S> to mark the beginning of the conversation. We also skip 0-tags (marking primarily user names at the beginning of reply tweets). Tables 4 and 5 show the top 5 bigrams and the most common starts of conversations, respectively. Table 6 compares the frequent bigrams for short and long conversations.

Bigram	Occurrence
INFORM, INFORM	135
ANSWER, INFORM	66
SETQUESTION, ANSWER	64
INFORM, AGREEMENT	63
AGREEMENT, INFORM	59

Table 4: Top five bigrams in the merged data.

5.3 Structure within tweets

Our analysis shows that despite their brevity, most tweets exhibit some internal structure. In 1213 tweets, we annotated altogether 2936 dialog acts. Table 7 shows the distribution of segments in tweets. It demonstrates that even though tweets are generally short, many contain more than just one dialog act. Even disregarding 0-segments (user names), which cannot be seen as true dialog acts, almost 500 tweets (more than 1/3) carry out more than one dialog act.

A tweet consists of at most 140 symbols. Since German words are on average six letters long[6], one German tweet consists of up to 23 words. Thus, in a tweet with five or six segments, each segment should have four to five tokens. Below we show two examples that have more than five segments, together with their annotations. Whereas some segments are debatable (e.g. the split-off dash in (7)), these examples show that Twitter turns can be quite complex, combining social acts with statements, questions, and emotional comments.

[6]Values around 6 are reported for the large Duden corpus `http://www.duden.de/suchen/sprachwissen/Wortlänge`, as well as for the TIGER corpus

Bigram	Occurrence
<S>, OPEN	40
<S>, TOPICINRODUCTION	32
<S>, INFORM	23
<S>, DSM	20
<S>, SETQUESTION	9

Table 5: Most common starts of the conversation.

Long conversations	Short conversations
INFORM, INFORM	INFORM, INFORM
INFORM, AGREEMENT	<S >, OPEN
AGREEMENT, INFORM	SETQUESTION, ANSWER
ANSWER, INFORM	ANSWER, INFORM
SETQUESTION, ANSWER	<S >, TOPICINTRODUCTION

Table 6: Bigrams in merged long and short conversations.

Number of segments per tweet	Tweets
1 segment	89 times
2 segments	671 times
3 segments	320 times
4 segments	114 times
5 segments	17 times
6 segments	2 times

Table 7: Distribution of segments.

(6) | @Marsmaedschen | Hey Mella, | sage mal, kocht ihr auf einem Induktionsherd? | Wenn ja, von welcher Firma ist die Grillpfanne? | Sowas suche ich! | :-) |
| 0 | GREET | QUESTION | SETQUESTION | INFORM | 0 |

(7) |@TheBug0815 @Luegendetektor @McGeiz | Genau, wir brauchen gar keine Grundlast, ist nur ein kapitalistisches Konstrukt | - | Wind/PV reichen? | Lol |
| 0 | AGREEMENT| 0| PROPQUESTION| DISAGREEMENT|

6 Discussion

In this paper we presented our attempt to annotate Twitter conversations with a detailed dialog act schema. We achieved only moderate inter-annotator agreement of $\pi = 0.56$ between three annotators on the DA labelling task, in contrast with work in other domains that achieved good agreement ((Stolcke et al., 2000) report $\kappa = 0.8$ for DA labelling of spoken data using 42 categories). Partially, annotation accuracy can be improved by better annotator training, e.g. to distinguish the different question types (see Table 9).

On the other hand, our data shows that the DA schema exhibits some inherent problems when applied to Twitter dialogs. For example, even though opening a conversation is rarely the main function of a tweet, every dialog-initial tweet could be argued to fulfil both the conversation OPEN function as well as a TOPICINTRODUCTION function, in addition to its communicative function (QUESTION, INFORM, etc.). Annotators found it hard to decide which dimension is more important. In the future, annotation in multiple dimensions should probably be encouraged, just like it was done for spoken human-human dialogs (Core and Allen, 1997; Bunt et al., 2010).

Many annotation problems are due to the fuzzy nature of INFORM and its relatives. Some INFORMs are shown in translation in (8–11). Even though all have been annotated with the same DA, they constitute very different dialog functions. Some are factual statements (8), some meta-commentary or discourse management (9), some opinions (10) and some read like statements or opinions, but are extremely sarcastic/ironic and thus do not have a primary "Information Providing" function (11). In order to properly analyse Twitter discussions, it seems necessary to make a clearer distinction between these kinds of dialog moves.

(8) *Coal 300 kWh, nuclear power 100 kWh*

(9) *The link still doesn't work.*

(10) *I'm going to end it right away, it got boring anyway.*

(11) *And the solar panels and wind power plants in the Middle Ages were great*

One implication of our DA annotation was that assigning single DAs to entire tweets is not sufficient. Not only does one utterance in Twitter dialogs often express several dialog functions as argued above, our data also shows that many tweets are composed of several successive dialog acts. This can be due to two discussion strands being carried out in parallel (like in text messaging), but often results from a combination of dialog moves as in this example:

(12) *True, unfortunately. | But what about the realization of high solar activity in the 70s and 80s?*

Finally, the non-linear structure of Twitter dialogs has interesting implications for their structural analysis, e.g. for DA recognition approaches

that take the context into account. In these cases, the initial tweet/DA will potentially be the first token of many DA bigrams. All answers taken together may provide context that helps determine what function the initial tweet was intended to fulfill. We leave these issues for further work.

Acknowledgements

The authors would like to thank the WebAnno development team for providing the annotation tool. We are extremely grateful to the participants in the Fall 2014 course *Dialogs on Twitter* at the University of Potsdam for their annotation effort. We are grateful to the anonymous reviewers for their detailed and helpful comments.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

John Langshaw Austin. 1975. *How to do things with words*, volume 367. Oxford university press.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Mark Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, pages 28–35. Boston, MA.

Eric N. Forsyth and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. pages 19–26.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. pages 152–161.

Jeroen Geertzen, Volha Petukhova, and Harry Bunt. 2008. Evaluating Dialogue Act Tagging with Naive and Expert Annotators. In *Proceedings of LREC*, pages 1076–1082.

Courtenay Honey and Susan C Herring. 2009. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE.

Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. 1995. Dialogue acts in Verbmobil. Technical report, Saarländische Universitäts- und Landesbibliothek.

Marco Lui 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*, (July):25–30.

Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Proceedings of NAACL*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, number 2010, pages 859–866.

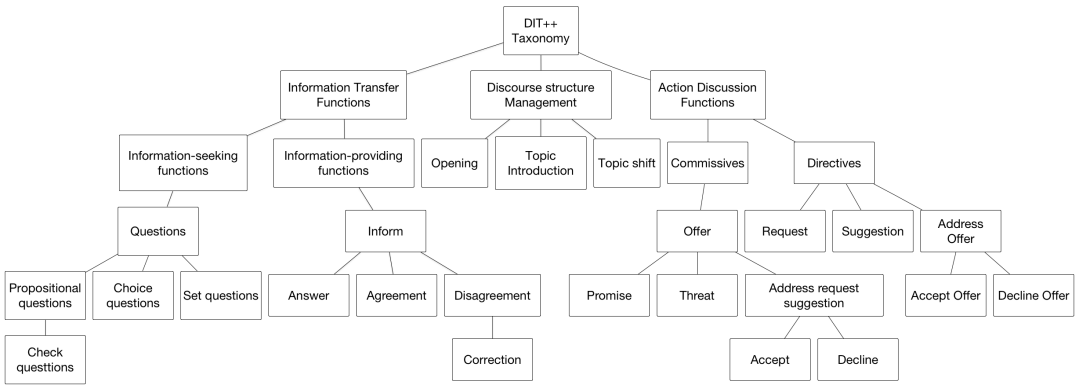
Tatjana Scheffler. 2014. A German Twitter snapshot. In N. Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andreas Stolcke, 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):339–373.

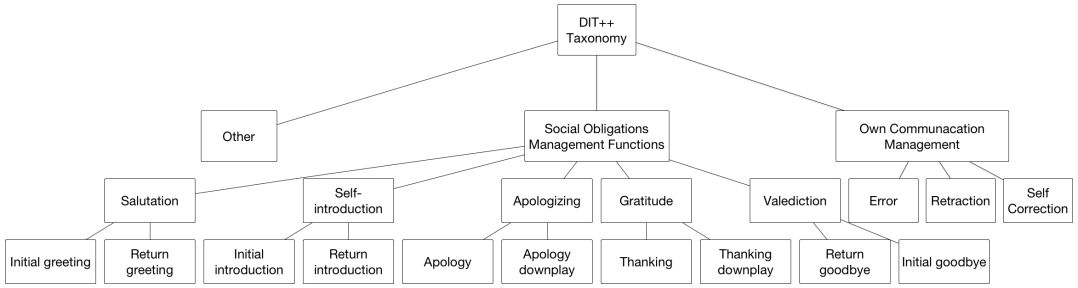
David R Traum. 2000. 20 questions on dialogue act taxonomies. *Journal of semantics*, 17(1):7–30.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.



(a) Adapted DIT++ taxonomy (1).



(b) Adapted DIT++ taxonomy (2).

Figure 3: Adapted DIT++ taxonomy.

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	DSM_OPEN	5	1	0	5	10	0	1	0	0
1	DSM_TOPICINTRODUCTION		0	0	0	9	1	1	0	0
2	DSM_TOPICSHIFT			0	3	17	3	8	1	5
3	IT				0	0	0	0	0	0
4	IT_IP					31	5	17	6	2
5	IT_IP_INFORM						26	45	31	15
6	IT_IP_INF_AGREEMENT							24	8	5
7	IT_IP_INF_ANSWER								14	8
8	IT_IP_INF_DISAGREEMENT									13
9	IT_IP_INF_DIS_CORRECTION									

Table 8: Annotation confusion matrix (1): Number of segments judged as both indicated dialog act labels by different annotators.

	PROPQUESTION_CHECKQ	SETQUESTION
PROPQUESTION	6	25
PROPQUESTION_CHECKQ		6

	PCM_COMPLETION	SOCIAL
INFORM	13	10
INFORM_AGREEMENT	2	15

Table 9: Annotation confusion matrix (2): Segments often confused within questions (top) or in other parts of the taxonomy (bottom).