A German Twitter Snapshot

Tatjana Scheffler

Department of Linguistics
University of Potsdam
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany
tatjana.scheffler@uni-potsdam.de

Abstract

We present a new corpus of German tweets. Due to the relatively small number of German messages on Twitter, it is possible to collect a virtually complete snapshot of German twitter messages over a period of time. In this paper, we present our collection method which produced a 24 million tweet corpus, representing a large majority of all German tweets sent in April, 2013. Further, we analyze this representative data set and characterize the German twitterverse. While German Twitter data is similar to other Twitter data in terms of its temporal distribution, German Twitter users are much more reluctant to share geolocation information with their tweets. Finally, the corpus collection method allows for a study of *d*iscourse phenomena in the Twitter data, structured into discussion threads.

Keywords: Twitter, corpus, German

1. Introduction

Twitter corpora have become a valuable source of data for linguistic and natural language processing (NLP) studies, due to the abundance of up-to-date, varied data. However, most existing research deals only with English tweets. This has several reasons. First, English dominates in the mix of languages on Twitter. According to different studies, more than 50% of tweets are written in English. Outside of the largest five Twitter languages (see Figure 1), other languages represent just under 1% of Twitter traffic each.

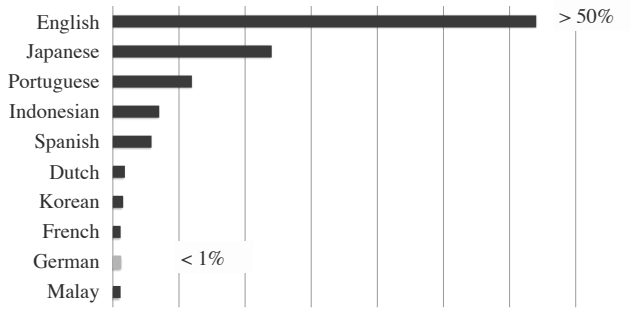

Figure 1: Top ten languages on Twitter. Data from (Hong et al., 2011).

English data is thus much easier to obtain and much more abundant than data in other languages. This fact plays into the second reason why research on English tweets predominates: The Twitter API (Twitter, 2013) offers several access methods to its data. The most commonly used access point is a random subset of tweets through the *gardenhose* stream (1% to 10% of tweets). Finding tweets of a particular non-English language in this stream of data is not trivial, and smaller languages may not be included in sufficiently large quantities over a short timespan. In addition, since the sampling method Twitter uses to reduce the stream is not entirely known, it is unclear whether corpora produced in this manner are inherently biased in some way.

Twitter is a communication channel (Dürscheid, 2003) that contains text of various different genres, registers and types. It includes both curated content (headlines, company dispatches, spam, bot-generated tweets), as well as spontaneous user-generated content (statements, discussions, small-talk, etc.). The exact make-up and relevance of Twitter content is still a target of study by communication scientists (among others), although for many NLP applications the data should be filtered to include more user-generated text and fewer automatic or spam messages.

The motivation for this work was the collection of a *representative* sample of German Twitter messages, that reflects the entirety of German Twitter content. Due to the relatively small number of German tweets, it is possible to collect a virtually complete snapshot of German twitter messages over a certain time span. In this paper, we present our collection method for our over 24 million tweet corpus. We prove that this corpus includes a large majority of all German-language tweets sent in April, 2013. Further, we give some initial analyses of this representative data set and present characteristics of the German twitterverse.

2. Related Work

Several attempts have been made to create Twitter corpora for reuse among NLP researchers. Of course, common corpora are essential for comparability of results and to reduce reduplication of effort. Work in this area is seriously restricted by the Twitter terms of service, which do not allow the sharing of aggregated resources of tweets. Several previously available Twitter corpora (for example, the Edinburgh Twitter corpus (Petrović et al., 2010)) have been retracted for this reason. A possible workaround for the NLP community is the distribution of only lists of tweet IDs, as is done for example in the TREC microblog shared task[1]. Another option is the distribution of only derivative data, such as n-gram counts, instead of the actual tweets themselves (Herdağdelen, 2013). However, this second approach makes certain kinds of linguistic analyses of the data

[1] http://trec.nist.gov/data/tweets/

impossible. Some analyses in the paper by (Herdağdelen, 2013), where this approach is proposed, such as the overall corpus analysis in section 4 (tweets per day of the week, etc.) are not possible with the aggregated n-gram corpus. Since we are particularly interested in tweets in context, we stick with the construction of an actual corpus of tweets including their metadata, which can then only be shared via the tweet IDs.

Previous Twitter corpora such as the Edinburgh corpus (Petrović et al., 2010), the Tweets2011 corpus from the TREC microblog shared task, as well as the Rovereto n-gram corpus (Herdağdelen, 2013), were collected using the public "gardenhose" setting of the Twitter streaming API. By this method, a certain (small) fraction of all tweets can be collected over a period of time. However, the sampling method Twitter uses to determine the random subset of tweets delivered is not clear, so a certain bias is possible. In addition, these existing corpora of social media data are almost always in English, since English data are most abundant and easiest to retrieve. In the current work, we are interested in German tweets.

In the web corpus construction community, sites for a particular language are often found using mid-frequency words as search terms (through a particular search API) (Baroni and Bernardini, 2004; Schäfer and Bildhauer, 2012). Here, we follow a similar approach, but using very high-frequency terms as keywords instead. This way, virtually all German tweets can be retrieved using a small list of search terms.

Another approach to collecting German tweets was followed in recent work such as (Rehbein et al., 2013). Here, tweets were founding using geolocation features and then filtered for language. However, as we will show using our corpus below, such data is very strongly biased for German, since only a tiny minority of Twitter users allow the public submission of their geolocation data. It must be assumed that the tweets retrieved this way are not representative of the larger sample, since the tweets of a user who has switched the geolocation feature to "on" will almost always be included in the dataset, whereas other kinds of users will never have their messages included. Furthermore, the density of tweets collected in this way is low and individual tweets are collected out of context. Finally, certain types of tweets (that do not often originate from smartphone clients) may be systematically excluded, such as curated content or in-depth political discussions.

3. Corpus Creation

The goal for this work is the collection of a representative sample of German Twitter messages. Tweets are very interesting for linguistic studies because they are almost limitless: Even though German tweets are relatively rare, they still make up more than 10 million words per day. Linguistic data on Twitter is characterized by a large mix of registers and a very useful set of metadata for each tweet. In order to study the particularities of German tweets, we aimed to collect if possible *all* German Twitter messages over a period of one month. Since virtually all German tweets contain at least one very high-frequency stopword, and German tweets are rare enough that all of them can be collected without hitting the Twitter rate limit, we tracked German high-frequency terms in order to collect a complete snapshot of German Twitter.

3.1. Data Collection

We collected the corpus using the Python package Tweepy (Tweepy, 2013) to access the Twitter Streaming API (Twitter, 2013). Our corpus collection pipeline is shown in Figure 2. The API allows simultaneous tracking of up to 400 keywords. The targeted access points to the Twitter streaming API (such as keyword tracking) differ from the gardenhose access points in an important way: As long as the number of tweets that match the query don't exceed a certain rate limit (standardly this rate limit is given as about 1% of tweets (Twitter, 2013)), Twitter returns *all* matching tweets. If the rate limit is exceeded, the user is notified of the number of omitted tweets.

We modified a German stop word list to exclude words that are also very frequent in other languages, especially English ('war', 'die'), because this would dilute the stream and also make it more likely to hit the rate limit[2]. Then we added other frequent uniquely German terms, such as number words, to yield a stop word list of 397 words.

The majority of tweets collected with this keyword list is nevertheless not German. After testing several existing language identification modules we settled on LangID (Lui and Baldwin, 2012), which achieves very good precision and recall on our data. The remaining tweets which are tagged as German by LangID make up our corpus of German tweets.

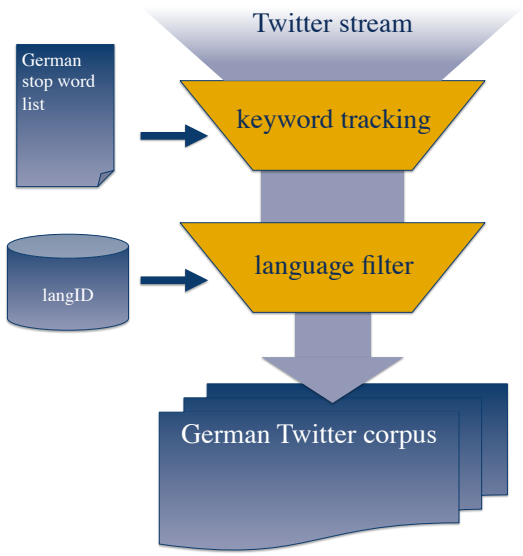

Figure 2: Corpus creation pipeline.

3.2. Completeness

Since we aimed to create a representative corpus of the entire German Twitter data, we evaluated how many German tweets were missed at each stage of the corpus creation: by

[2]Twitter has a rate limit in place for accessing the stream, omitting some messages if too many tweets match the given query. The number of omitted tweets is then transferred instead.

using the stop word list, by language filtering, and through rate limiting. In order to evaluate the coverage of the stop word list, we ran several collection methods in parallel: the stop word tracking method, a geolocation-based stream with a box encompassing approximately the area of Germany, and a user-based stream with a set of 500 user IDs which were semi-automatically determined to have posted recently and at least sometimes post in German. This pretest was carried out in December 2011, when rate limiting was not an issue yet for German data, since German tweets were so infrequent. Over the same period of four days, we collected almost 1.8 million tweets through the stop word list (*track* stream), 365,000 through the user list (*follow*), and less than 30.500 through the geolocation restriction (*loc*). The user stream and the geolocation stream contained so few messages that it is clear they returned *all* messages that fit the query: tweets that were written by users from our list or were sent from locations within our bounding box. respectively.

We then checked how many of the *follow* and *loc* tweets were also included in the *track* corpus, in order to assess the coverage of the stop word list. The coverage was 97.2% for *loc* and 94.6% for *follow*. This means that only around 5% of potential German tweets do not contain one of the keywords on our stop word list for accessing the Twitter stream. Since not all of those missed tweets are actually in German, the real number of target tweets that are missed is likely much lower than this upper bound.

For the language identification module, we carried out a small manual evaluation. It yielded a precision of 97.3% on the streaming data. Another package with very good results was the Google language detector which is part of Google Translate (McCandless, 2011). In addition, our data suggests that the two modules make complementary errors and are therefore even slightly better in combination, especially with regard to precision. This suggests further language filtering on our corpus as a way to clean up the data.

Finally, the stop word list matches on a very large number of tweets, which leads to restrictions based on Twitter's rate limiting. We estimate that over the course of the month of April, less than 4.5 million tweets were missed in the tracking stage due to rate limiting. However, only a small percentage (around 16%) of these are actually German. This means that up to about 700,000 German tweets may have been missed due to rate limits, or under 3% of the data.

Taken together, this means that our collection method enables us to collect more than 90% of all German-language tweets over a given time period, disregarding the recall of the language filter used. The other less than 10% of German tweets missing from the corpus were lost either due to the lack of coverage of the stop word list or due to rate limiting by Twitter. For future corpus construction, it may be useful to optimize the stop word list by tracking which keywords are good discriminators of German tweets (retrieve many German tweets but few foreign-language messages). This would simultaneously reduce the risk of rate limiting (and losing messages) as well as improve the recall of German tweets.

4. German Twitter Data

In total, we collected 24,179,189 tweets during the month of April, 2013, a little more than 800,000 per day. Figure 3 shows the temporal distribution of tweets during the time period, binned by hours. A better view of the distribution of tweets throughout the day is shown in Figure 4, where the average of all 30 days is depicted.

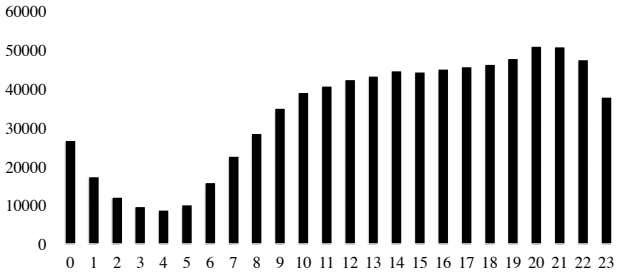

Figure 4: Number of tweets by hour, averaged over all 30 days.

It can be seen that German twitter users are most active during "office" and evening hours, although a significant number of tweets (almost 10,000 per hour) are also sent through the night. In this data, all messages are shown in the Central European time zone. Similarly to (Herdağdelen, 2013), we see the slowest Twitter traffic at 4 a.m. local time, and the peak around 8 or 9 p.m.
Some of the night-time activity may also be due to spam or automatic posts. We analyzed all Twitter users included in the corpus to find the distribution of frequent and infrequent twitterers. The distribution is markedly Zipfian, as shown in Figure 5. Out of the more than 1.9 million unique user ids in our corpus, more than 1.1 million users only wrote one tweet over the entire one-month period. In contrast, the most prolific twitterers send up to 1 tweet per minute (28,500 within the month). These hyper-active twitterers were usually bots dispersing spam or automatic sensor data.

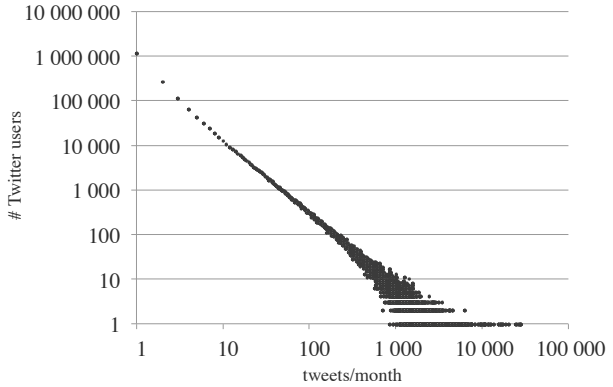

Figure 5: Log-log scale graphic of users vs. tweets per user.

We have also made some effort towards automatically identifying good quality tweets from spam and automatic posts. We have identified the originating client as a very good indicator of poor quality data: The vast majority of human-

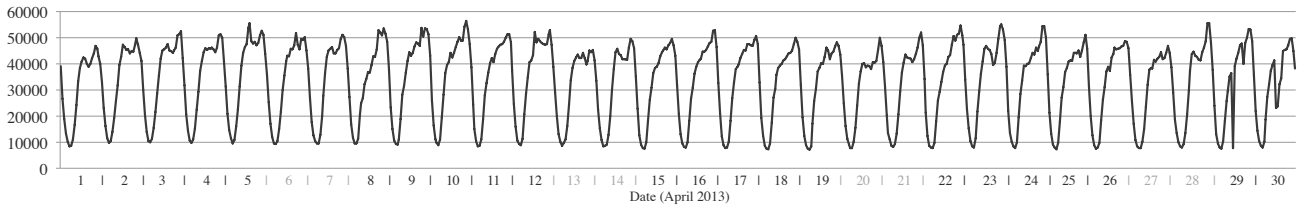

Figure 3: Number of tweets in the corpus, binned by hours. Weekends are marked in light grey. Server problems affect the last two days.

authored, genuine Twitter data originates in fewer than 20 common clients, including Twitter's own websites, mobile clients, and so on (see Table 1 for the top 10 German language Twitter clients). The other 12905 sources for Twitter data are mostly made up of customized bots or APIs which distribute spam and auto-posts. These types of clients account for up to one fifth of Twitter data which can be excluded in order to obtain higher-quality Twitter data, for applications where bot-generated content is problematic. We have not purged suspected spam from the corpus since we aim here for completeness in order to allow for different uses of the data. Spam detection may be one possible application which can be carried out with our data (although a gold standard for evaluation would have to be constructed).

# of tweets	client	% of tweets
5679380	web	23.5%
3311068	Twitter for Android	13.7%
2966427	Twitter for iPhone	12.3%
1955509	Twitterfeed	8.1%
1232017	The Tribez for Android	5.1%
1211910	TweetDeck	5.0%
1058326	Facebook	4.4%
807320	Tweetbot for iOS	3.3%
544675	Google	2.3%
491480	Tweet Button	2.0%
19258112	**total**	**79.6%**

Table 1: Ten most frequent Twitter clients in our data.

Geolocation features. One fact that distinguishes German Twitter data from other languages is the reluctance of German users to share their geographic location publicly. In the corpus, only 1.1% of all tweets contain geolocation information (see Table 2). In addition, many tweets with geo information are mere check-ins ("I'm at ...") or automatically posted tweets ("Now playing on XYZ radio ...") without any real linguistic content. In consequence, even in Berlin the existing geolocated tweets track the movements of a very small number of Twitter users, without giving a reliable indication of German Twitter users' whereabouts in general (Figure 6).

Twitter discourses. In addition to spam and celebrity news, Twitter also contains many discussions between humans. In fact, 21.2% (5,133,544) of the tweets in our corpus are *replies* to a previous tweet. The vast majority of these replies are human-authored linguistic content (which may be used in spam filtering). Our corpus, since

total tweets in the corpus	24179189	
geo-tagged tweets	263364	1.1%
distinct users in corpus	1907891	
distinct users in geo-tagged tweets	46559	2.4%

Table 2: Numbers of tweets in the corpus.

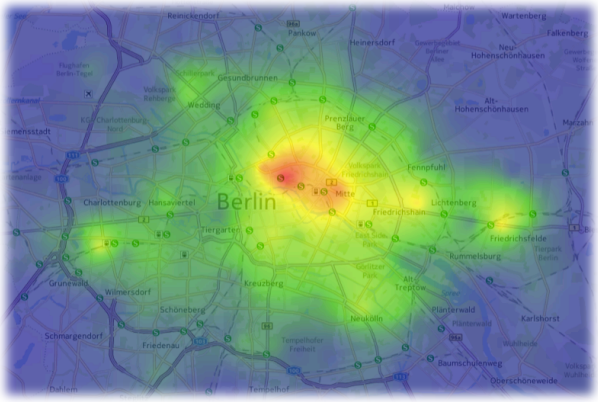

Figure 6: Heat map of tweets in Berlin. Two red points show probable homes of the two Berlin twitterers who are most happy to share geotagged tweets.

it contains largely complete German Twitter data, allows the further study of Twitter conversations. Previously, this has only been possible using customized corpora (Ritter et al., 2010), since randomly sampled Twitter corpora are not guaranteed to contain connected conversations. In our corpus, connected threads of Twitter discussions can be retrieved using the "in_reply_to_status_id"-links.

The vast majority of these discussions is only two tweets long (one initiating tweet and one reply), but they can be up to hundreds of tweets in length. Figure 7 shows a scatter plot of the length vs. depth (maximum level of embedding for a reply in the discussion) of the discussion threads on April 1, 2013. It can be seen that at the extremes, two types of discussions exist: First, in the lower right corner of the plot, posts that got many answers (presumably from different users) but whose answers didn't in turn yield further discussion. Celebrity statements ("I'm finally at home. Where are you right now?") are typical for this type. Second, in the diagonal are discussions whose depth and length is exactly the same, indicating that each new tweet in the discussion is a reply to the previous one. This structure is typical for

conversations between few users that go back-and-forth.

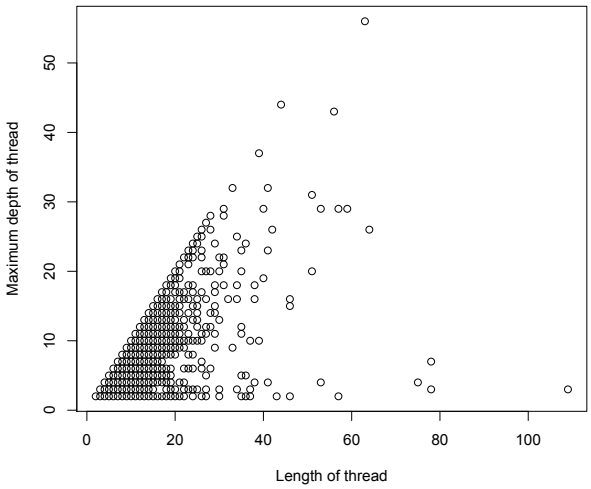

Figure 7: Length vs. maximum depth of discussion threads on April 1, 2013.

5. Example Corpus Application

To demonstrate a potential use of the corpus, we carried out a small linguistic study to look into the particular "Twitter style" of language. It is sometimes claimed that (some) Twitter messages exhibit a more oral style than other written text. See also (Rehbein and Ruppenhofer, 2013) for a look at some other features of orality. We looked into the distribution of different causal connectives in order to test this hypothesis. In German, there are three conjunctions roughly expressing 'because': *weil*, *denn*, and *da*, plus an adverbial *nämlich*. In addition, many phrasal expressions can be used to indicate causality or reasons, but they were excluded from the present study. Of these connectives, especially *denn* and *da* have been claimed to belong mostly to the written register, while being extremely rare in spoken German (Wegener, 1999).

In Figure 8, we show the relative frequencies of German causal connectives in different text types, including Twitter. Since different text genres use discourse connectives in different frequencies in general, we normed the observed frequencies of each connective to the frequency of *weil*, the connective with the broadest meaning and distribution.[3]

The comparison shows that while *denn* and *da* are very common in the two written corpora, *bmp* and *Rudolph*, *weil* is much more prevalent in both the spoken corpora and on Twitter, dwarfing out all other kinds of causal connectives. This shows that on this measure at least, German Twitter messages do indeed show a more "oral-like" style, probably due to their short and often dialogic structure. This observation can be confirmed when looking at only the replies in the Twitter data, which are always part of a discussion. All

[3]In the Twitter corpus and the spoken corpus *FOLK*, the causal uses of *denn* and *da* had to be estimated by manually examining a smaller number of items, since both are highly ambiguous and have unrelated meanings.

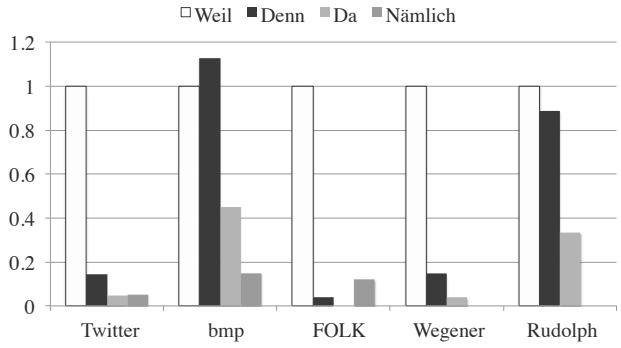

Figure 8: Relative frequencies of *denn*, *da*, and *nämlich*, compared to *weil* in different types of corpora, including Twitter. Twitter = 253172 German tweets about former president Wulff; bmp = Berliner Morgenpost-subsection from the COSMAS II newspaper corpus; FOLK = Forschungs- und Lehrkorpus Gesprochenes Deutsch – dialogs; Wegener = spoken corpora 1980-1999 from (Wegener, 1999, Tab. 1); Rudolph = written texts (Rudolph, 1982) cited in (Wegener, 1999)

studied connectives are more common in replies than non-replies. The adverbial *nämlich*, which can be used across turns, is more than twice as common in replies as it is in the general corpus (Table 3).

nämlich	14431	in 0.059% of tweets
nämlich in replies	6336	in 0.123% of tweets

Table 3: Prevalence of *nämlich* in the corpus.

6. Conclusion

In this paper we introduce a new, comprehensive corpus of German Twitter data. We present our corpus collection method which is based on a language-specific stop word list and aims to collect a representative chunk of all German language tweets. The coverage of this collection method is above 90% before language filtering. In addition, we characterized the obtained 24 million tweet corpus in part to show the specific make-up of German Twitter data. The corpus will be made available in a format complying with Twitter's Terms of Service (tweet ID list). It can serve as a basis for linguistic studies of German social media as well as a training corpus for NLP applications.

7. Acknowledgements

The author would like to thank the student researchers Kira Eberle and Norman Rosner, as well as Wladimir Sidorenko for their participation in this project. I am grateful to the three anonymous reviewers for their helpful comments. This work is part of the collaborative project Analysis of Discourse in Social Media (project number 01UG1232A), funded by the German Federal Ministry of Education and Research.

8. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal.

Christa Dürscheid. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für angewandte Linguistik*, 38:37–56.

Amaç Herdağdelen. 2013. Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, 47(4).

Lichan Hong, Gregorio Convertino, and Ed Chi. 2011. Language matters in Twitter: A large scale study. In *International AAAI Conference on Weblogs and Social Media*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea.

Michael McCandless. 2011. Accuracy and performance of Google's Compact Language Detector. blogpost. http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ines Rehbein and Josef Ruppenhofer. 2013. Investigating orality in speech, writing, and in between. Talk presented at the Corpus Linguistics 2013 conference. Lancaster, UK.

Ines Rehbein, Sören Schalowski, Nadja Reinhold, and Emiel Visser. 2013. 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). Potsdam, Germany.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 172–180.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 486–493. European Language Resources Association (ELRA).

Tweepy. 2013. Twitter for Python. https://github.com/tweepy/tweepy.

Twitter. 2013. Twitter Streaming API. https://dev.twitter.com/docs/streaming-apis.

Heide Wegener. 1999. Syntaxwandel und Degrammatikalisierung im heutigen Deutsch? Noch einmal zu weil-Verbzweit. *Deutsche Sprache*, 27(1):3–26.