N-gram Frequencies for Dutch Twitter Data

Gosse Bouma G.BOUMA@RUG.NL

University of Groningen, Groningen, The Netherlands

Abstract

This paper presents n-gram frequency data obtained from a large sample of Dutch tweets, covering a period of 4 years. After filtering of re-tweets, (near-) duplicates, and non-Dutch tweets, more than 2.6 billion tweets remained. These were tokenized, and frequencies were collected for n-grams of up to 5 words. A web interface allows users to obtain frequency information for spelling variants, grammatical phenomena (as reflected in n-gram patters), monthly trends, and word clusters. All the underlying n-gram frequency data as well as the word clusters are available for download.

1. Introduction

There can be no doubt that the language used on social media differs considerably from other linguistic registers (Baldwin et al. 2013, Eisenstein 2013). Given the widespread use of Twitter, Facebook, whatapp, and other applications in almost all parts of society, it is not surprising that linguists are interested in studying the exact nature of language use and change in such media. However, platforms like Twitter provide a continuously fluctuating stream of data, which cannot readily be used as corpus. Most researchers therefore work with samples, using various restrictions to ensure that only tweets from a given region, language, user group, or topic are collected. Such corpora can take considerable effort to compile, and thus are valuable resources to be shared, also with an eye on replicability of results. Copyright laws make sharing of this kind of data impossible, however, and in some cases researchers were forced to withdraw data-sets from the public domain (i.e. the Edinburgh Twitter Corpus described in Petrović et al. (2010)). An alternative that can be valuable for many researchers, is to give access to n-gram frequencies derived from large samples of Tweets. While obviously less informative than the original data, frequency information can be of use for many applications and research questions, and furthermore has the advantage that it can be made fully public and results based on these frequencies can be verified and replicated.

This paper presents n-gram frequency data obtained from a large sample of Dutch tweets, covering a period of four years. After filtering of re-tweets, (near-) duplicates, and non-Dutch tweets, more than 2.6 billion tweets remained. These were tokenized, and frequencies were collected for n-grams of up to five words. The web interface allows users to obtain frequency information for spelling variants, grammatical phenomena (as reflected in n-gram patters), monthly trends, and word clusters. All the underlying n-gram frequency data as well as the word clusters are available for download.

There are at least two reasons why this data is of interest. First of all, search in a corpus consisting of 2.6 billion tweets (with an average token length of 10), can be challenging. A researcher interested in the frequency with which *jij/hij/zij wordt* and *jij/hij/zij word* (and similarly for other frequent verbs with a stem ending in *-d*) occur in four years of tweets, may have to write scripts that take considerable time to complete. Database solutions help for quick access to infrequent and medium frequent patterns, but can still be slow for highly frequent phenomena, such as bigrams consisting of a pronoun and a (frequent) verb.

Second, access to large Twitter corpora is restricted. The University of Groningen has been collecting Dutch language tweets since the end of 2010 (Tjong Kim Sang 2011). A similar corpus has been constructed as part of the project TwiNL, carried out at the Dutch eScience center (Tjong

hits	word	hits	word
116,388	zehma	2,024	zeghma
54,312	ze3ma	2,022	zema
35,987	zegma	1,409	zeh3ma
3,123	ze3hma	209	zechma
2,289	zemma	171	ze7ma

Table 1: Frequency of frequent spelling variants (converted to lower case) of the Moroccan adverb *ze'ma* in 2.5 billion Dutch tweets (2011-2014). Less frequent forms include *ze'ma, zehmma, ze3gma, zehhma, zehgma, zehgma, zeqma, zeggma* and *zeg3ma*.

Kim Sang and van den Bosch 2013).[1] For most researchers, access to this material is restricted to the functionality offered by a web-interface.[2] These interfaces are primarily designed to provide information (examples, frequency, geographical distribution) on tweets containing a given keyword or n-gram. The University of Groningen web-interface is restricted to searches covering at most 1 month. The eScience center interface allows unlimited search, but response times depend on the processing load of the Hadoop cluster that is used to execute queries and can take up to several hours. Unlimited re-distribution of the underlying corpus data is not possible, as re-distribution of (large) samples of tweets is not allowed.

The current data set fills a gap for users that require quick and easy access to frequency of n-grams in the Twitter corpus as a whole. By making the n-gram frequency data available for download, use of this material is not restricted to the functionality of our web-interface, and results obtained using this material can be checked and replicated.

Below, we provide examples of linguistic questions that can be answered using n-gram frequency information alone and some pointers to other collections of n-gram frequency data. Next, we describe the construction of the data, the functionality of the web-interface, representativeness, and options for further work.

2. Motivation

Many questions about language use and language change on social media like Twitter require frequency information about the words or phrasal patterns of interest. For instance, the use of the Moroccan particle *ze'ma* (Boumans 2003) is one of the characteristics of an ethno- or sociolect that is sometimes being referred to as *'Moroccan flavored Dutch'* (Nortier and Dorleijn 2008) and which is used by a substantial number of Twitter users in the Netherlands. The Arabic orthography does not carry over to the Latin alphabet, and thus the number of spelling variants is considerable. Collecting tweets containing this particle requires that one knows at least its most frequent spelling variants. As this type of language use is largely undocumented, and subject to change, researchers will not always have access to this type of information. By searching in the n-gram frequency data for all words of the form `ze%ma` (the % matches arbitrary character sequences) we find the variants shown in Table 1 (along with a number of false hits such as *zendschema* (*program schedule*) or *zeikprogramma* (*shitty program*)). Whereas Boumans (2003) mentions a number of spelling variants but is unclear about their frequency, our results clearly suggest that *zehma* is, or is becoming, the predominant form.

1. Note, that while the technology used in corpus creation is similar for both corpora, they have in fact been created independently. It would be interesting to investigate to what extent both collections overlap or could supplement each other.
2. Groningen: `http://www.let.rug.nl/~kleiweg/bin/dagtwform.py`, eScience center: `http:www.twiqs.nl`

verb	gloss	wel eens		wel is		nog eens		nog is	
		count	%	count	%	count	%	count	%
doen	to do	4,721	76.9	1,423	23.1	10,950	78.3	3,036	21.7
gebeuren	to happen	2,745	80.5	666	19.5	1,512	86.5	237	13.5
horen	to hear	4,994	80.2	1,240	19.8	1,378	87.1	205	12.9
komen	to come	3,744	75.5	1,215	24.5	1,653	78.8	446	21.2
kunnen	can	12,668	90.1	1,407	9.9	2,960	85.5	502	14.5
kijken	to look	6,152	77.6	1,778	22.4	10,490	89.2	1,279	10.8
voorkomen	to happen	440	94.3	27	5.7	166	83.5	33	16.5
weten	to know	13,597	75.9	4,340	24.1	441	87.7	62	12.3
zien	to see	34,051	76.9	10,257	23.1	8,610	73.8	3,060	26.2

Table 2: Frequency of the trigram *wel/nog eens* VERB (standard) and *wel/nog is* VERB (substandard) for a number of infinitival verb forms.

As Halevy et al. (2009) note, for many typical Machine Learning and Natural Language Processing tasks, *'invariably, simple models and a lot of data trump more elaborate models based on less data'*. The same is true to some extent for research in Corpus Linguistics. Carefully composed and annotated corpora are available for many languages, but even the largest annotated corpora are several orders of magnitude smaller than corpora consisting of more or less *ad hoc* samples of printed books (i.e. the Google Books corpus, Michel et al. (2011)), web pages (i.e. the WaCKy[3] and COW[4] corpora, Baroni et al. (2009) and Biemann et al. (2013)) or tweets. The fact that meta-data and linguistic annotation is largely missing from such corpora, is compensated by the fact that there is a lot of it. As a consequence, simple n-gram queries that approximate the linguistic structure of the phenomenon of interest can be effective. For instance, given a tweet like (1), one might wonder how often the adverb *eens* (*once*) is written as *is* in Dutch tweets.

(1) dit kan uiteraard wel is voorkomen in de statistiek
this can obviously positively sometimes happen in the statistics
this can obviously happen sometimes in statistics

As *is* is also a highly frequent form of the auxiliary *zijn* (*to be*), a simple count of *eens* vs. *is* will not do. Manual annotation of a sample of tweets containing the word *is* or *eens* would be tedious for the same reason, as the adverb use of *is* is extremely rare compared to its use as auxiliary. In the trigram *wel is/eens voorkomen*, however, *is* has to be an adverb. In searching a small corpus, this would not be of help, as the counts for both *wel is voorkomen* and *wel eens voorkomen* will probably be close to 0. However, in the Twitter corpus a reasonable number of hits for both trigrams can be found, as wel as for the trigram *nog eens/is V-en* (*again once Vinf*), which can then be taken as estimates of the proportion of *is* vs. *eens* in this context. Table 2 gives results for the verb *voorkomen* and a number of frequent other verbs.[5] It shows that for most verbs, in this context, using *is* as an informal spelling variant of *eens* is by no means exceptional.

Vocabulary and spelling on Twitter are subject to rapid change (Eisenstein et al. 2012, Kulkarni et al. 2014). Therefore, it can also be interesting for linguists to study trends in the relative frequency with which certain forms occur. The use of the word form *ri* as abbreviation for the prepositional use of *richting* (*direction, towards*), for instance, appears to have been a brief trend in Dutch tweets,[6] as

3. `http://wacky.sslmit.unibo.it/`
4. `http://corporafromtheweb.org/`
5. Using the query `[wel,nog] [is,eens] %en`, where we use `%en` to find infinitival verb forms. Apart from some false hits, this also returns many valid pairs of the construction.
6. Thanks to Erik Tjong Kim Sang for pointing this out.

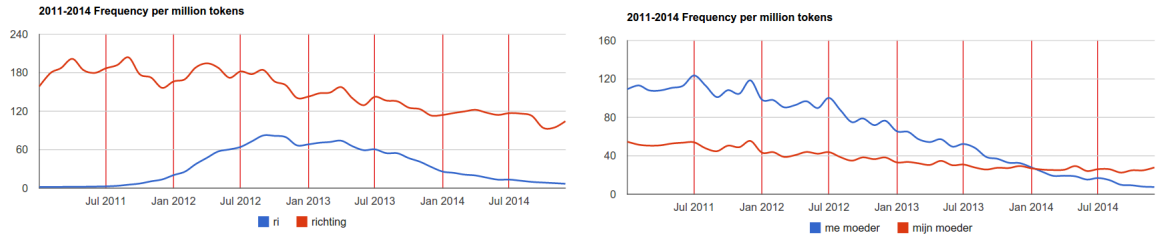

Figure 1: Frequency over time of *richting* and *ri* (*direction*) (left) and the bigrams *mijn moeder* (*my mother*) and *me moeder* (*my mother, informal*) (right).

illustrated in Figure 1 (left pane). The right pane compares the relative frequency of the standard form *mijn moeder* (*my mother*) with its informal counterpart *me moeder*. It shows that while the informal variant was more frequent in 2011-2013, this is no longer the case in 2014. In section 6, we will offer a tentative explanation for this trend.

3. Related Work

There is a tradition in corpus linguistics of publishing frequency data for corpora. For the British National Corpus (100 million words), unigram frequency information[7] is available, and for the Corpus of Contemporary American English (450 million words), unigram frequencies with part-of-speech information, as well as n-gram frequencies[8] are available. The Google web 1T 5-grams data[9] provides frequencies for n-grams up to length 5 based on English text consisting of over 1,000 billion tokens indexed by the Google search engine (Brants and Franz 2006). This source has been used among others for spelling correction (Islam and Inkpen 2009), language modelling (Talbot and Brants 2008), information extraction (Tandon and De Melo 2010), and lexical acquisition of gender and animacy information (Ji and Lin 2009). The Google Books n-grams and n-grams viewer (Michel et al. 2011), based on digitized books covering a period of several centuries, is another large scale resource. It allows trends in n-gram frequency over time to be studied. Recently, part of speech and grammatical (dependency) information have been added (Lin et al. 2012). This data has attracted a lot of attention, not only from linguistics but also as a resource for broader socio-cultural studies, such as Twenge et al. (2012). The Rovereto Twitter Corpus[10] is an n-gram dataset based on almost 75 million English tweets, along with aggregated information on the gender of the authors of the posts and the time of the posting (Herdağdelen 2013). Although the addition of gender information is interesting, it should also be noted that this data-set is derived from a relatively modest source corpus, and thus for many rare phenomena and less frequent longer n-grams the corpus may not provide relevant information.

For Dutch, the most widely used resource for word frequency information has been the CELEX[11] electronic lexical database (Baayen et al. 1993). It provides unigram frequency information based on corpora compiled by the Institute for Dutch Lexicography. This institute also provides an alternative, more recent, data-set with frequency information for the 5,000 most frequent words in several corpora (up to 100 million words).[12] The Corpus of Spoken Dutch (Oostdijk 2000) is a 10 million word speech corpus, for which unigram frequencies are available. SUBTLEX-NL (Keuleers et al. 2010)

7. http://www.kilgarriff.co.uk/bnc-readme.html
8. http://www.ngrams.info/
9. https://catalog.ldc.upenn.edu/LDC2006T13
10. http://clic.cimec.unitn.it/amac/twitter_ngram/
11. https://catalog.ldc.upenn.edu/LDC96L14
12. http://tst-centrale.org/producten/lexica/frequentielijsten-corpora/7-51

word	frequency per 1000 words			
	twitter	web	SUBTLEX	CGN
ik	27.08	2.16	39.88	22.75
je	16.07	3.15	36.60	10.85
de	15.40	16.86	24.26	24.86
en	13.07	10.32	13.98	22.61
een	12.76	7.45	17.97	17.20
het	12.08	7.48	24.43	22.14
is	11.56	4.19	21.66	14.14
niet	10.46	2.24	18.32	11.26
van	9.76	10.90	10.41	13.26
dat	8.53	2.29	22.07	26.26

Table 3: Relative frequency of frequent words on Twitter, the Web, subtitles, and speech. Twitter and web frequencies are summed over lower and upper case word forms. In CGN, frequencies for *ik* and *'k* and for *het* and *'t* are summed.

is a database of Dutch word frequencies based on 44 million words from movies and television subtitles. The authors argue that the frequencies of subtitles reflect everyday language use better than frequencies obtained from written corpora, and for that reason, are better predictors of reading times in psycholinguistic experiments. Google has also published N-gram frequencies for 10 European languages other than English and including Dutch, based on minimally 100 billion tokens of web pages per language.[13]

In table 3, we compare the frequency of 10 highly frequent words in the Twitter corpus, the Google Web corpus, SUBTLEX, and the Corpus of Spoken Dutch (CGN). The first an second person pronouns *ik* and *je* are very frequent in the Twitter, subtitles, and speech data. To the extent that a pattern can be detected in these frequencies, it seems that Twitter frequencies are closer to those of the subtitles and speech corpus, than to those of the web corpus. This suggests that the former three all share some properties of informal and everyday language use that is less visible on the web, which might be closer to news, books, and other written genres.

4. Creating Twitter N-gram Data

Since the end of 2010, the Information Science department of the University of Groningen has been collecting Dutch language tweets along with readily available meta-data such as time of posting, the Twitter profile of the user (including a short user description, and location), and, if available, the geographic coordinates of the posting. Tweets are collected using the Twitter API on the basis of a list of Dutch keywords, as described in Tjong Kim Sang (2011) and Tjong Kim Sang and van den Bosch (2013). The keywords list contains frequent Dutch words which do not or rarely occur in other languages. The number of collected tweets varies between 0.5M and 1.5M per day. Since approximately january 2013, the number of collected tweets is steadily decreasing, which is probably due to less traffic on Twitter and/or changes in the download limits set by Twitter.

4.1 Cleaning and Filtering

Using only the text of the tweets, per month of data (typically between 30 and 40 million tweets), the following steps are performed to arrive at a filtered corpus of tokenized tweets:

13. `https://catalog.ldc.upenn.edu/LDC2009T25`

Filter retweets
⇓
Replace URLs and usernames with a generic placeholder
⇓
Remove duplicates
⇓
Filter non-Dutch tweets
⇓
Tokenize

All tweets containing the token 'RT' are considered retweets. By substituting a generic keyword for URLs and usernames, we can ensure that n-grams that differ only in the URL or username they contain, are seen as identical. Substitution has the additional advantage that no usernames are exposed in the data. Hash tags are left as is, as these sometimes carry meaning that might be relevant for linguistic purposes. Even after removing retweets, many duplicate tweets remain. While some of these might be considered original (i.e. especially one word tweets like *morgen* (*tomorrow*) or *slapen* (*sleep*), the majority of these are to be considered as retweets (i.e. copies of a tweet that went viral, popular sayings, quotes from songs, messages from twitter bots, etc.). Note that we process the data per month, and that we do not attempt to detect duplicates over longer periods of time. As duplicate removal is done after normalization of URLs and usernames, tweets that differ only in this respect are also included only once.[14]

The keyword-based method for collecting tweets is not 100% precise, in that sometimes foreign language tweets are included. Initially, we used an letter n-gram-based approach to filter non-Dutch tweets. The recently added language code in the meta-data provided by Twitter is very accurate in our experience, and so in more recent data, this method is used to filter non-Dutch tweets. To ensure that language identification is applied consistently to all the data, we run a Bayesian language guesser trained on several thousand manually annotated tweets, as an additional filter on all collected tweets. While this does not exclude the possibility that some Dutch language tweets are excluded during the first classification step, it does ensure with reasonable high accuracy[15] that all tweets that remain after language identification are indeed Dutch. Finally, tweets are tokenized using the Alpino tokenizer (van Noord 2006).[16] While the Alpino tokenizer has not been tuned for social media text in particular, it is generally quite robust in dealing with URLs, e-mail addresses, and, to some extent, emoticons. Note that case distinctions are preserved.

In total, we obtain a filtered and tokenized corpus covering a period of 48 months and consisting of 2.685 billion tweets and almost 28.954 billion tokens.

4.2 N-gram creation

N-gram frequencies are computed per month for n-grams up to length 5. All tweets are suffixed with the keyword `XXX_TB_XXX`, indicating the end of a tweet. Next, we compute n-gram frequencies using basic Linux commands as explained in Church (1994). N-grams that start or end with the keyword `XXX_TB_XXX` are kept (to facilitate search for n-grams occurring at the beginning or end of tweet), but all other strings containing the separator keyword are removed.

For the n-gram frequency database that lists counts over the full period of 48 months, counts from all monthly frequency lists are summed and a frequency cut-off of 10 is applied. After application of this cut-off, the database contains 6.65 million distinct unigrams, 61 million bigrams, 135 million trigrams, 147 million 4-grams, and 125 million 5-grams.

14. Various kinds of twitter spamming consist of retweeting more or less random tweets while adding a URL or username to attract traffic to a site of attention of a user.
15. The Naive Bayes classifier was trained on a set of 51658 tweets divided in 33841 Dutch tweets and 17817 non-Dutch. The classifier achieves an accuracy of about 98%.
16. `http://www.let.rug.nl/vannoord/alp/Alpino/AlpinoUserGuide.html`

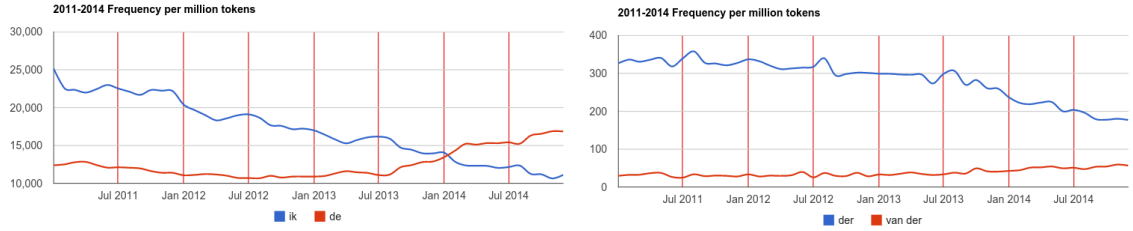

Figure 2: Frequency over 48 months of the highly frequent words '*ik*' ('*I*') and '*de*' ('*the*') and the informal pronoun '*der*' ('*her*'), and the bigram '*van der*' which is part of some last names.

5. Web access

The n-gram frequency data has been made accessible[17] in three different ways, each catering for slightly different use of the data.

5.1 Global n-gram frequencies

The summed n-gram frequencies have been indexed using software originally developed for web access to the Google web n-grams data (Evert 2010). This interface supports (limited) use of pattrn matching and allows search results to be exported in a variety of formats. This interface is particular useful to find spelling variants. The pattern `ge%dt`, for instance, finds (among others) instances of spelling errors where a past participle is written with *-dt* (*gehadt* (e.g. *had*), *gebeurdt*, (*happened*), *gehoordt* (*heard*), and *gezegdt*,(*said*)). It can also be used to find variation in grammatical patterns. The pattern `de %je`, for instance, finds how often a diminutive (which is always neuter) is preceded by a common, non-agreeing, definite determiner. The pattern `[de,het] %je` includes counts for the agreeing case as well, and thus makes it possible to compute the percentage of non-agreement per diminutive. Similarly, `het %je [die,dat]`, approximates the frequency with which a diminutive is followed by either a neuter (agreeing) or common (non-agreeing) relative pronoun.

5.2 Trends

The monthly n-gram frequencies have been used to create an interface with the same basic functionality as the Google Books N-grams viewer.[18] For visualisation, we use Google Charts.[19] One or more n-grams can be entered, and relative monthly frequencies are displayed per selected year or for the whole period of 4 years. Some examples are shown in Figure 2. The left pane shows that, contrary to what Tjong Kim Sang (2011) observed in 2011, in recent Dutch tweets *ik* is no longer the most frequently used word. The right pane shows that the informal spelling of the pronoun '*der*' (mostly used as an informal alternative for '*haar*' ('*her*')), is losing popularity. As '*der*' may also be part of proper names (usually preceded by '*van*'), we include the trend for '*van der*' as well. As the latter shows a slightly increasing trend, decrease of *der* cannot be due to a decrease in occurrence of proper names.

5.3 Word clusters

We provide 1000 word clusters created on the basis of a sample of approximately 25% of all the available data (594 million tweets, 5.8 billion tokens). Only the (769,629) types occurring at least

17. `www.let.rug.nl/gosse/Ngrams`
18. `https://books.google.com/ngrams`
19. `https://developers.google.com/chart/`

'family' moeder ouders vader zusje broertje broer zus pa tante nichtje vriendje neefje oom mams neef nicht ex mannetje vrouwtje zusjes broers broertjes buurvrouw paardje vaders nichtjes neefjes buurmeisje broeder prinsesje zussen mammie schoonzus buurjongen pappie zwager schoonmoeder broertje broer zus pa tante nichtje vriendje neefje oom mams neef nicht ex mannetje vrouwtje zusjes broers broertjes buurvrouw paardje vaders nichtjes neefjes buurmeisje broeder prinsesje zussen mammie schoonzus buurjongen pappie zwager schoonmoeder

'location' trap boot brug berg paal dijk dam ingang rotonde stoep poort tunnel parkeerplaats boerderij ijsbaan plas bossen kruising toren boulevard duinen molen balk maas pont glijbaan stoplichten rivier begraafplaats gracht sluis pier fietsenstalling heuvel roltrap kade bankjes uitgang steiger fontein fakkel noordpool ark ruiter kabelbaan stijger vuurtoren stoeprand stalling helling

'yes' yeah yes Yes hehe Yeah Hehe jeej Jeej yess Yess yeahh Hoppa yeaah yay Yeahh Yeaah yesss jeeej Yesss Yay Jeeej hoppa heuj yeaaah jippie Jaaaaa joepie Joepie woehoe Jippie jeuj Bam Aight yeahhh Jeuj Yeaaah jeeeej Heuj Woehoe Jeeeej yessss Yessss yeaahh Yeahhh jeah Jaaaaaa Officieel Joehoe Enn yeey

'straattaal' ('streetlanguage', sociolect) wollah eey jo aii eeh noh ewa eej swa ej wallah aai joo denkje eyy eeey eeeh jongee juh jow ait vallah ewaa =d eehh jeh alls rustigg saff eeyy aaii aye bruur aaai eee eeej jongeee goos eeeeh swaa jongu walla eeeey ewaja jonguhh saf eyyy mimang hoh teh

clitics (Brabantic) zijt kwas kzal kzou hedde kunde benk moogt geraak is't waart tzal kmis Kzou gade Kzal komde kank suckt kmoe kbn gaak khaat tgaat eeft hebde ebt ebn isda Kmis Ti tga kzen Weeral denkte doede moeje Tzal tzijn gak zedde moej Kdoe tzou Hedde ziede wilk wask kenk eje

Table 4: Fragments of word clusters created using word2vec.

50 times in the corpus are included in the clustering results. These frequent types cover 5.09 billion tokens.

There are two kinds of clusters:

- Flat clusters created on the basis of word vectors created using Google's word2vec[20] (Mikolov et al. 2013) (with the `-cbow` option, window-size 5, vector dimensionality 200 and `-classes 1000`). Word2vec uses K-means clustering.
- Hierarchical clusters created using Percy Liang's implementation of Brown clustering.[21]

The resulting word clusters are formed on the basis of varying semantic or distributional dimensions. The examples in Table 4 show, for instance, that some conceptual dimension can play a role (i.e. clusters of family relations or locations), spelling (spelling variants of *yes* and *no*), but also sociolect (various clusters for informal sociolects) or dialect (among others a cluster with many word forms consisting of a verb and cliticized pronoun, which is typical for Brabantic Dutch).

Clusters like this can be useful for Twitter spelling normalisation (Gouws et al. 2011) or POS-tagging (Owoputi et al. 2013).

6. Representativeness

The trends in the Twitter data highlight a limitation in the data as we collected them. The examples we presented (i.e. Figure 1 and 2) show that the frequency of the first person personal pronoun is decreasing, and that informal forms such as *me moeder* and the pronoun use of *der* are decreasing in frequency as well. We have observed similar tendencies for frequent spelling errors (i.e. incorrect use of verbal inflection), and for other informal forms, such as the use of *hun* as third person plural subject or the use of a reduced and cliticized form of the first person pronoun (i.e. *kheb* (*I have*)).

20. `code.google.com/p/word2vec/`
21. `https://github.com/percyliang/brown-cluster`

The most obvious explanation for these trends is that the Twitter population is changing over time. The use of the first person pronoun as well as the use of informal forms and frequent production of spelling errors is typical for younger users, who use Twitter primarily as a 'chat' medium. While there appears to be a general downward trend in the number of Twitter users, we suspect that this trend is stronger for these young, informal, users. As a consequence, more recent material contains (relatively) larger amounts of tweets produced by professional users. The latter type of user (i.e. professionals representing news agencies or companies in general) primarily use Twitter for information exchange, and are less likely to use the first person pronoun or informal forms.

Trends therefore need to be interpreted with some care. Similar concerns have been expressed concerning trends in other diachronic corpora. Davies (2010), for instance, argues that a diachronic corpus has to consist equal amounts of text from all genres of interest in each period in order to be useful as *monitor* corpus. Pechenick et al. (2015) note that recent decades of the Google Books corpus contain large amounts of scientific text, resulting in a proliferation of phrases referring to time as used in scientific citations.

For traditional diachronic corpora, it seems that at least the addition of meta-data (i.e. genre and author information) is required in order to be able to distinguish between trends in language use and shifts in corpus composition. For social media corpora, addition of author meta-data (i.e. type of user, age, gender) would help to see whether there is an actual change going on in the data (i.e. does the frequency with which teenagers use *me moeder* on Twitter change over time or not?).

7. Future Work

We see several ways in which the current work can be expanded. First of all, web access can be improved by providing more powerful regular expression support in the n-grams interface, by including regular expression support in the trends interface, and by adding the possibility to perform basic arithmetic in search queries (i.e. give the ratio of *me moeder* over *mijn moeder + me moeder* over time).

Another extension would be to combine the current interface, which provides only n-gram frequency data, with a link to (a sample of) the underlying tweets containing this n-gram. We have recently indexed large portions of the corpus as a suffix array, a technique that can be used to support very efficient search over large corpora.

A challenging extension would be the inclusion of user meta-data. While shallow methods can be used to classify users by gender and age (i.e. `laura1985` → `female, birthyear 1985`), such methods often have low recall. Using machine learning helps (Nguyen et al. 2014, van Halteren and Speerstra 2014), but the accuracy of the resulting automatic classifiers is not perfect, while scalability is still hindered by the fact that only users for which a substantial number of tweets are available can be classified with reasonable accuracy. Apart from trying to detect attributes like gender and age, it can also be interesting to perform a wider classification of users, where twitterbots and automatically generated tweets (as produced by wheather stations and many sports apps for instance) can be detected (Chu et al. 2010).

A final advanced project is to add part-of-speech information. This would obviously open up novel ways to search for linguistic patterns. Automatic part-of-speech tagging of Dutch tweets is currently an active research area (Aminian et al. 2012).

References

Aminian, Mehdi, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, and Menno van Zaanen (2012), Assigning part-of-speech to Dutch tweets, *Proceedings of the LREC workshop:@ NLP can u tag# user generated content*, pp. 9–14.

Baayen, R. H., R. Piepenbrock, and H. van Rijn (1993), *The CELEX Lexical Database (CD-ROM)*, Linguistic Data Consortium, UPenn, Philadelphia, PA.

Baldwin, Tim, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang (2013), How noisy social media text, how dffrnt social media sources, *International Joint Conference on Natural Language Processing*.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta (2009), The WaCky wide web: a collection of very large linguistically processed web-crawled corpora, *Language resources and evaluation* **43** (3), pp. 209–226, Springer.

Biemann, Chris, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch (2013), Scalable construction of high-quality web corpora, *Journal for Language Technology and Computational Linguistics* **28** (2), pp. 23–60.

Boumans, Louis (2003), Ze3ma. een Noordafrikaans epistemisch partikel dat zich verspreidt, *Gramma/TTT* **10** (1), pp. 1–26.

Brants, Thorsten and Alex Franz (2006), Web 1t 5-gram version 1, Linguistic Data Consortium, Philadelphia.

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia (2010), Who is tweeting on Twitter: human, bot, or cyborg?, *Proceedings of the 26th annual computer security applications conference*, ACM, pp. 21–30.

Church, Kenneth Ward (1994), Unix for poets, *Notes of a course from the European Summer School on Language and Speech Communication, Corpus Based Methods*.

Davies, Mark (2010), The Corpus of Contemporary American English as the first reliable monitor corpus of English, *Literary and linguistic computing* p. fqq018, ALLC.

Eisenstein, Jacob (2013), What to do about bad language on the internet, *Proceedings of NAACL-HLT*, pp. 359–369.

Eisenstein, Jacob, Brendan O'Connor, Noah A Smith, and Eric P Xing (2012), Mapping the geographical diffusion of new words, *arXiv preprint arXiv:1210.5268*.

Evert, Stefan (2010), Google Web 1T 5-Grams made easy (but not for the computer), *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics, pp. 32–40.

Gouws, Stephan, Dirk Hovy, and Donald Metzler (2011), Unsupervised mining of lexical variants from noisy text, *Proceedings of the First workshop on Unsupervised Learning in NLP*, Association for Computational Linguistics, pp. 82–90.

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009), The unreasonable effectiveness of data, *Intelligent Systems, IEEE* **24** (2), pp. 8–12, IEEE.

Herdağdelen, Amaç (2013), Twitter n-gram corpus with demographic metadata, *Language resources and evaluation* **47** (4), pp. 1127–1147, Springer.

Islam, Aminul and Diana Inkpen (2009), Real-word spelling correction using google web 1tn-gram data set, *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, pp. 1689–1692.

Ji, Heng and Dekang Lin (2009), Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection., *PACLIC*, pp. 220–229.

Keuleers, Emmanuel, Marc Brysbaert, and Boris New (2010), SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles, *Behavior research methods* **42** (3), pp. 643–650, Springer.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena (2014), Statistically significant detection of linguistic change, *arXiv preprint arXiv:1411.3315*.

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov (2012), Syntactic annotations for the Google Books Ngram Corpus, *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 169–174. http://dl.acm.org/citation.cfm?id=2390470.2390499.

Michel, J.B., Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. (2011), Quantitative analysis of culture using millions of digitized books, *science* **331** (6014), pp. 176–182, American Association for the Advancement of Science.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013), Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119.

Nguyen, Dong, Dolf Trieschnigg, and Theo Meder (2014), Tweetgenie: Development, evaluation, and lessons learned, *Proceedings of COLING*.

Nortier, Jacomine and Margreet Dorleijn (2008), A Moroccan accent in Dutch: A sociocultural style restricted to the Moroccan community?, *International Journal of Bilingualism* **12** (1-2), pp. 125–142, SAGE Publications.

Oostdijk, Nelleke (2000), The Spoken Dutch Corpus: Overview and first evaluation, *Proceedings of LREC 2000*, pp. 887–894.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith (2013), Improved part-of-speech tagging for online conversational text with word clusters, Association for Computational Linguistics.

Pechenick, Eitan Adam, Christopher M Danforth, and Peter Sheridan Dodds (2015), Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution, *arXiv preprint arXiv:1501.00960*.

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2010), Streaming first story detection with application to Twitter, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 181–189. http://dl.acm.org/citation.cfm?id=1857999.1858020.

Talbot, David and Thorsten Brants (2008), Randomized language models via perfect hash functions., *ACL*, Vol. 8, pp. 505–513.

Tandon, Niket and Gerard De Melo (2010), Information extraction from web-scale n-gram data, *Web N-gram Workshop*, Vol. 7.

Tjong Kim Sang, Erik (2011), Het gebruik van Twitter voor taalkundig onderzoek, *TABU: Bulletin voor Taalwetenschap* **39** (1/2), pp. 62–72.

Tjong Kim Sang, Erik and Antal van den Bosch (2013), Dealing with big data: The case of Twitter, *Computational Linguistics in the Netherlands Journal* **3**, pp. 121–134.

Twenge, Jean M, W Keith Campbell, and Brittany Gentile (2012), Increases in individualistic words and phrases in American books, 1960–2008, *PloS one* **7** (7), pp. e40181, Public Library of Science.

van Halteren, Hans and Nander Speerstra (2014), Gender recognition on dutch tweets, *Computational Linguistics in the Netherlands Journal* **4**, pp. 171–190.

van Noord, Gertjan (2006), At last parsing is now operational, *in* Mertens, Piet, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42.