Mapping German Tweets to Geographic Regions

Tatjana Scheffler Johannes Gontrum Matthias Wegel Steve Wendler
Department of Linguistics
University of Potsdam
firstname.lastname@uni-potsdam.de

Abstract

We present a first attempt at classifying German tweets by region using only the text of the tweets. German Twitter users are largely unwilling to share geolocation data. Here, we introduce a two-step process. First, we identify regionally salient tweets by comparing them to an "average" German tweet based on lexical features. Then, regionally salient tweets are assigned to one of 7 dialectal regions. We achieve an accuracy (on regional tweets) of up to 50% on a balanced corpus, much improved from the baseline. Finally, we show several directions in which this work can be extended and improved.

1 Introduction

Tweet collections are becoming more and more valuable as language resources due to their abundance, and the range of styles and topics they cover. Another interesting factor of Twitter data is the fact that it is much more than just text – metadata such as time stamps, user profile information and network data can be explored in NLP applications as well. Geolocation information is also sometimes present, most notably in the form of GPS coordinates of the origin of the tweet. However, while for some languages, geolocation data is commonly included in tweets, German twitterers are very reluctant to include geolocation coordinates. Of German tweets, which only make up less than 1% of all Twitter traffic, less than 2% are geo-tagged (Scheffler, 2014). In this paper, we show a data driven approach that can learn regionally salient words from seed data, and subsequently classify incoming tweets into geographic regions. Our method could be applied to other languages as well.

The aim of this study is to place German tweets geographically within a region of origin, despite the frequent lack of geolocation information. Tweets that do contain geolocation metadata (see Figure 1) are used as "gold standard" data in our work. The geolocation metadata of tweets is usually obtained from the GPS coordinates of the Twitter user (the author of the tweet) at the time of writing.

1.1 Regional expressions in tweets

Tweets that do not contain explicit geolocation metadata can still indicate where they originate from. In this first approach, we consider only the text of a tweet in order to place it geographically, and we ignore other information (for example, the authoring user and the user's given profile information). The text of a tweet can be regionally influenced in at least two ways: First, by the dialectal region of origin of the author (Twitter user). Such dialect regions could be reflected in the text by the use of regionally salient words and dialectal expressions (example (1a)). In German tweets, dialects are also often represented orthographically (e.g., by writing *ned* instead of *nicht*, 'not' example (1b)). Second, the current location of the twitterer induces the mention of location names, locally relevant person names, local events, etc

```
place (
| country =  "Germany"
| place_type =  "city"
| country_code =  "DE"
| name =  "Stuttgart"
| full_name =  "Stuttgart, Stuttgart"
| url =  "http://api.twitter.com/1/
    geo/id/e385d4d639c6a423.json"
| id =  "e385d4d639c6a423"
| bounding_box (
| | coordinates => Array (1) (
| | | ['0'] => Array (4) (
| | | | ['0'] => Array (2) (
| | | | | ['0'] =  9.038755
| | | | | ['1'] =  48.692343 )
| | | | ['1'] => Array (2) (
| | | | | ['0'] =  9.315466
| | | | | ['1'] =  48.692343 )
| | | | ['2'] => Array (2) (
| | | | | ['0'] =  9.315466
| | | | | ['1'] =  48.866225 )
| | | | ['3'] => Array (2) (
| | | | | ['0'] =  9.038755
| | | | | ['1'] =  48.866225 ) ) )
| | type =  "Polygon" )
| attributes ( )
)
```

Figure 1: Geolocation metadata of a tweet (JSON).

(example (2)). Both kinds of regional influences on tweet texts can of course pertain at the same time and possibly independently of each other, as when a person from Bavaria (region of origin) visits Berlin (current location). In this case, a mix of Bavarian terms and Berlin-specific names may occur.

(1) a. *Jep, der Lütte ist inzwischen 4,5 Jahre alt. . . .*
Yup, the little-one [regional Northern term] is now 4.5 years old. . . .
b. *Weiß ned, was ich lustiger finde. . .*
Don't know what's funnier to me. . .

(2) *Falls ihr jemanden mit einer Zwergenmütze durch Berlin laufen seht- winkt mir doch!*
If you see anyone walking through Berlin with a gnome hat, wave at me!

Although both kinds of regional influences are partially independent of each other, in this first attempt we have not tried to tease them apart systematically. Instead, we take geo-tagged tweets as accurately reflecting their origin and try to recover this geographic information in untagged tweets. Our basic assumption is that regionally diverging tweets (where regional origin and current location don't match) should be relatively rare compared to converging tweets, so that the basic signal does not get obscured for machine learning. In addition, our probabilistic model of regional salience (introduced below) allows for tweets and lexical items to be associated with several regions at the same time. With enough training data (and ignoring sparse data problems for the moment), this would allow for a tweet to be identified as associated with Bavaria and Berlin in equal measure.

1.2 German dialect regions

In this work, we defined dialect regions by hand based on existing classifications. For this purpose, we split the German-speaking European area into seven non-overlapping regions, along dialectal and structural boundaries (see Figure 2). We determined the regions based on the data in the *Atlas zur deutschen Alltagssprache* (de Liege and Salzburg, 2013). We also had to take some Twitter-specific properties into account. For example, the data of the *Atlas* also showed a small region around Saarland and Luxemburg to have characteristic idiosyncrasies, but we did not split it off because there would be too few tweets from such a small region.

1.3 Outline of this paper

In the following section, we give a brief overview of previous work with regard to processing German Twitter data and geolocation data encoded in tweets. Section 3 presents the data used in this work. Subsequently, we discuss our approach to finding the geographical origin of tweets and present our results. In the final section, we discuss the approach used and present several possible directions for further research.

2 Related Work

2.1 German Twitter

There is very little previous work on German Twitter data. Social media NLP research has largely concentrated on English, because English data are much more abundant (about 40–50% of

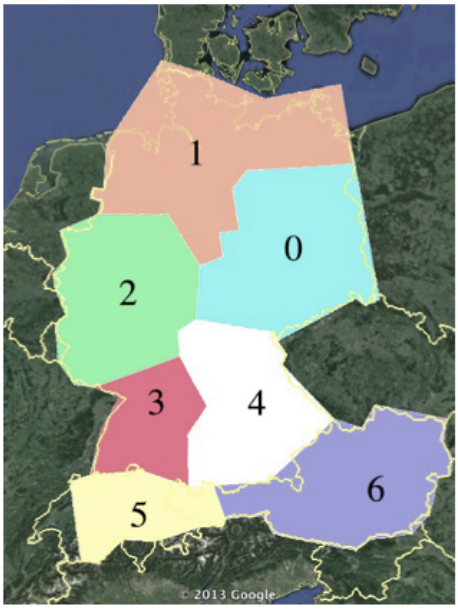

Figure 2: Map of the regions and the index of their feature used in the vectors represented as polygons.

all tweets) and thus easier to obtain. (Scheffler, 2014) introduces a large-scale corpus of German tweets, part of which is used in this work. Scheffler shows that in her corpus, which is an almost complete collection of all German-language tweets sent in April, 2013, less than 2% of these tweets contain public geolocation metadata.

There has been some work on adapting common NLP applications to German Twitter data, such as POS tagging (Rehbein et al., 2013b) and normalization (Sidarenka et al., 2013). And though certain linguistic phenomena have been studied using German Twitter data, including the specific style present on Twitter (Rehbein et al., 2013a), to our knowledge, no previous work has analysed the geographic origin or distribution of German tweets.

2.2 Tweets and geolocation

For other languages, the relationship between tweets and their location of origin has been looked at in several different ways. For example, (Arakawa et al., 2011) propose a three-tier search algorithm to find location dependent words. Their goal is to find place names and other terms (e.g., store names) to aid a predictive Japanese text-entry system. (Eisenstein et al., 2012) present a sociolinguistic study and model that shows how neologisms spread between US cities based on tweets. They used only data which included public geo-tags, while (Arakawa et al., 2011) devised a method to find geographically anchored Twitter data, even when those geo-tags are set to "private" by the users (they still show up in geographic Twitter searches). Recent work by (Grieve, 2014) on the regional distribution of variants in English also makes use of tweets with geolocation metadata.

Previous work on localizing tweets has for example built on language models (Kinsella et al., 2011), and has often tried to classify the location of users instead of a single tweet (Cheng et al., 2010; Hecht et al., 2011). In a different approach, (Leetaru et al., 2013) applied an algorithm developed for geocoding Wikipedia articles (Leetaru, 2012) to tweets. Since this approach is based on finding explicit location names in the text, it cannot be used to find the geographic origin of the vast majority of tweets.

3 Data

Our study is based on a corpus of German tweets collected in April 2013. It was collected by filtering the Twitter stream using a list of 397 common German words as key words (any tweet containing any word on the list is returned). The filtered stream was further narrowed down using the language identification module LangId (Lui and Baldwin, 2012), which yields very good results for our German data. The remaining data covers upwards of 90% of all German-language tweets sent during that period. We collected on average about 800,000 German tweets per day, for a total of 24,179,872 (see (Scheffler, 2014) for more detail on the corpus and the collection method). Out of these, only 254,874 tweets contained geolocation attributes. We eliminated tweets authored by two spam bots, all retweets, as well as automatically created tweets with the hashtags "#nowplaying", "#np", and "#4sq". After holding out 150 tweets from each region as a test set, the remaining 174,011 tweets formed our training corpus (geo-174k).

Since the regions were not represented equally in the training data (the smallest region, Austria, had only 8637 tweets, excluding the test set),

we built several balanced sub-corpora to measure the influence of the size of the training corpus: balanced-60k (the maximal balanced corpus with 60,459 tweets), balanced-21k with 3000 tweets from each region, and balanced-39k, all 39,459 tweets in the former sub-corpus but not the latter.

We performed almost no pre-processing on the data beyond the filtering described above. The tweets were tokenized using Christopher Potts' Twitter tokenizer[1], which recognizes such social media-specific entities as URLs, emoticons, etc. The resulting tokens were converted to lower case, yielding the final list of tokens for each tweet.

4 Geo-Mapping German Tweets

Our basic method is to represent each word in a corpus of tweets as a region vector representing the probability of that word originating from that region. Following the two kinds of regional influences on language mentioned above, we devised two approaches to train the initial region vectors from our training data: an approach based on dialectal expressions found in the Atlas zur deutschen Alltagssprache, and one trained directly from tweets that are tagged with geolocation information.

4.1 Regional words approach

The first attempt uses a seed word list of hand-selected regional expressions. As a source for the regional expressions we used the Atlas zur deutschen Alltagssprache (de Liege and Salzburg, 2013), which contains maps aggregating survey data on dialectal variants.

We included terms from the Atlas based on the following factors. Variants not reflected in the written form (such as vowel qualities) were excluded, as were multi-word expressions (e.g., *viertel vor*, a variant for the temporal expression 'quarter to'). We also excluded terms that showed too much overlap (did not adhere to clear dialect boundaries) or covered almost the entire language area (e.g., *Backofen*, 'oven'). A word was only included in our seed list of regional terms if it appeared in a maximum of four out of our seven regions. Furthermore, homonyms and polysemes were inappropriate for our purposes, so for example most of the regional words for 'attic', including *Boden*, *Speicher* and *Bühne*, were ruled out. We also went without very short expressions like *wa* (Berlin dialect for the question tag 'right?') because of the high chance of coincidence with abbreviations and cropped words.

In total, we selected a list of 209 regionally dependent terms from the Atlas, and split the probability mass uniformly between the regions in which the term is attested in order to yield seed vectors. E.g., the region vector for *Porree* ('leek') is (.33 .33 .33 0 0 0 0), since this word is only used in East, North, and West Germany (in the South, the variant *Lauch* is used). The disadvantage of this approach is the sparseness of the data, especially with regard to the kinds of terms not found in the Atlas (which contains mostly food related and outdated terms).

4.2 Training from geolocated tweets

In the second approach, we trained the seed vectors directly from tweets that have been tagged with GPS geolocation metadata by their authors. We used the following algorithm (Algorithm 1) to assess the probabilities of a certain term originating from a certain region by directly observing geo-tagged training data. For each tweet, we determined its originating region using the point-in-polygon algorithm from (Lawhead, 2011) and initialize the tweet vector as 1 for the originating region, and 0 for all others. For each term in the tweet, excluding stop words, we then added this tweet vector to the word vector for the term. After all tweets in the training corpus have been processed, these word vectors (essentially, counts of how often a word originated from each region) were then normalized to yield probabilities.

Following the initialization of the word vectors by one of the above methods, we included a bootstrapping step during which the vectors could be adjusted using additional data without geolocation information. In a nutshell, first a tweet vector is calculated for each tweet in the bootstrapping corpus based on the existing generation's word vectors (classification), and then a new generation of word vectors is calculated for the corpus based on the tweet vectors for all the tweets that a particular word occurs in (bootstrapping step).

[1] http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

Data:
tweets: Corpus of geo-annotated documents
stopwords: List of stopwords
Result:
WV: normalized word vectors, representing the probability distribution for each word

$WV \leftarrow \emptyset$;
foreach *tweet in tweets* **do**
  $region \leftarrow \texttt{Classify}(tweet)$;
  $\vec{tweet} \leftarrow \texttt{CreateVector}(region)$;
  forall the *token in tweet* **do**
    if *token* $\notin$ *stopwords* **then**
      $WV(token) \leftarrow WV(token) + \vec{tweet}$;
    end
  end
end
foreach $\vec{word}$ *in WV* **do**
  $\vec{word} \leftarrow \texttt{normalize}(\vec{word})$;
end
return *WV;*

Algorithm 1: Obtaining regional probabilities for words.

Finally, after training and bootstrapping, the word vectors can be used to classify tweets into regions. For classification, we used the cosine similarity between the tweet vector and the average tweet vector over the entire bootstrapping corpus. A tweet would be assigned to the dimension (region) in which the difference vector between the current tweet vector and the average tweet vector is maximal. Note however, that a huge majority of German tweets are written in standard German without any signs of regional influence whatsoever, or are very short. In order to alleviate this problem, we used a variable threshold of "non-regional tweets", below which we did not attempt to classify a tweet. This threshold (called "guess" in Algorithm 2) was set experimentally as the minimum difference (maximum cosine similarity) between a tweet vector and the average tweet vector, reasoning that a tweet that is very similar to the average of all tweets doesn't show any clear regional trends. Algorithm 2 computes the "average tweet" vector to compare each tweet with during classification, as well as the cosine similarity threshold beyond which a tweet is recognized as sufficiently "different" from the average. This threshold is computed based on a preset percentage of assumed regional tweets. We

Data: *tweets*: Set of geo-annotated documents
guess: guessed percentage of regional tweets
WV: Set of word vectors
Result: *threshold*: cosine similarity threshold

$tweetvectors \leftarrow \emptyset$;
foreach *tweet in tweets* **do**
  $\vec{tweet} \leftarrow (0,0,0,0,0,0,0,0)$;
  forall the *token in tweet* **do**
    if $token \in WV$ **then**
      $\vec{tweet} \leftarrow \vec{tweet} + WV(token)$;
    end
    $tweetvectors \leftarrow tweetvectors \cup \{\vec{tweet}\}$;
  end
end
$\vec{average} \leftarrow (0,0,0,0,0,0,0,0)$;
foreach $\vec{tweet} in tweetvectors$ **do**
  $\vec{average} \leftarrow \vec{average} + \vec{tweet}$;
end
$\vec{average} \leftarrow \frac{\vec{average}}{l(tweetvectors)}$;
$vectorlist \leftarrow \emptyset$;
foreach $\vec{tweet}$ *in* $tweetvectors$ **do**
  $similarity \leftarrow sim(\vec{tweet}, \vec{average})$;
  $vectorlist \leftarrow append(similarity)$;
end
$vectorlist.sort()$;
$threshold = vectorlist[int(guess \times l(vectorlist))]$;
return *threshold*;

Algorithm 2: Cosine similarity algorithm.

discuss below how this threshold is set.

The final parameter influencing the results is the length of the stop word list. We compiled a custom stop word list by excluding the most frequent N words in the training corpus. The best value for N was determined experimentally.

5 Results

Here, we first report the results of the approach using geo-tagged data for estimating the initial word vectors. A naïve random baseline for tweet classification on the balanced test set should yield an accuracy of 1/7 = 0.14 for seven regions.

First, we evaluated the best data set combinations for the training and bootstrapping stage; all numbers are accuracy scores on the held-out test set of 1050 tweets (150 from each region). For subsequent experiments, we used the best data sets determined above: For training, the balanced-

39k corpus, and for any bootstrapping steps, the entire (unbalanced) geo-tagged corpus of 174k tweets.

Next, we assessed the effect of the number of stop words excluded. Figure 3 shows that performance decreases again after 200 words, maybe because some regional words are very common.

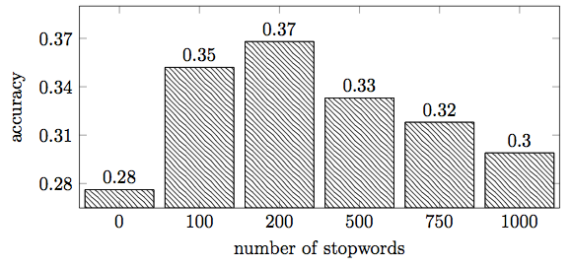

Figure 3: Accuracy based on size of stop word lists.

To determine the optimal cosine similarity threshold ("guess") to distinguish "standard German" from regional tweets, we varied the number of regional tweets we attempted to classify in steps of 10%. Clearly, the accuracy rises the fewer tweets are deemed "regionally salient". The optimal result on the test set is reached with only 20% of tweets deemed sufficiently different from the average to be classified. The overall accuracy on this setting reaches 0.506 (see Figure 4).

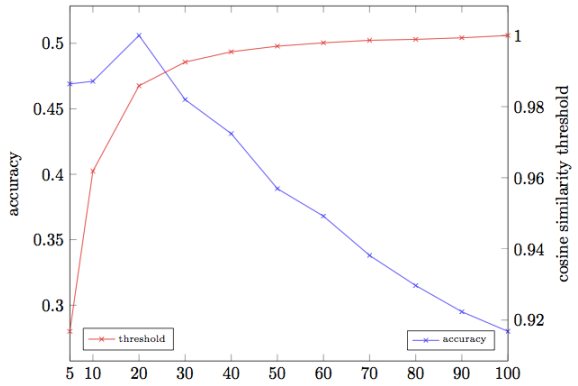

Figure 4: Relation between percentage of regional tweets and accuracy.

Finally, we estimated the effect of the number of bootstrapping loops included in the calculation. Any number of bootstrapping steps actually decreases the overall accuracy. We suspect that this happens because during bootstrapping, all vectors are assimilated more and more to the average vector.

The best result of our classification algorithm is obtained with the balanced-39k training corpus and the geo-174k corpus used in order to compute the overall average tweet vector (the bootstrapping step is skipped), with 200 stop words excluded and 20% of tweets deemed regionally salient (this corresponds to a maximum cosine similarity value of 0.94). With these settings, we achieve an accuracy of 0.53 on the test set.

Using the regional words approach, the results were much worse, reaching only up to an accuracy of 0.3 in the best case. We kept the percentage of regional tweets (20%) and the stop word list (200 words) constant.

6 Discussion

In this paper, we have shown a data-driven method to regional classification of German tweets. Our approach is trained on a medium-sized corpus of geographically tagged German tweets by deriving regional probabilities for each word in the corpus. Though most tweets are standard German and cannot be assigned to one particular region, we automatically identify the 20% most significantly regionally influenced tweets. Our classification accuracy on these 20% is 0.53 with optimal settings, a significant improvement over the 0.14 random baseline.

Our second approach based on a seed set of regionally salient words yields a much lower accuracy of less than 0.3 due to sparse data problems. An obvious idea for future work is the combination of the two methods, since they capture different intuitions: the geolocation metadata used in the geolocated tweets approach is based on the current location of the twitterer (usually, GPS location obtained from a mobile phone). In contrast, the regional and dialectal expressions covered in the Atlas zur deutschen Alltagssprache more likely reflect the regional origin of the twitterer (no matter her/his current location). It could also be worthwhile to amend the regional word seed list, which is currently very small (only 209 terms). Then, it could be combined with additional geo-tagged Twitter data in a bootstrapping step as outlined above.

In addition, the current scoring scheme is very rigid and does not reflect the fact that some re-

	Assigned region						
True region	*0*	*1*	*2*	*3*	*4*	*5*	*6*
0 = East	.18	.41	.12	.06	.00	.06	.18
1 = North	.12	.65	.00	.12	.00	.06	.06
2 = West	.09	.23	.45	.14	.05	.05	.00
3 = Southwest	.04	.22	.13	.52	.09	.00	.00
4 = Bavaria	.05	.29	.05	.00	.57	.00	.05
5 = Switzerland	.02	.16	.04	.08	.00	.68	.02
6 = Austria	.05	.27	.00	.05	.05	.18	.41

Table 1: Confusion matrix for final run.

gions are more similar to each other than others, as is also visible from the confusion matrix in Table 1. The table also indicates that most misclassifications are assigned wrongly to region 1 (North), indicating a problem with the definition of that region or with the corpus training data we have for it.

Another obvious extension to the work reported here, as suggested by one of the reviewers, is a qualitative evaluation of regional and non-regional German tweets with respect to linguistic and lexical features. This may lead to an improved regional seed word list, possibly a new region assignment, and new insights for the localization of tweets.

Acknowledgments

We are very grateful to the four reviewers' detailed comments and questions. This work has been supported by the collaborative project "Analysis of Discourse in Social Media" (project number 01UG1232A), funded by the German Federal Ministry of Education and Research.

References

Yutaka Arakawa, Shigeaki Tagashira, and Akira Fukuda. 2011. Spatial statistics with three-tier breadth first search for analyzing social geocontents. In A. König, A. Dengel, K. Hinkelmann, K. Kise, and R. J. Howlett, editors, *Proceedings of the 15th international conference on Knowledge-based and intelligent information and engineering systems (KES'11)*, volume Part IV, pages 252–260. Springer.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

Universite de Liege and Universität Salzburg. 2013. Atlas zur deutschen Alltagssprache. online resource. http://www.atlas-alltagssprache.de/.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. *CoRR*, abs/1210.5268.

Jack Grieve. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speec*. Walter de Gruyter, Berlin/New York.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA. ACM.

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 61–68, New York, NY, USA. ACM.

Joel Lawhead. 2011. Point in polygon 2: Walking the line. http: //geospatialpython.com/2011/08/point-in-polygon-2-on-line.html.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

Kalev Leetaru. 2012. Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia. *D-Lib Magazine*, 18(9):5.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In

ACL (System Demonstrations), pages 25–30. The Association for Computer Linguistics.

Ines Rehbein, Sören Schalowski, Nadja Reinhold, and Emiel Visser. 2013a. Uhm... uh.. filled pauses in computer-mediated communication. Talk presented at the Workshop on "Modelling Non-Standardized Writing" at the 35th Annual Conference of the German Linguistic Society (DGfS).

Ines Rehbein, Emiel Visser, and Nadine Lestmann. 2013b. Discussing best practices for the annotation of Twitter microtext. In *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, Sofia, Bulgaria.

Tatjana Scheffler. 2014. A German Twitter snapshot. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Uladzimir Sidarenka, Tatjana Scheffler, and Manfred Stede. 2013. Rule-based normalization of German Twitter messages. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*.