Hans van Halteren & Nelleke Oostdijk

Word Distributions in Dutch Tweets

A quantitative appraisal of the distinction between function and content words

Abstract – In this paper, we investigate the distinction between function words and content words in the light of their distribution over domains and/or topics. More specifically, we investigate whether this distinction should be viewed as a dichotomy or rather as a continuum. Observing that function words ought to be generally applicable while content words are domain/topic-dependent, we measure how widely words are being used, by examining their distribution over the 1,000 most frequent hashtags on Twitter. Based on the results of these measurements, we conclude that a continuum ranging from fully grammatical words to fully content-bearing words is the more promising viewpoint.

1 Introduction

Since classical times, linguists have attempted to devise classification systems for the words found in various languages. Much like the early biologists, linguists have had to rely on limited observations and their intuition. In biology, the Linnaean system has only partially survived after the advent of DNA-based comparisons and access to a much larger evidence base. In linguistics, conventional ideas about what constitute the basic descriptive units are similarly challenged now that linguists can tap into and examine the written language production of the general public through their presence on social media platforms, and computer systems and statistical methods have grown powerful enough to process amounts of text that are beyond anything anyone could have imagined before.

In this paper, we investigate the extent to which words are usable equally well with various domains and topics, or reversely how bound they are to specific domains and/or topics. In linguistics, this question has been addressed mostly in terms of syntactic parts of speech.[1] Some parts of speech have been declared to be *function words*, that is, words which provide the grammatical glue with which sentences (and texts) can be built. Their meaning is concentrated within the text and it builds on the other words that they link together. These other words are the *content words* (or: *lexical words*), which refer to the extra-linguistic entities, qualities and events about which the text makes propositions. In the most extreme view of this division, function words should be equally usable within any domain and about any topic, whereas the content words should be restricted to the domains and topics that they refer to. This extreme view, however, is not subscribed to by many linguists, as it hardly ever survives a confrontation with actual language data and

[1] In information retrieval, where it is useful to exclude topic-independent words from the retrieval processes, hand-crafted lists (so-called *stoplists*) of words with specific parts of speech have been largely replaced by the use of inverse document frequencies (IDF; Spärck Jones 1972).

has been the subject of many discussions and proposals for amendments. It has already been suggested that the split between the function and content words should not necessarily be made on the part-of-speech level, but also within specific parts of speech (Haspelmath 2001), and that '[F]unction words and lexical words are not sharply distinct categories but rather form a continuum' (van Gelderen 2005).

In order to evaluate the merit of the various viewpoints, we measure the presence of various classes of words in different domains and with different topics. We take our measurements on texts which have been produced by many different types of authors, namely tweets posted on the social media platform Twitter. We rely on the hashtag to identify the topic (and hence also the domain). In other words, our research question in this paper is

> To what degree does the part of speech of a word influence that word's distributions in texts covering various topics/domains and produced by the general public?

Below (in Section 2), we first take a closer look at the notion of function and content words, as well as at doubts that have been raised as to whether there is a strict dichotomy between the two classes. Then we will argue for the use of hashtag-marked data on Twitter as the basis of our word distribution measurements (Section 3) and describe the data collection[2] eventually used (Section 4). In Section 5, we move on to the manner in which we measured the word distributions, after which we can turn to an examination of the distribution of some specific word types (Section 6). In our conclusion (Section 7) we summarize the most important findings and make suggestions for future research. The paper also includes three appendices describing the creation of the experimental data collection (I), technical details on measuring how widely words are used (II), and a comparison of our measure with existing dispersion measures (III).

2 Function words and content words

When speaking of word classes, commonly a distinction is made between *function* (or: *functional* or *closed class*) and *content* (or: *lexical* or *open class*) words:

> Word classes are normally divided into two. The open or lexical categories are ones like nouns, verbs and adjectives which have a large number of members, and to which new words are commonly added. The closed or functional categories are categories such as prepositions and determiners (containing words like *of*, *on*, *the*, *a*) which only have a few members, and the members of which normally have a clear grammatical use (Manning & Schütze 1999: 82).

Grammatical use is also the main motivating argument for Quirk et al. (1985: 71f) for making the distinction. After defining various word classes, they go on to distinguish between open and closed word classes.[3] For the latter they also intro-

[2] The reader should be aware that data collection can refer to both the process of collecting and the result of the collecting. However, in this paper, we use the term exclusively for the resulting data.

[3] With numerals and interjections as minor types – and thus in-between open-class words and closed-class words.

duce the terms ‘grammatical words’ and ‘structure words’, since ‘These terms also stress their function in the grammatical sense, as structural markers: thus a determiner typically signals the beginning of a noun phrase, a preposition the beginning of a prepositional phrase, a conjunction the beginning of a clause’ (idem: 72).

For Dutch, the object language of our study, the same division between function words (‘*functiewoorden*’) and content words (‘*inhoudswoorden*’) is made, and, just as for English, there is variation in both the description and in the composition of the two groups, as we will see in some examples. Van Wijk & Kempen (1979) claim to present a full inventory of function words. In their introduction, they place the nouns, adjectives, verbs and a large number of adverbs under the content words, these being the words that refer to a non-linguistic reality and for a large part determine the meaning of a sentence. The other group (not specified in their introduction), the function words, are said to play a role mostly inside the language, not referring to anything outside it but contributing to an important degree to the grammatical structure of the sentence. The function words are listed as comprising prepositions, conjunctions, articles, (indefinite) numerals, pronouns, proforms (e.g. *sindsdien* (‘since then’)), auxiliary and copula verbs, connective adverbs (e.g. *integendeel* (‘on the contrary’)), intensifying adverbs (e.g. *nogal* (‘rather’)), and qualifying adverbs (e.g. *ongeveer* (‘about’)). Van Wijk & Kempen provide a full list of function words, derived from an authoritative dictionary, a grammar and a corpus-based frequency list. In a study of words entering the Dutch language, Van der Sijs (2002) distinguishes between function words, content words, and interjections. The function words are said to be grammatical words that serve to express certain relations between parts of the sentence. They have no meaning by themselves, but form the cement of the sentence and as such cannot be left out. The word classes listed within this group are the articles, numerals, conjunctions, pronouns, and prepositions. The content words are not characterized. Elbers & Wijnen (1989) investigate the development of the use of function words by a two-year-old child. They list the content words (being proper names, nouns, adjectives, lexical verbs, and adverbs that can also be used adjectivally), and state that they ignored the interjections. The function words are then defined as ‘the rest’, being ‘pronouns, prepositions, auxiliary verbs, articles, etc.’.

There are also those that propose that the distinction between function and content words should be viewed more as a continuum. For example, Haspelmath (2001) looks at word classes and parts of speech from a cross-linguistic perspective. He observes that while there is sometimes disagreement of the assignment of words and even entire word classes to either the category of content words or function words, ‘their usefulness and importance is not in doubt’ (idem: 16539).[4] It is ‘the precise delimitation of function words and content words’ that Haspelmath finds is often problematic. The explanation he offers is that function words enter into a language as the result of an ongoing grammaticalization process in which over time content words develop into function words which in turn at a later stage may develop into clitics and eventually affixes. In this light the distinction between function words on the one hand and content words on the other

[4] Haspelmath lists nouns, verbs, adjectives, and adverbs as content words, and adpositions, conjunctions, articles, auxiliary verbs and particles as function words.

should not be viewed as a strict dichotomy; rather, it would appear that there is 'a gradient from full content words to clear function words' (ibid.). A view similar to Haspelmath's is upheld by van Gelderen (2005) who, also from a cross-linguistic perspective, observes that there appears to be a continuum, such that 'certain classes of words can share features with prototypical lexical words and prototypical function words'.

Given that function words are most often characterized as primarily serving to structure a sentence or text, one would expect such words to be generally usable, i.e. in texts of all domains and topics. Content words on the other hand should be more restricted in their use, in that they represent specific notions, and can therefore only be applied with a topic/domain in which those notions play a role. This is supported by observations about the distributional properties of words, e.g. in Fagan & Gençay (2011: 140):

> While function words appear to occur fairly uniformly throughout documents, content words appear to cluster. Zipf (1932) noted this phenomenon by examining the distance (D) between occurrences of a given word, and then calculating the frequency (F) of these distances. … This implies that most content words occur near other occurrences of the same word.

The clustering of content words is most likely due to the fact that their related topics are only discussed at specific points in the documents.

While the association of function words with general applicability and content words with topicality would seem appropriate at first sight, there is also evidence that suggests that this point of view is too simplistic. Thus, the number of topics/domains in which content words can be used will vary greatly. Moreover, content words may be used metaphorically. A word like *bal* ('ball'), for example, appears to have a rather narrow focus, mostly related to sports, but gains a wider usability by metaphorical uses such as *aan de bal* (lit.: 'on the ball', i.e. 'having to act') and incorporation into proverbs such as *de bal is rond* (lit: 'the ball is round', i.e. 'anything can happen'). Therefore, we prefer to introduce a measure, the *ubiquity score* (or U-score for short), that does not imply an a priori binary classification, but rather permits measurements involving a numerical scale.[5] This scale ranges from 1 to 0 and indicates how widely the word can be applied. For each word, a U-score can be computed. A U-score of 1 means that the word is generally applicable, whereas 0 means that the word is extremely restricted as regards its applicability across different topics and domains.

Below, we will investigate whether indeed the words traditionally classified as function words all have high U-score values, and which content words have comparable scores, showing that they appear to have general applicability.

[5] The U-score will be introduced in more detail in Section 5 and Appendix II, and compared to some existing measures of dispersion in Appendix III.

3 Twitter data as the basis for language research

The perfect data source for our investigation can be found in the social media platform Twitter. It has all the qualities that we require: it contains so much text that we can take adequate measurements for a substantial number of words, the texts are produced by the general public rather than by a select few professional authors, and a substantial percentage of the texts is explicitly marked as belonging to a specific topic/domain by means of so-called hashtags. However, using Twitter as a source for linguistic research is as yet rather uncommon, and may therefore encounter some skepticism, especially by those already questioning the use of internet in general as an alternative to traditional corpora. For this reason, we will start by discussing the various objections that might be made against Twitter as a data source, and argue that they are either not valid, or not strong enough to outweigh the advantages. Then, we continue by discussing the appropriateness of the sample we took from Twitter.

3.1 The appropriateness of Twitter data

In recent years, with the rise of the internet and the advent of the social media, linguists interested in studying actual language use have gained access to an unprecedented wealth of data. In the past, many linguistic studies would relate to the data present in the various corpora. The traditional sample corpora were typically relatively small scale, carefully designed, balanced collections of data produced by professional and semi-professional authors. With the advancements in storage and processing capabilities, and the wider availability of digitized data, corpora grew increasingly larger and eventually were surpassed by data collections harvested from the internet. What followed was a heated debate about the status of the internet as corpus, a debate which has kept linguists divided to date. A central issue here pertains to the origin and authenticity of the data. Admittedly, the internet holds huge amounts of data, obviously more than one would find in any corpus, and each day more data are added. However, often we have no way of knowing exactly where the data originated from, and information which is generally considered vital in the context of linguistic research into language use is missing. For example, when were the data produced and by whom? How do we know if the author was a native speaker? With corpora such information is usually available in the metadata, with the internet it often is not. Well, all these caveats are valid for the internet as a whole, but they do not necessarily apply to the special case of Twitter.

First, tweets (the rather short, 140-character text messages sent on Twitter) do in fact come with some kind of metadata. With each tweet there is a time stamp which tells us exactly when a tweet was posted. Moreover, it is noted when tweets are simply copied ('retweeted'). Where (part of) a tweet is retweeted, this is often indicated in the text, so that we are able to remove most of the non-original material, something which is much harder with data harvested from the internet in general.

As for authorship of the text, the tweet metadata contains a user name. We may not have any detailed information about the identity of each author, but we do know whether tweets were produced by the same author or by different authors.[6] As we will see below (Section 4, Appendix 1), this will enable us to filter out many tweets we would rather exclude from our research, such as those produced by software ('bots') rather than by people. If in our research we wanted to target only authors that were native speakers of the language, then this would be a problem as this information is not available. However, in the present context we do not think that uncertainty about the authors' nativeness is an issue. One should realize that for Dutch, other than for English, the percentage of non-native authors will be relatively low generally, and even more so when we consider the social media. Not being able to control for nativeness will hardly influence large-scale measurements such as those featuring in the present study.[7] What we do want to avoid are tweets which in fact are not meant to be Dutch, but these as well we are able to filter out to a very large degree (Section 4, Appendix 1).

Obviously, as information about the author's age, educational background, etcetera is missing from the metadata, the data cannot be used for studies where such information is key to the research questions addressed. For the research question in this paper, we do not think that the missing metadata is prohibitive.

Now, our choice for Twitter does have some disadvantages, the most important of which is that it leads to a restriction in the text types that are observed. Obviously, Twitter's limitation of posted messages to 140 characters has its repercussions in the kinds of text that can be produced. However, we expect that the influence on our measurements is minor, given that we are only looking at distributions of lexical items.[8] Furthermore, hashtag-marked tweets are also clearly only a subspace of the full text type space, which means we may have to be careful in generalizing any conclusions we draw on the basis of our collection. But all in all we think that using Twitter data is an excellent option for addressing our research question, and certainly far better than any other available corpus or data collection.

3.2 The appropriateness of the sampling strategy

Now that we have established that Twitter provides an acceptable basis for our research, we can proceed to show that it is possible to develop a strategy that can provide us with a representative sample.

The first point we need to address is that we will not be sampling directly from Twitter, but from an existing sample of Dutch tweets. This sample is the TwiNL collection, originating from the TwiNL project that ran from September 2012 until February 2013 and which 'focused on developing a website, twiqs.nl, which enables researchers, students and other interested people to search through Dutch tweets and examine visualized summaries of the search results.' (http://ifarm.nl/erikt/twinl/about-twinl/)

[6] Barring the rare situations where a new user starts using a user name, previously used by another author who has since stopped using it.

[7] Although they may crop when looking for extreme examples.

[8] But there will certainly be influences. For example, we expect that subordinating conjunctions may be observed less often than in longer texts, because 140 characters are not conducive to complex sentence structures.

The collection is now available from the Dutch eScience Centre (Tjong Kim Sang & van den Bosch 2013), and comprises tweets dating from the end of 2010 onwards. Dutch tweets were harvested by two strategies. One strategy is by keyword search, using 229 Dutch words and hashtags. The other strategy was 'to gather all messages from a ranked list of 5,000 users who post messages in Dutch most frequently. This data stream does not reach the maximum number of messages that can be retrieved. However, 5,000 is the maximum number of users that may be tracked this way, which is substantially less than the estimated number of Dutch users on Twitter (about one million)' (idem: 123).[9] The harvested tweets are subsequently subjected to two language filters, viz. libTextCat[10] and Twitter's own language identifier [11] and tweets were accepted as being Dutch if either filter accepted them as such. This strategy is estimated to lead to a minimal amount of noise in the form of non-Dutch tweets (about 2.5%), at the cost of removing almost 9% of genuine Dutch tweets (idem: 124). The total TwiNL collection is estimated to contain about 40% of all public Dutch tweets (idem: 132).

The issue here, then, is how representative the TwiNL sample is for the population of Dutch tweets. The sampling procedure introduces several biases. The first harvesting strategy demands that one of 229 specific words and hashtags is present, which may well rule out specific tweet types. The same is true for the language filters. Especially tweets written in dialect or street language are likely to be underrepresented, but this will not affect the current investigation as we are interested in standard Dutch, and not its local varieties. The second harvesting strategy demands that the tweet is produced by a prolific author. This may bring back some of the missing tweet types, but it is likely that prolific authors are also experienced authors, leading to an underrepresentation of less experienced authors. We do not know the relative numbers of tweets stemming from the two strategies, nor do we know how many of the prolific authors will be removed by our own filters (Section 4, Appendix 1). After a cursory inspection of our final collection, we do not hold the impression that the data collection is compromised. Apart from the harvesting strategies, tweets may also be lost due to Twitter's restrictions on the amount of data that can be downloaded. However, to our knowledge, the filtering there is completely random, and should not influence representativeness.

Thus TwiNL provides an acceptable basis, so that we can now proceed to compile a 'controlled' subset of tweets on which we can take the measurements needed in order to answer our research question. We selected this set on the basis of a number of criteria.[12] The first of these were on principle. As we only wanted

[9] 'There are regular estimates about how many people in The Netherlands exactly use Twitter. The most recent estimate was released this week from the German market research institute GfK (Gesellschaft för Konsumforschung). They report that in The Netherlands 3.5 million people have a Twitter account and these produce about 1.7 million messages per day. The user count is similar to the estimate of TwitterMania of June 2013: 3.4 million users. The message estimate is lower than our own estimate for messages in Dutch in the past week in December 2013: 2.4 million messages per day. But our estimate includes the messages from Flanders. If these amount to about 0.7 million messages per day and the number of Dutch messages from outside The Netherlands equals the number of non-Dutch messages from The Netherlands then these numbers match' (Post 13-12-2013 on TwiNL).

[10] http://software.wise-guys.nl/libtextcat/.

[11] https://blog.twitter.com/2013/introducing-new-metadata-for-tweets.

[12] See Appendix I for a full description of the selection of our dataset.

tweets produced by human Dutch users, we needed to filter out any foreign language tweets picked up by accident in the TwiNL harvesting procedure (the 2.5% mentioned above), as well as tweets produced by bots. Also, in order to avoid an a priori bias in the word distributions in our data set, users that just kept repeating the same or very similar text over and over again (such as marketeers) had to be excluded.

The next criterion was driven by the current paper's focus on word distributions. We wanted to investigate to what degree words are either used with a wide variety of topics, or appear only with a small number of topics. We took the hashtag as a proxy for determining a tweet's topic and in our data set only included tweets containing at least one hashtag.[13] To ensure that the data set would include as large a number of tweets as possible for each topic, we restricted ourselves to the 1,000 most productive hashtags.

The final criterion was a purely pragmatic one: the type of investigation pursued here requires ample amounts of data, while at the same time, in order to keep things practical, we obviously need to restrict ourselves to a dataset that can be processed within a reasonable amount of time. Therefore, we decided to use those tweets from TwiNL which had a date stamp from January 1, 2011 to June 30, 2013.

4 Our data collection: Dutch tweets containing frequent hashtags

As described above, we built our data collection of Dutch tweets for the current investigation on the basis of the TwiNL data collection.

4.1 Selection and preprocessing[14]

From the TwiNL collection, we selected all tweets with a date stamp from January 1, 2011 to June 30, 2013, and containing at least one of the 1,000 most productive hashtags. We filtered out tweets which we did not consider to be 'normal' original Dutch Twitter language use, e.g. tweets not in Dutch, retweeted, produced by bots, or overly repetitive. Furthermore, we tokenized all text samples with our own specialized tokenizer for tweets. Finally, as the use of capitalization and diacritics is found to be quite haphazard in tweets, we stripped all words of diacritics and converted them to lower case.

The initial data collection contained about 3.8 million users, producing a total of about 1.5 billion tokens. Of these, about 670,000 users (18%), collectively producing a total of about 85 million tokens (5.5%), were removed because their tweets on the whole did not appear to be written in Dutch. A further 173,000 users (5%), collectively accountable for a total of about 327 million tokens (22%), were removed because vocabulary measurements showed that they produced text which deviated significantly from the average.

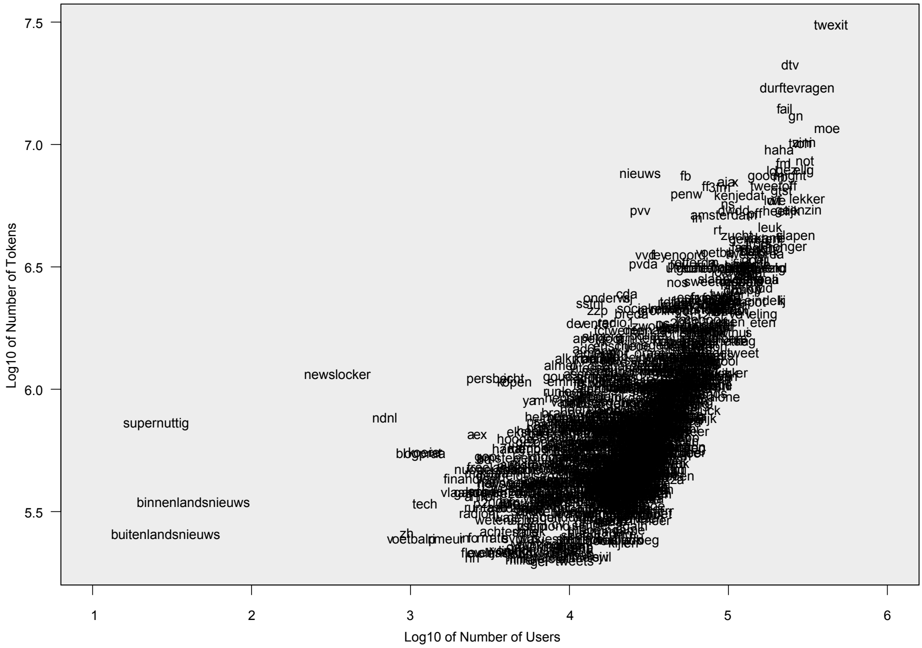
[13] We also considered automatically classifying tweets as to the topics addressed in them. This idea was discarded as reliable classification on the basis of 140 characters was deemed too hard a task.

[14] A more complete description of the data selection and preprocessing can be found in Appendix 1.

4.2 *Statistics*

Our final data collection, then, comprises Dutch tweets posted from January 1, 2011 to June 30, 2013. It contains about 1 billion tokens, published under 2.95 million different user names. It is divided into 1,000 datasets, each representing a single hashtag.[15] As can be expected, the size of the datasets varies considerably, as can be seen in Figure 1. Most of the datasets have between 10,000 and 100,000 contributors. As can be expected, the number of tokens generally goes up with the number of users, from 300,000 to 3,000,000. On the left of the main cloud, we find hashtags where a smaller number of users produce relatively more tokens. These include hashtags relating to the various news feeds (e.g. #binnenlandsnieuws ('national news'), #buitenlandsnieuws ('foreign news'), #newslocker, #persbericht ('press release'), #voetbalprimeur ('football scoop') and, higher up, #nieuws ('news')). Also high, but closer to the cloud, we find hashtags that are related to some major Dutch political parties (#cda, #pvda, #vvd and #pvv). Towards the right of the plot, we find the hashtags with the largest number of users. Remarkable here are the (often short) personal status reports (#thuis ('home'), #eten ('eat'), #slapen ('sleep')), and in the top right corner of the plot a group of metatags (#twexit, #dtv ('dare to ask'), #durftevragen ('dare to ask'), #fail).

Fig. 1 Total number of tokens and users associated with each hashtag



[15] If a tweet contains more than one sampled hashtag, the tweet will be repeated in each dataset. For the investigation in this paper, such repetition will have minimal effect on the statistics, as each dataset is examined separately. For later research where repetition may be undesired, we will have to correct for it.

5 Measuring Word Distributions

After compiling the datasets comprising Dutch tweets found with the 1,000 frequent hashtags, we continued by investigating the distribution of individual words, that is, we needed to establish, for each word, whether it is being used with all hashtags or only with some hashtags. Given a well-balanced data collection in which each dataset is sufficiently large, we would only have to check if each type actually occurs with all hashtags.[16] However, our data collection is not all that balanced. As we have just seen, the datasets for the various hashtags vary in size and some datasets are relatively small. Simply checking for the ubiquitously present types would only allow the investigation of a small group of extremely frequent word types. We therefore approached the problem in another way.[17]

For each individual word under investigation, we took a random sample of 100,000 of the word's observations. We based the sampling on relative frequencies per hashtag dataset, so that the samples would not be biased by the difference in size between the datasets. We then used the number of observations of the word within each dataset in this random sample as the frequency of that word with the corresponding hashtag. We then divided the various possible frequencies (from 0 to 100,000) into frequency bands, namely 0 (band 0), 1 (band 1), 2-3 (band 2), 4-7 (band 3), etc. With a maximum frequency of 100,000, the highest frequency band is 17. Now, for each frequency band, we count how many datasets there are in which the word has a frequency in that band.

In theory, a fully generally applicable word would be equally likely for all hashtags, and can therefore be expected to be selected an equal number of times in each dataset, which for 100,000 observations is 100 times. This means that for such a word, all hashtag datasets fall in frequency band 7 (frequency 64-127 tokens). However, this distribution with all hashtag datasets in band 7 is an ideal one, and is unlikely to be found for actual words.[18] If we look at the words with at least 100,000 observations in our collection, and with the highest peak at band 7, we see that the top-5 is formed by *aan* ('to'), *met* ('with'), *en* ('and'), *een* ('a') and *om* ('in order to', 'at (time)'). We plot their distributions in Figure 2, with green lines. The five distributions are almost identical, with almost 800 hashtag datasets falling in band 7. There is some variability which is partly due to the fact that in the datasets we find slight differences in language use, and is partly a result of the random sampling.

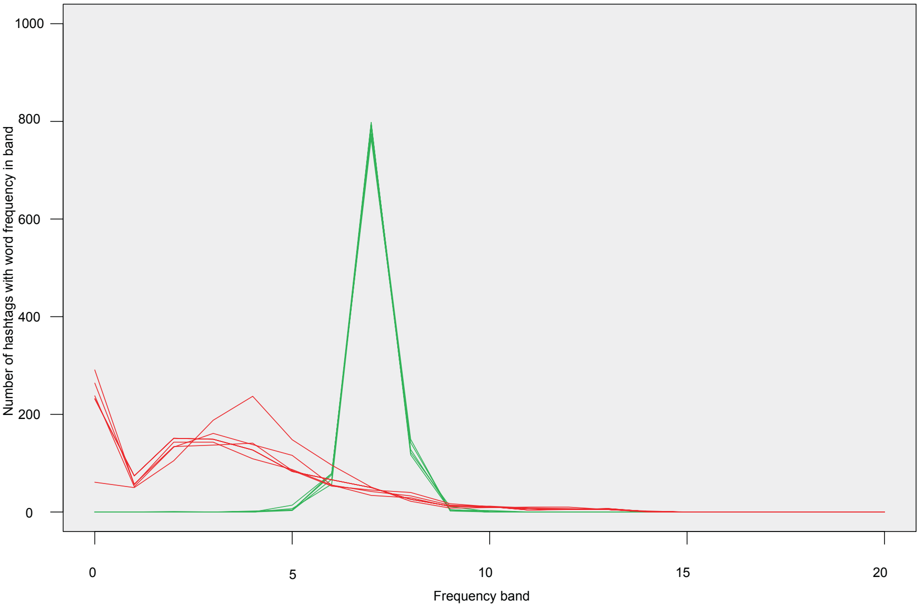
The most restricted word would be applicable with only a single hashtag, and would have that hashtag in frequency band 17 and all other hashtags in frequency band 0. This is also an extreme that we do not find in our data collection. If we take the words with at least 100,000 observations and with the lowest number in band 7, we see that the top-5 words are themselves all hashtags, four for political

[16] Which would correspond to measuring each word's inverse document frequency (IDF; Spärck Jones 1972).

[17] A more complete description of the measurement can be found in Appendix II.

[18] This ideal distribution is the one that most existing measures for word dispersion compare against, such as Juilland et al.'s D (1971) and Gries' DP (2008). The main innovation of our measure is that we compare against a linguistic reality rather than a mathematical ideal. A more detailed comparison between the various measures can be found in Appendix III.

Fig. 2 Frequency distributions for various words over the 1,000 examined hashtags*



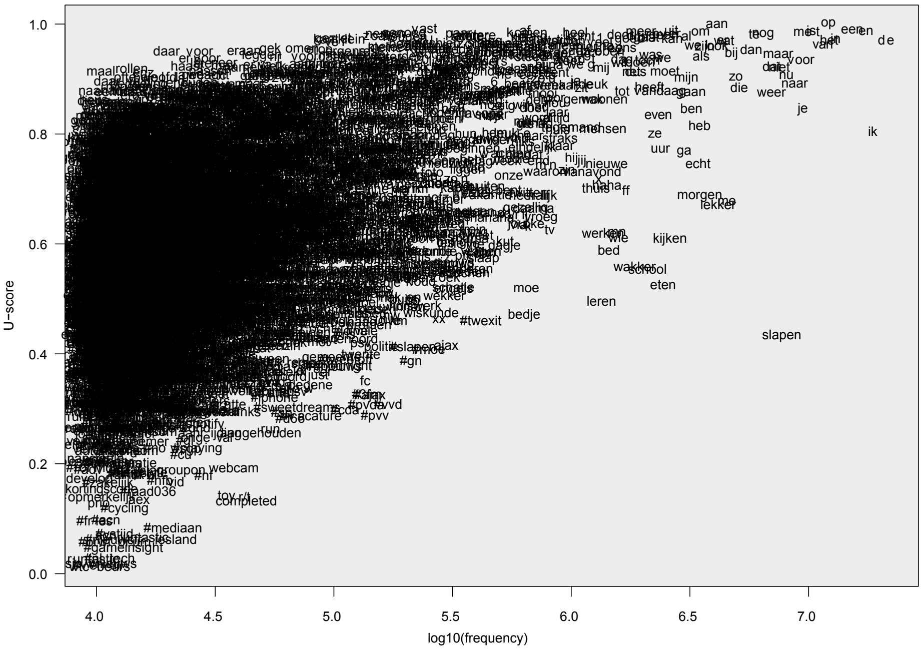
* i.e. the number of hashtag datasets where a word has its observed frequency falling in the given frequency band. The green lines represent a random sample of 100,000 occurrences of 5 common function words (aan ('to'), met ('with'), en ('and'), een ('a'), om ('in order to', 'around'). The red lines represent a random sample of 100,000 occurrences of 5 hashtags (four political parties: #cda, #pvda, #pvv, #vvd; one football team: #ajax).

parties (#cda, #pvda, #pvv, #vvd) and the last one for a football team (#ajax). We plot their distributions in Figure 2 in red. These distributions differ somewhat more, although the four political hashtags are again very similar. In these distributions, we see three local peaks. The first of these is at zero. This peak represents the hashtag datasets where the hashtag words are not present at all. The second peak represents the bulk of datasets. Here the hashtag words do occur, but much less frequently so than the function words we saw above (*aan, met, en, een* en *om)*. Finally, there is a small peak at the high frequency bands, which relates to a few hashtag datasets that apparently represent the home domains/topics for these hashtag words, and where these hashtag words are used much more frequently than the function words. For the political parties, these are mostly datasets for hashtags for other parties, politicians, elections and politics-oriented television programmes. For #ajax, they are other Ajax-related hashtags, such as the supporters' association (#afca) and the home stadium (#arena), other football teams, such as AZ (#az) and FC Twente (#twente), matches, such as FC Twente-Ajax (#twea-ja), competitions, such as the Dutch Premier League (#eredivisie) and football-oriented television programs and periodicals, such as Voetbal International (#vi).

On the basis of the frequency band distributions, we calculated a *ubiquity score* (U-score) for each word.[19] The most widely applicable words that we actually ob-

[19] A more exact technical explanation of the calculation can be found in Appendix II.

Fig. 3 Frequencies and U-score measurements for all 5,520 examined words



served, namely *aan* and *op*, received a U-score of 1.[20] The theoretically most restricted word received a U-score of 0.[21]

The full set of measured words is shown in Figure 3.[22] On the horizontal axis, we find the number of actual observations in our dataset, on the vertical axis the U-score measurement. We see that the most frequent words almost all have a high U-score, but a high U-score does not imply a high frequency. Another observation to be made in relation to Figure 3 is that there is a group of words, such as *slapen* ('sleep'), *eten* ('eat') and *ik* ('I'), that is positioned quite far below the diagonal, which means that they are rather more frequent than one might expect on the basis of the U-score. The words found in this group are clearly linked to typical, popular topics encountered on Twitter, which implies that there is an infl -ence of the text type on the resulting measurements. Still, we expect this influence not to have a major impact on the research we present in this paper.

In the next section, we will examine some types of words in more detail.

[20] For the maximum value, we preferred actual words to an unattainable ideal here, and we could assume that truly ubiquitous words ought to be observed in a data collection as large as ours, so that we could indeed use *aan* and *op*.

[21] For the minimum value, we could not assume that it would be present in our collection; therefore we chose the theoretical extreme here.

[22] The source data for this plot can be found at http://cls.ru.nl/staff/hvhalteren/VanHalteren_Oostdijk_TNTL2015_Data.csv.

6 Distributions of specifi word types

With the use of our U-score measurement, we can now investigate how various groups of words are being used by the Twitter authors. We base most groupings on their word class (part of speech), as listed in the *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al. 1997). For several parts of speech, such as prepositions, the ANS does not provide an exhaustive list; in such cases we restrict ourselves to the listed words. In addition, we concentrate on standard modern Dutch, and leave out archaic forms, regional variants and pronunciation-inspired variants (e.g. clitic forms).

As we were not able to perform actual part-of-speech tagging on the tweet collection, the unit of examination is the word form, with all its potential uses taken together. If a word is listed under more than one part of speech, e.g. *zijn* ('to be') as both possessive pronoun and as auxiliary, copula and lexical verb, we include it under all appropriate parts of speech. Only in those cases where the attribution of a word to a specific word class is expected to be rare, and hence is unlikely to contribute to the measurements in any significant way, will we ignore it.

6.1 Grounding of the U-score

In order to check that the U-score does indeed represent the general applicability of words, we examine the scores for some words which we expect to find at the extreme ends of the scale. For words with a very high U-score, we look to the words with a predominantly grammatical function and the least contribution to content. The two most obvious candidates are the infinitive marker *te* ('to') and the existential marker *er* ('there').[23] All other words carry at least some content, but we judge the most grammatically oriented words to be the articles and coordinating conjunctions. By contrast, we expect to find a very low value for the U-score for words (with and without a hashtag) relating to politics as these words are highly specialized and can be used only with specific topics. Thus we examined the distribution of the names of political parties and their leaders.

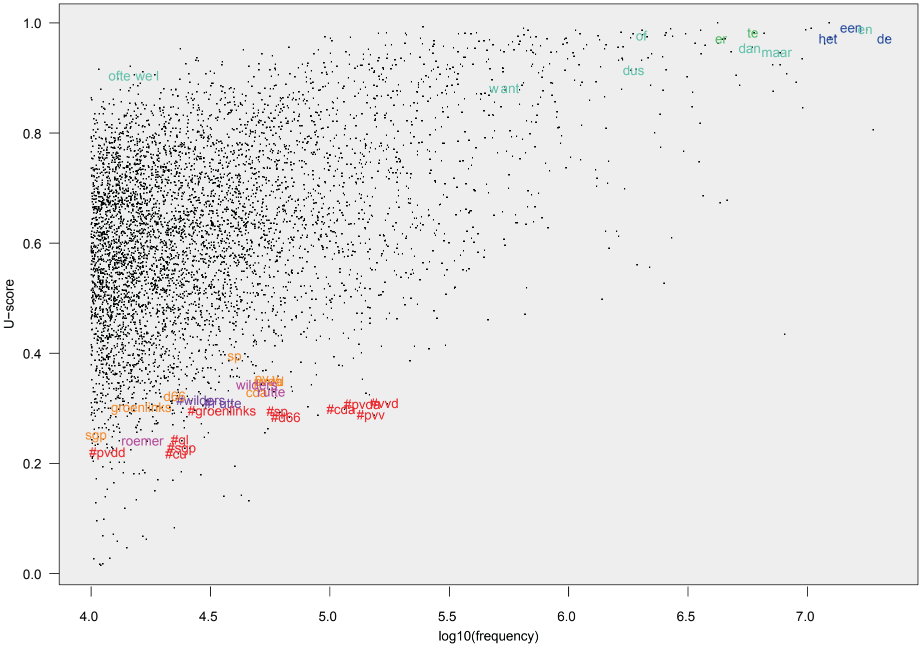
Figure 4 shows the placement of these words in the cloud (cf. Fig. 3). Both *te* and *er* (plotted in green), as well the articles *de* ('the'), *het* ('the') and *een* ('a') (dark blue) are found in the top right corner, which is where words are situated that are clearly generally applicable as well as very frequent. The coordinating conjunctions (light blue) are also all high up as we would expect them to be, even the more elaborate form of *of* ('or'): *oftewel*.

All words related to politics show up with very low U-scores. The relative placements of various groups of words also appear appropriate. The versions of the names marked with a hashtag have a lower U-score than those without. The smaller political parties (*cu*, *sgp*, *pvdd*) are placed lower than the major parties (*cda*, *d66*, *pvda*, *pvv*, *sp*, *vvd*).[24] The names of the party leaders *wilders* and *rutte*

[23] Both of these can also be used in ways in which they do carry content, but such multi-applicability should only increase their U-score further.

[24] With a strange phenomenon happening for Groen Links, which has two hashtags, and where the hashtag #*groenlinks* is found among the major parties, but the hashtag #*gl* among the smaller ones.

Fig. 4 Frequencies and U-score measurements for words expected to be either very widely applicable or very restricted in their use



are at the same level as their respective parties *pvv* and *vvd*, both for the versions with hashtags and those without.[25]

All in all, the U-score values appear to correlate well to our intuitions about general applicability. As a result, we feel confident that the U-score can be used successfully to investigate the general applicability of words other than these extreme ones.

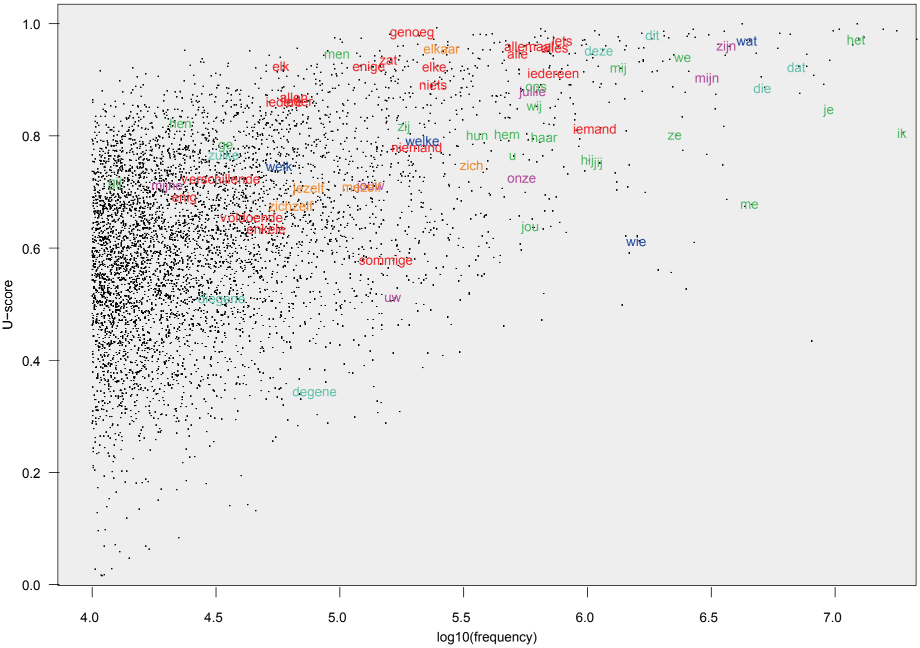
6.2 Pronouns

Another word class that is generally placed among the function words is that of the pronouns. However, when we examine their U-scores (Fig. 5), we see a tendency towards higher U-score values, but a much less sharp grouping than we saw for the fully grammatically oriented words in the previous section (Fig. 4).

Especially the personal pronouns (dark green) have much lower values than we expected beforehand. But on second thought, we must conclude that our expectations were unfounded. The first and second person personal pronouns (*ik* ('I'), *me* ('me'), *jij* ('you') and *jou* ('you')), are not grammatical glue at all, but are primarily references to the people participating in the conversation. The third person ones can either be anaphora for persons introduced earlier, or also references to specific people known by the participants. In the context of many Twitter conversa-

[25] This is not true for the party leader *roemer* and his party *sp*. There is no apparent reason for this exception.

Fig. 5 Frequencies and U-score measurements for the pronouns



tions, the latter is often more likely. The same can be said for the possessive pronouns (pink). Here, we must also comment on the two most extreme U-scores. The highest value, for *zijn* ('his'), is most likely enhanced by the fact that *zijn* is also the infinitive and plural present tense of the verb 'to be'. The lowest value, for *uw* ('your'), is probably so low because this is the polite form of 'your', which on Twitter is rather restricted to specific user groups and hence hashtags. The reflexive pronouns (orange) are also somewhat restricted in their use, probably because they can only be used with specific verbs. Only the reciprocal pronoun *elkaar* ('each other', also orange) has a very high U-score.

The demonstrative pronouns (light blue) show two groups and a singleton. The main group shows high U-scores, as expected, in several cases supported by the fact that they can also be used as relative pronouns. The singleton *zulke* ('such') takes an intermediate position. The reason for this is unclear, but the word does contain a component indicating comparison, which might well limit its applicability. We find very low scores for *degene* ('the one') and *diegene* ('that one'). This is caused by the fact that there is a hashtag *#degeneonderdezetweet* ('the one under this tweet'), that always contains *degene* or *diegene*, which causes a severe disturbance in the distribution measurement. The interrogative/relative pronouns (dark blue) appear to have lower scores than expected, apart from *wat* ('what'), which can also be used as an indefinite pronoun ('something'). The low scores might be caused by the fact that relative pronouns need more complex sentences than can be found in most tweets, and that interrogative sentences are also only used with a limited number of Twitter topics.

The final group here is formed by the indefinite pronouns (red). These are mostly found with high U-scores. The lower values for words such as *sommige* ('some') and *enkele* ('a few') seem to point to less trendy (possibly old-fashioned?) language use.[26]

Our investigation of the pronouns alone, a class almost always proclaimed to comprise function words, already supports the view that the distinction function-content words takes the form of a continuum. The range of U-scores for the pronouns stretches so far down that the classification of the whole class of pronouns as function words becomes untenable.[27] Transferring some subclasses of pronouns to the content words does not seem a viable solution either, as all traditional subclasses show a substantial spread in U-scores. Alternatively the whole class of pronouns might be moved to the content words, and the very high U-scores of some words in the plot explained by pointing out that these high scores are due to the ambiguity of the words in question.[28] However, considering the number of high-scoring pronouns, this is at best a questionable approach.

6.3 Prepositions and subordinating conjunctions

Prepositions and subordinating conjunctions have sometimes been put in a kind of intermediate position between function and content words:

> … some words share characteristics of both word classes. Prepositions are an example. We haven't added any new prepositions to the language in several hundred years, so in that sense prepositions form a closed class. And although their primary function seems to express information about the direction, location, and such of a following noun in English (near/on the table), many prepositions have quite complex, 'contentful' meanings. … So, although we will assume here that prepositions are function words, note that this distinction is not entirely clear and that prepositions present an interesting case of possible overlap between function and content word classes (Denham & Lobeck 2009: 144-5).

> For instance, while the conjunctions if, when, as, and because are unequivocally function words, this is less clear for words like suppose, provided that, granted that, assuming that. And while the adpositions in, on, of, at are clearly function words, this is less clear for concerning, considering, in view of. In the case of adpositions, linguists sometimes say that there are two subclasses, 'functional adpositions' and 'content adpositions', analogous to the distinction between content verbs and function verbs (=auxiliaries) (Haspelmath 2001: 16539).

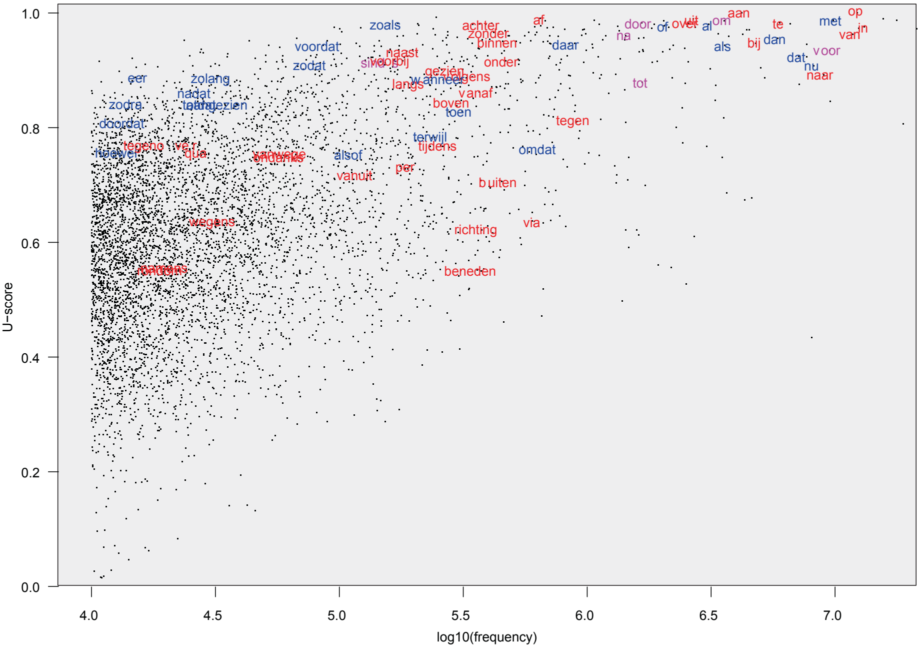
This characterization is justified if we examine the U-scores of these classes (Fig. 6). Prepositions are plotted in red, subordinating conjunctions in blue, and words that have interpretations in both classes in pink.

[26] This notion of a low number of something is much more often expressed with the alternative *paar* (lit.: 'pair', i.e. 'few'), for which we do see a very high U-score of around 0.98.

[27] At least as long as the lack of reference to an extra-linguistic world is the main indicator for a word being a function word, as opposed to, e.g. closedness of the class, which is not in question for pronouns.

[28] A high U-score represents a quite general applicability. This cannot only be caused by a word being mostly grammatical, and therefore not restricted in topic, but also by the word being very ambiguous and being very widely applicable because the various senses cover many topics.

Fig. 6 Frequencies and U-score measurements for prepositions and subordinating conjunctions



The prepositions are for the most part found in the upper ranges of the plot and contain some of the highest U-scores measured. It must be observed, though, that some of these forms can also occur as adverbs or as split-off parts of separable verbs, which obviously contributes to their wide applicability. However, this is also true for many of the lower scoring forms. Among these lower scoring forms, we find many specialized spatial and temporal prepositions, such as *beneden* ('below'), *richting* ('in the direction of'), *via* ('via'), *buiten* ('outside'), *tijdens* ('while') and *rondom* ('around'). Another low scoring group is related to third persons' relations to a described event, such as *namens* ('on behalf of') and *wegens* ('because of').

The subordinating conjunctions are generally higher up on the scale. Still, here too we see a spread that confirms the suggestion that there is room for a subclassification within the subordinating conjunctions (cf. Haspelmath 2001). However, for both prepositions and subordinating conjunctions, we do not see a clear separation between high scoring and low scoring words, but a more or less unbroken continuum. This would indicate that, here too, the proposed subdivision is not a solution. As for pronouns, adherence to a dichotomy implies classification of all prepositions and subordinating conjunctions as content words. And again high scores would have to be explained by ambiguity rather than by word type.[29]

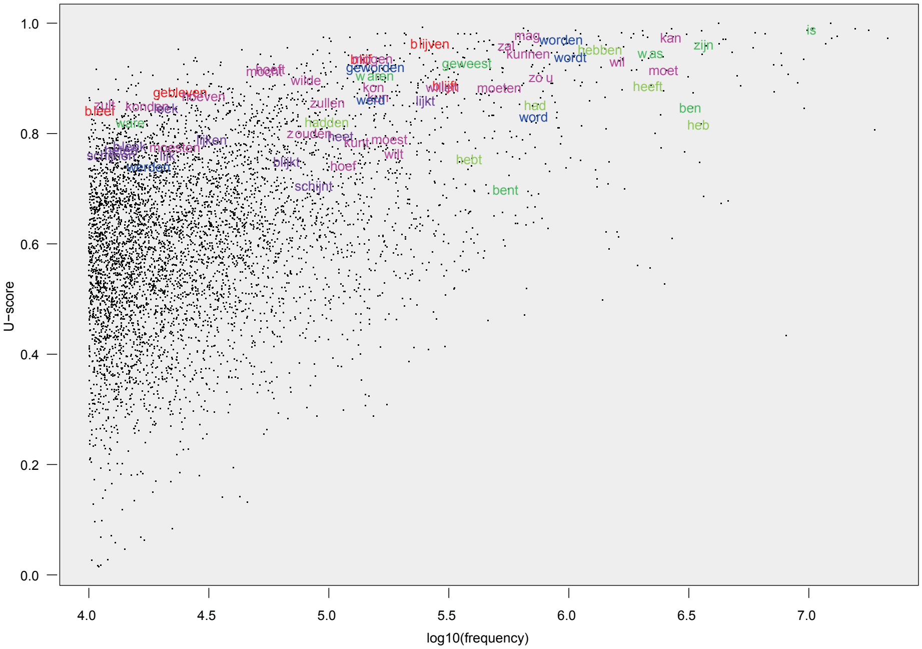
[29] Seeing which words do score very high, this explanation presents itself rather strongly and has more value than with the pronouns.

6.4 *Auxiliary verbs and copulas*

For the verbs, which as an overall group are usually assigned to the content words, it has already often been argued that some subgroups are rather function words, as was already mentioned in the above quote (Haspelmath 2001). Likely candidates for classification as function words are auxiliaries and copulas.[30] We examined the verbs that are listed explicitly by the ANS as auxiliary verb or copula.[31] The measurements for these subgroups are shown in Figure 7.

Starting with the verb *zijn* ('to be', dark green), which can be used as an auxiliary with past participles (both for pluperfect and for passive), we see mostly the same situation as for the personal pronouns: a mostly high U-score, except for first and second person singular (*ben* and *bent*). In addition, the form for the irrealis mood, *ware*, has a slightly lower U-score.[32] The verb *hebben* ('to have', light green), auxiliary for the pluperfect as well as lexical verb, shows the same pattern,

Fig. 7 Frequencies and U-score measurements for auxiliary and copula verbs



[30] In Dutch many verbs can be used both as an auxiliary verb and as a copula. Furthermore, some of these can also function as lexical verbs.

[31] There are many more verbs that can function in the verb cluster of a Dutch sentence as so-called 'grouping verbs' rather than as lexical verbs, but we wanted to restrict ourselves here to the clearest cases. Note also that, from the verbs listed explicitly as auxiliary or lexical verb, we exclude a few potential auxiliaries (e.g. *voorkomen* ('appear to someone')) whose use is dominated by lexical rather than auxiliary occurrences.

[32] The score is actually higher than might be expected given the unusual mood it is connected too. However, the word *ware* is also an adjective meaning 'true' and a noun meaning 'someone's true love'.

with first and second person forms here being *heb* and *hebt*. The second auxiliary for the passive, *worden* (dark blue), shows a less pronounced lower U-score for the first/second person, *word*, but appears to have restrictions on the plural past tense, *werden*.

We next move to a group of pure modal auxiliaries (pink), namely *hoeven* ('have to' in a negation context), *moeten* ('have to'), *mogen* ('be allowed to'), *kunnen* ('can'), *willen* ('want to') and *zullen* ('will'). Again we see high U-scores overall, with some lower ones for first/second person singular, such as *hoef* ('have to'), *wilt* ('want'), and *kunt* ('can'), and for plural past tense, such as *moesten* ('had to') and *zouden* ('would'), but here also the singular past tense *moest* ('had to').

The last group (purple) consists of modal auxiliaries that can also be used as copula (purple), namely *blijken* ('prove to'), *lijken* ('appear to'), *schijnen* ('appear to'), and *heten* ('be called'). Even though these words can be used in two different grammatical constructions, we still see U-scores that overall are lower than the U-scores for the other auxiliaries.

The observations for auxiliaries and copulas are not as supportive for the continuum view of the function-content word distinction as the observations in the previous subsections. The measurements for the auxiliary/copula group appear to show that an assignment to the content words, together with all verbs, is justified. The other groups do seem to allow a classification as function words, although this would necessitate a more thorough examination of what is going on with the first and second person forms.[33]

6.5 A function-content continuum

In the previous subsections, we have observed that many words that are traditionally seen as function words, or have been suggested later as possible additions, show U-scores that are incompatible with their function word status (given the distinction criterion of lack of content).[34] On the basis of these observations, we pose that the defence of a dichotomy between function words and content words becomes untenable, not only at the level of word classes, but even at the level of individual words (i.e. lemmas). Therefore, we support the view that each individual word is positioned on a continuum ranging from fully grammatical in nature to fully content-bearing. In fact, it is not so much the individual word that should be positioned, but rather the combination of the word sense and the word form. That the form is of consequence has become clear with the first and second person auxiliaries. We have, in our measurements, not been able to take the sense into account, but we are convinced that different senses will be positioned on different positions in the continuum. Note, by the way, that the function-content continuum is certainly not the same as our measured U-score, which refers to width of use rather than to content born. Surely, deriving values for the various words will be an interesting task.

[33] A classification per word form instead of per lemma seems by itself undesirable, but the differences here between different word forms belonging to the same lemma should be cause for reflection.

[34] The criterion of the openness of the class is not under discussion here, but is clearly different from the criterion of lack of content.

In the remainder of this subsection, we will attempt to derive some more information about our proposed continuum. We will not discuss the words with low U-scores, as we expect that these can safely be characterized as (almost) totally content-bearing. Nor will we examine the middle ranges, as positioning there will as yet be too subtle. Instead we will look at the unexpected, and examine words that were traditionally seen as content words but that now turned out to have high U-scores. For this, we selected the words with a U-score higher than 0.9, that is 243 words out of the 5,520 words that we measured (4.4%). Of these, 78 were already covered in the examination above. For the other words, we will discuss some more or less coherent groups.

The largest group (about 50 words) is that of the adverbs. This should not come as a surprise. After all, in classifications of part of speech, the class of adverbs is often a rest category. As such it contains words along the whole function-content continuum, and therefore also ones that are mostly grammatical in nature. In our top-243 we first find some expected groups, such as intensifying adverbs, e.g. as *heel* ('very'), *nog* ('yet'), *erg* ('very'), and *best* ('quite'), the negative adverb *niet* ('not') and members of the group that are sometimes referred to as pronouns, such as *hier* ('here'), *waar* ('where'), *ergens* ('somewhere'), and *overal* ('everywhere'). Then, we see a number of words used for hedging, such as *misschien* ('maybe'), *bijna* ('almost'), *waarschijnlijk* ('probably'), and *ongeveer* ('about'). A final more or less coherent group is formed by various adverbs relating to time, such as *eerst* ('first'), *meteen* ('right away'), *later* ('later on'), *steeds* ('all the time') and *eerder* ('earlier').

The relation to time can also be observed for various words of other word classes (in total about 25 time-related words). For nouns, we see *tijd* ('time') itself, *begin* ('beginning', also verb), *dag* ('day'), *moment* ('moment'), and *eind* ('end'). A clearly verbal form of *begin* is also present in the present tense *begint*. Then there are the adjectives *laat* ('late'),[35] *late* ('late'), *lang* ('long'), *kort* ('short') and *snel* ('fast'). And, finally, we find two more prepositions that were not in the example list in the ANS, namely *geleden* ('ago') and *rond* ('around'). The high U-scores for time-related words, though, do not stem from a mostly grammatical role. For, obviously, they do refer to an extra-linguistic reality. However, the aspect to which they refer, time, is on a different dimension than the one covered by our split into topics. There will be very few topics indeed in which time does not play a role, and this is even more true for the typical topics discussed on Twitter. In this case, then, the relation between general applicability and lack of content does not hold, which also implies that using the U-score for a first estimate of the position on the function-content continuum should only be done with extreme caution.

A next large group is that of the numerals and related words (about 35 words). For the cardinal numerals, we mostly find numerical forms, such as *1*, *2*, and *10*, but also the lexical form *twee* ('two').[36] In addition, there are quantifiers like *geen* ('no'), *meer* ('more'), *paar* ('pair'), *veel* ('many'), *zoveel* ('so many'), *weinig*

[35] The form *laat* can also be used as a tense of the grouping verb *laten* ('let'), but that use is less frequent than the adjective.

[36] The lexical form for the number one was already covered as the numeral in question is homonymous with the article *een*.

('few'), and *hoop* ('a lot'). For the ordinal numerals, we find *2e* ('2nd') and *eerste* ('first'), and related words like *andere* ('other'), *volgende* ('next'), *laatste* ('last'), and *vorige* ('previous'). Also numerical in nature are *keer* ('times') and *maal* (also 'times'). As was the case for the time-related words, this group too covers a special dimension in the content space. Counting and ordering/ranking can also be applied in the context of almost any topic. Here, however, we would assume a position at the grammatical end of the continuum, just as these words are sometimes already assigned to the function words, either as a category (e.g. van der Sijs 2002) or in their role as determiners (e.g. Manning & Schütze 1999).[37]

Less numerous than we expected are spelling variants of frequent function words. In the top-243, they are limited to *'n* and *n* for *een* ('a'), *'t* and *t* for *het* ('it'), and *nr* for *naar* ('to').[38] Apparently, the use of spelling variants itself is linked to a restriction in discussion topics. This implies that even spelling variants might have to be treated as separate word forms when assigning a position on the function-content continuum.

In the top-243, we find six nouns that do not belong to previously mentioned groups. The highest U-score, 0.97, is found for *kant*. Apart from a few observations of the rare sense 'lace', we mostly see the sense 'side'. About 25% of the observations are compositional constructions referring to a side of something, and about 45% are fixed combinations referring to a physical side such as the side of the road or movement to the right side. The remaining observations are lexicalized constructions, such as *aan de andere kant* ('on the other hand'), *mijn sterkste kant* ('my forte'), or *kant en klaar* ('ready'). Here then, the high U-score is a combination of a very widely useable spatial sense and fixed expressions. The importance of expressions is even stronger for *idee* ('idea', U-score 0.94), where only about 35% of the observations are compositional constructions referring to actual ideas. About 25% are instances of *geen idee* (lit.: 'no idea', i.e. 'I really don't know'), 15% *heeft iemand een idee?* (lit.: 'does anyone have an idea?', i.e. 'who can help?'), 10% *het idee hebben dat* (lit.: 'have the idea that', i.e. 'think that'), and the remainder a large group of fixed expressions, such as *het gaat om het idee* (lit.: 'it is about the idea', i.e. 'it is the thought that counts') and *mijn idee!* (lit.: 'my idea!', i.e. 'you're right'). Next is *soort* ('kind', U-score 0.92), where we find less than 5% *soort* used for species of animal. All other observations concern expressions with a similar type of meaning in English, such as 'that sort of people' (50%) or generally 'sort of' (45%), which fits in with the group of hedging words mentioned above. *kop* (U-score 0.92) is used in its literal sense slightly more than half the time ('head' 35%, 'cup' 15%, 'headline' 5%). The other meanings are more or less metaphoric (e.g. 'have a song in your head' 10%, 'in first place' 5%) or one of an impressively large group of completely lexicalized expressions, such as *hou je kop* (lit.: 'hold your head', i.e. 'shut up') or *de spijker op zijn kop slaan* (lit.: 'hit the nail on its head', i.e. 'state the exactly right thing'). The word *moeite*

[37] Note that this group is also a closed class. After all, its apparent openness stems from the fact that it is possible to build numerical words that have never before been used in sentences. However, the class is not so much open as well as infinite in size, as all meaningful numerical words can be predicted.

[38] The interpretation *naar* for *nr* is appropriate in about three quarters of the observations; in the remaining cases, it represents *nummer* ('number').

('effort') is special in that it has already almost lost its original sense for centuries, and occurs practically always in fixed expressions, such as *de moeite waard* (lit.: 'worth the effort', i.e. 'worth doing', about 25%) or *moeite hebben met iets* (lit.: 'have effort with something', i.e. 'find something difficult'). Less than 5% of the observations appear to be compositional constructions referring to effort. Finally, there is *gang*. Here we find about 33% compositional constructions, in which *gang* has various meanings, namely 'corridor' (25%), 'gang' (5%), and 'course of a meal' (2%).[39] All fixed expressions refer to yet another sense of *gang*, related to movement. Frequent examples are *aan de gang* ('in progress', almost 25%) and *op gang komen* ('build up speed', 20% when we include other verbs as well). All in all, we can conclude that a primary factor in the high U-scores of these nouns is their metaphorical use and/or presence in fixed expressions. This means that, in the inventory of senses of words for the purpose of determining their position on the function-content continuum, fixed expressions should be treated separately. Looking at the examples given, we can also conclude that, if we would take only the observations where words are being used in their literal sense, these words would certainly end up with lower U-scores, demonstrating that the various senses of words (by themselves or in combinations) can assume greatly varying positions in the continuum.

Finally, we look at the verbs in the top-243, of which there are 36 that are not already mentioned above. Most widely usable appears to be *komen* (lit.: 'come'), with the infinitive and plural present tense *komen,* as well as the third person singular present tense *komt* having U-scores of 0.98. Now, the concept of coming is already quite widely usable on Twitter, but even so, this literal meaning is found for only about 45% (*komen*) and 40% (*komt*) of the observations. The grouping verb interpretation (see above; cf. semi-auxiliary) is found for about 5% of the observations of *komen* and 15% of *komt*. Then there are the observations where *komen* is the verbal part of a separable verb, such as *tegenkomen* ('meet') or *omkomen* ('die'), both forms accounting for about 25%. All remaining cases concern fixed expressions, such as *het komt doordat* ('it is caused by') and *in beeld komen* ('starting to come to attention'). For *nemen* (lit.: 'take'), also with a U-score of 0.98, the percentage of literal interpretations is only 15%, divided over three senses, 'take possession of', 'choose', and 'eat'. Here, the percentage of parts of separable verbs is about 35%, with examples like *meenemen* ('take along') and *opnemen* ('record', 'answer the phone'). For *nemen*, half the observations concern fixed expressions, such as *afscheid nemen* ('say goodbye') and *op sleeptouw nemen* (lit.: 'take onto a towing cable', i.e. 'help'). For the verb *laten* ('let'; U-score 0.98), roughly 95% of the observations ask for a grouping interpretation, e.g. *mobiel achter bed laten vallen* ('let mobile phone fall behind the bed') or *laten we naar kantoor gaan* ('let's go to the office'). Only a few observations show the literal interpretation, e.g. *laat mijn haar lang* ('leave my hair long'), separable verbs, e.g. *uitlaten* ('walk (e.g. a dog)'), or expressions, e.g. *in de steek laten* (lit.: 'leave in the fight', i.e. 'abandon'). Taking one more example at a slightly lower U-score (0.94), *leggen* ('put down'), we see that nearly 55% of the observations concern separ-

[39] Note that many Twitter users are still going to school, and therefore are doing things in corridors and are possibly forming gangs.

able verbs, mostly *uitleggen* ('explain') and *wegleggen* ('put away'). The literal interpretations account for only 30% of the observations, and about one quarter of these are actually a spelling variant for *liggen* ('lie down'). The rest are fixed expressions, such as *in de watten leggen* (lit.: 'lay in wadding', i.e. 'pamper') and *knopen leggen* ('make knots'). All in all, as none of the verbs are extremely ambiguous in their literal sense, we must conclude that their high U-scores are derived from either their status as grouping verb (*komen* and *laten*) or their role in many different separable verbs and fixed expressions. Now we have seen the effect of fixed expressions already for the nouns, but there the expression was generally contiguous around the noun. With the verbs, the other components of the expression can range more widely in the clause, and often the fixed expression as a whole will be discontinuous.

7 Conclusion

In many places in the literature, we find that a distinction is made between function words, being characterized as grammatical glue, and content words, being characterized as the words contributing most to the content. Some authors present this distinction as a pure dichotomy, and assign entire word classes to either of the two groups. However, it is unclear whether these authors actually have such strong views on the distinction, or whether this is merely a manner of presentation.[40] It is clear that a strict dichotomy is difficult to uphold, seeing that the exact assignment of word classes or individual words varies in the literature. It will therefore not come as a surprise that there exist alternative views of the function-content distinction, such as one in which the dichotomy is replaced by a continuum.

In this paper, we have measured how widely, i.e. with how many different topics/domains, words are being used on the social media platform Twitter. For this purpose we have developed the so-called U-score, a measure which can be computed for each word and which tells us something about the applicability of a word. The assumption is that function words ought to be generally applicable, whereas content words are restricted to specific topics/domains. If the notion of a dichotomy is correct, the measurements should show a clear division in scores for the two groups as well, whereas the presence of a continuum would also become visible in the form of more spread-out scores.

We found that most word classes which are traditionally placed under the function words show a substantial spread in U-scores, which confounds attempts at a clear division. Moreover, many words traditionally placed under the content words prove more widely usable than many words traditionally placed under function words. All in all, the U-score measurements support the view that words should be distributed along a continuum rather than that they are divided over two distinct classes. At the extreme ends of the scale, we find on the one end the fully grammatical words, while on the other end the fully content-bearing words will be positioned.

[40] Furthermore, a second much mentioned criterion is the closedness or openness of the class, which lends itself better for a strict dichotomy.

From the observations in Section 6, we can also derive a number of desiderata for the application of this scale. The unit of measurement should be a specific word form, used in a specific sense. When dealing with language use on social media, spelling variants should best be treated as separate word forms, as they show different behaviour. As to the senses of a word, multi-word units including the word in question should also be treated separately, and hence factored out from the measurements of the word by itself. In this, it should be noted that such multi-word units may well be discontinuous, as for example the components of separable verbs in Dutch.

Obviously, these suggestions give rise to a substantial amount of future work, of which we ourselves can only cover a fraction. We have already started an investigation into the use of fixed expressions in Dutch tweets. We are also considering the use of linguistic and extra-linguistic context for word sense disambiguation. Both of these would help us to derive U-scores for word form – word sense combinations instead of just word forms as was the case in this paper. Furthermore, we will be investigating other applications of the U-score introduced in this paper. As an example, we would like to examine how the U-score relates to the use of words as features in either authorship profiling on the basis of idiolect (preferably high U-scores) or topic recognition (preferably low U-scores). Similarly, the corpus presented here provides opportunities for further study.

Looking back at our investigation in this paper, we found that Twitter proved to be a very rich and valuable source of information about the language use by a representative sample of the native speakers of a language. The sheer number of different authors and of different discussion topics/domains have allowed us to measure something that would not have been possible with traditional corpora, nor would it have been possible with experiments other than the most ambitious crowd-sourcing ones. Therefore, we emphatically encourage the use of Twitter and other social media as an exciting new resource for linguistic research.

Appendix 1 Compilation of the Data Collection

We built our data collection of Dutch tweets for the current investigation on the basis of the TwiNL data collection. From this material, we selected tweets with a date stamp from January 1, 2011 to June 30, 2013, and containing at least one of the 1,000 most productive hashtags. Moreover, we filtered out tweets which we did not consider to be 'normal' original Dutch Twitter language use, e.g. tweets not in Dutch, retweeted,[41] produced by bots, or overly repetitive. As we expected to lose quite some data to filtering, we started by including the 1,100 most frequent hashtags, which were later reduced to the 1,000 which were the most frequent when considering only the remaining tweets.

[41] Note that we only excluded retweets that were explicitly marked as such.

1.1 Preprocessing

We tokenized all text samples with our own specialized tokenizer for tweets.[42] This tokenizer recognizes words, numbers and dates as other tokenizers do, but it is also able to recognize a wide variety of types abundant in social media texts. Thus the tokenizer is able to identify hashtags and Twitter user names to the extent that these conform to the conventions used in Twitter, i.e. the hash (#) resp. at sign (@) are followed by a series of letters, digits and underscores. Many (but certainly not all) URLs and email addresses are recognized as well.[43] Finally, as the use of capitalization and diacritics is found to be quite haphazard in tweets, the tokenizer strips all words of diacritics and converts them to lower case.

1.2 Filtering out foreign language tweets

As observed in Section 3, the TwiNL collection still contains some 2.5% of tweets in a foreign language. Given the automatic collection procedure, this is inevitable. The common Dutch words used as anchors sometimes also occur in other languages. Especially German tweets are often captured. This is due to shared common frequent words like *als* and *hier,* such as in[44]

> *#job #psychologe/in Psychologe als Teamleiter: Munchen, Bayern – hier finden Sie freie Stellen, die Arbeitgeber … <url>.*

An explanation for other erroneous captures is that the often short Dutch function words are prone to being used as abbreviations or shortened words in other languages, such as *op* in the tweet

> *—¡OP! —¿OPPA? —¡Oppa gangnam style!.* ┌(-_-)┘ └(-_-)┐ └(-_-)┘ —.

It is the language filters' task to stop such tweets from getting into the final collection, but finding the right balance between excluding as many non-Dutch tweets as possible and including as many Dutch tweets as possible is not easy. The makers of the TwiNL collection are aware of this, and have changed tactics a few times in order to address the problem (Tjong Kim Sang: personal communication). For older sections of the collection, only tweets clearly recognized as Dutch have been included. Later parts include more tweets, together with an indication of the language proposed by the filtering process. Where this is another language, such as **french** or **english**, or where this is UNKNOWN, we automatically excluded the tweet from our selection. However, there is also a marker **notdutch**, which we found with both foreign language tweets, e.g. English, and also Dutch language tweets containing multiple non-Dutch tokens. Only in the latter case would it be appropriate to include such tweets in our dataset. This would require that all tweets marked as **notdutch** be inspected somehow so as to determine whether the tweet can be considered to be a Dutch language tweet. As we found that frequently even tweets marked **dutch** appeared to be suspect, these also needed to be checked. This meant that we had to build an additional language filter for the tweets marked as **dutch** or **notdutch.**

[42] We intend to merge our tokenizer in the near future into the open source tokenizer Ucto (http://ilk.uvt.nl/ucto/).

[43] The tokenizer relies on clear markers for these, e.g. http, www, or domain names such as .nl and .eu. Assuming that any sequence including periods is likely to be a URL proves unwise, given that spacing between normal words in tweets is often irregular.

[44] The words *in* and *die* in the example are also shared common frequent words, but are not included in the TwiNL anchor list.

As we did not expect to be able to construct a significantly better language filter at the individual tweet level, we decided to address the problem at the user level. For each user in our dataset, we examined all tweets and collected the twenty most frequent words. These are most likely to be highly frequent function words, although they may well include other words, such as proper names.[45] We then determined how many of these were present in a Dutch word list provided by the OpenTaal project.[46] The percentage was checked against the average percentage of all users. Since this average is dependent on the size of the sample (which we measure in terms of the number of tokens; cf. Table 1) for a specific user, we calculated it separately for a number of size bands: 1 token, 2-3 tokens, 4-7 tokens, etc. up to 2,048 tokens or more. There is quite a bit of variation between the different bands in the proportion of words found to be listed in the OpenTaal word list. Thus, at the highest band (where the size of the sample is 2,048 tokens or more), we found a mean of 97.2% listed words in the top-twenty, with a standard deviation of 7.3%. In the band of 128-255 tokens the mean has gone down to 91.1% with a standard deviation of 13.2%. We decided to filter out those users whose listed word frequency was more than 2.5 standard deviations below the band mean. For the highest band, this means that at least about 79% of the top-twenty words should be Dutch words that were identified as such. Furthermore, even though they would pass the 2.5 standard deviation threshold because of the low word count, we decided to filter out also any users with less than 50% listed words in their top-twenty.

The initial data collection contained about 3.8 million users, producing a total of about 1.5 billion tokens. Of these, the language filter removed about 670,000 users (18%), collectively producing a total of about 85 million tokens (5.5%). Rejected users with large amounts of tweets tend to be professional Twitter feeds. These include many German ones, but also Dutch feeds where most content is in specialized formats rather than in the form of text, e.g. feeds from weather stations and bookmakers. Rejected users with lower amounts of tweets are mostly human foreign users, but also native speakers of Dutch with extremely low word counts. We do lose some Dutch native speakers in the range just above the threshold of 2.5 standard deviations. However, given our current goal, investigating regular Dutch Twitter language, we decided to remove them as we would rather err on the side of caution.[47]

1.3 Filtering out non-standard Dutch language use

The next step in the process was to filter out tweets that show signs of not being produced by the general public so that we avoid our statistics being compromised by (often large volumes of) text posted on specialized professional twitter feeds, created by bots or possibly also human marketeers. As with the language filter above, we decided to be cautious rather than permissive in this 'normality' filter. We took three measures (viz. type-token ratio, proportion of hapax legomena and non-zipfiness, see below) which can be used to characterize language use and removed all users whose measurement was 2.5 standard deviations lower or higher than the mean for the size band at hand, in other words, when the

[45] Punctuation, numbers, URLS, @-names, hashtags, (symbolic) emoticons and such are not counted here, but only words (with 'word' being defined as a sequence of letters, dashes and apostrophes, so that emoticons like *xxx* are included in the count).

[46] OpenTaal is a project directed by the Dutch Language Union which aims to make available for free (written) Dutch language resources for use in open source projects (e.g. OpenOffice.org). One of the resources is the OpenTaal word list that was compiled for use with, for example, spelling checkers and grammar checkers. The word list we used for the research described in this paper is version 2.10g. It includes some 350,000 word forms, including many frequently used abbreviations and common Dutch proper names. For more information see http://www.opentaal.org/opentaal.

[47] And, obviously, the sheer number of users involved does not allow for manual intervention.

Table 1 Characteristics of user tweet samples per size band*

band	*#users*	*TTRm*	*TTRsd*	*HPXm*	*HPXsd*	*NZPm*	*NZPsd*	*DUTm*	*DUTsd*
1	22,732	1.000	0.000	1.000	0.000	0.000	0.000	0.453	0.498
2-3	73,551	0.970	0.117	0.947	0.207	0.218	0.060	0.613	0.414
4-7	275,417	0.968	0.100	0.962	0.126	0.235	0.035	0.657	0.319
8-15	608,212	0.930	0.118	0.926	0.139	0.226	0.037	0.654	0.291
16-31	764,756	0.862	0.140	0.849	0.196	0.209	0.037	0.666	0.273
32-63	529,367	0.743	0.169	0.709	0.274	0.182	0.038	0.758	0.237
64-127	418,354	0.683	0.140	0.713	0.185	0.153	0.035	0.845	0.184
128-255	350,501	0.603	0.113	0.688	0.135	0.124	0.029	0.911	0.132
256-511	285,973	0.516	0.093	0.658	0.110	0.096	0.023	0.945	0.099
512-1023	208,876	0.431	0.077	0.633	0.096	0.072	0.019	0.963	0.078
1024-2047	130,822	0.354	0.068	0.610	0.094	0.054	0.016	0.970	0.070
2048+	106,808	0.256	0.078	0.565	0.127	0.038	0.017	0.972	0.073

* Mean (m) and standard deviations (sd) for type-token ratio (TTR), fraction of hapax legomena (HPX), non-zipfiness (NZP, see text for explanation) and fraction of recognized Dutch words among the twenty most frequent words (DUT).

user's z-score with respect to a measurement was higher than 2.5.[48] In addition, we looked at all measurements together, i.e. the three measures we just mentioned plus the percentage of recognized Dutch words in the user top-twenty. We removed users whose total deviation was judged to be too high (the sum of the absolute value of the z-scores higher than 6), even when each of the four individual measurements was below the threshold.

The type-token ratio (TTR) is defined as the number of different tokens in a text sample, divided by the total number of tokens. The TTR is mostly known from its use as a measure of vocabulary richness in authorship studies, but can also be used in other text classification tasks, such as genre recognition. As for recognizing non-normal language use, a very low ratio usually indicates that the text sample is overly repetitive. But a high ratio also points to language use which does not conform to what we may expect in a regular text, as users tend to reuse standard building blocks like function words quite often. In Table 1, we can see that a decent size text sample (over 2,000 words) has an average type-token ratio of about 0.25. For smaller samples, the measurement becomes less exact, as smaller samples will obviously also contain fewer different tokens. In general, the TTR is known to be sensitive to sample size, and various approaches have been proposed to deal with this sensitivity. We will refrain from using derived measures and stay with the well-known original TTR, but use the simple mechanism of dividing our samples into size bands in order to adjust for sample size.

Another measure which is known from text classification is the number of word types that occur only once in a given sample, the so-called *hapax legomena.* In the context of authorship studies, these have been linked to the use of rare, obsolete or innovative words, and also to the degree of use of productive morphology. For our current goal, the exact reasons why hapaxes occur are less important. We will focus on the fact that their number will normally occur in a certain range, and that very high or very low numbers signal an abnormal source. In Table 1, we see that the proportion of hapaxes tends to represent close to 50% of the tokens for large samples, which is in line with what we find in the literature on tradi-

[48] The z-score is the number of standard deviations the measurement differs from the mean. For normally distributed measurements, 95% of the data points are expected to have a z-score in the range -2 to 2. Even though linguistic data is often not distributed normally, the use of the z-score has still proven fruitful.

TABLE 2 Filtering statistics for three sample normality measures

measure	*#users rejected*	*total #words by these users*	*#users rejected by this measure, but not by the other two*	*total #words by these users*
TTR	75,644 (2.5%)	273M (20%)	10,234	59M
HPX	89,718 (2.9%)	238M (17%)	38,789	27M
NZP	54,740 (1.8%)	12M (1%)	26,449	3M
One or more of these three measures	155,609 (5%)	311M (22%)		

tional text types. As with TTR, corrections for sample size have been proposed, but again we will stay with the original and use sample size bands.

Both the type-token ratio and proportion of hapaxes are useful measures, but they are rather crude, in the sense that they represent only a single aspect of the type distribution curve for a text sample. In an attempt to model this distribution in a more detailed manner, we developed a measure on the basis of Zipf's Law (Zipf 1935), which more or less states that the frequency of any type in a text sample is inversely proportional to its rank in the frequency table. In other words, the rank times the frequency is taken to be a constant. Our measure, dubbed *non-zipfines* , represents the deviation from Zipf's Law and expresses the degree to which the observed curve for a sample (here all tweets produced by a user) differs from the curve predicted by Zipf's Law. First, for each type, we multiply its frequency by its rank, after which we normalize this for sample size by dividing it by the total number of tokens in the sample. Then we determine the apparent constant for the sample at hand, by taking the mean over all types. Next, we go back to the individual types and determine the deviation by measuring the distance between the actual (normalized) multiplication result and the result predicted on the basis of the constant. Finally, we average the deviation over all types. The result may be less intuitively interpretable than the other two measures, but Table 1 shows that non-zipfiness behaves regularly enough to be used as a measure for sample normality. As with the other two measures, the sample size influences the measurement. The general rule is that with higher number of tokens, the measurement becomes more exact, but also deviations will have less effect. This can also be seen in the standard deviations. The measure fails at the smallest sample sizes. When there is only one word, the prediction will obviously be always correct, and this likelihood of a good prediction stays high at two and three words. Again, we use sample size bands as a correction for sample size.

If we look at the various users' z-scores with regard to these three measurements, we find a Pearson correlation (Pearson 1895) of 0.79 between the type-token ratio and the proportion of hapaxes. The non-zipfiness has a correlation 0.71 with the type-token ratio, and of 0.41 with the proportion of hapaxes. Each measures non-normality in a different way, and this is also visible when we examine how many users are rejected by each method (cf. Table 2).

Of the 3.1 million users remaining in our data collection after the language filter, producing a total of 1.4 billion words, the normality filter removed about 156,000 users (5%), collectively producing a total of about 311 million tokens (22%). These removed users are on average much more productive than those removed on the basis of language. This is not surprising, as the main group here are Dutch news feeds, which show high productivity, but very low variation (i.e. low TTR and very few hapaxes).

The final step of the user filtering, where all users were removed for which the total of all four z-scores was higher than 6, led to the removal of another 17,261 users, collectively producing a total of about 16 million words. Upon closer examination, we see that these users are of the same type as the users that were removed in the previous step.

We next applied the same filtering methods to the collection of tweets per hashtag. Of the original 1,100 hashtags, #sportheadlines had already been discarded, as it was filled completely by a single user, @Sport_Heads, which was removed. Five were rejected on the basis of the language filter: #voetbalheadlines, #sporttweets_, #nicovideo, #gameinsight, and #lt. This was correct for hashtags with genuinely foreign tweets (e.g. the hashtag #nicovideo contained mainly Japanese tweets),[49] but could be debated for, for example, a hashtag like #voetbalheadlines which contained mostly Dutch tweets. However, this hashtag was rejected because it contained more proper names than normal words. Considering our goal for the filtering, we decided that we indeed preferred the removal of such hashtags. Another 45 hashtags were rejected on the basis of the normality filter. Of the remaining 1,049 hashtags, we selected the 1,000 with the highest remaining token counts.

Appendix II Calculation of the U-score

For each individual word, we calculated a U-score, ranging from 1 (present in practically equal relative amounts in each examined hashtag dataset) to 0 (present in only one examined hashtag dataset).

II.1 Step 1: Determining the frequency distribution

First, we created a virtual random sample of the word's observations.[50] For the example in Figure 2 above, the sample size was 100,000; below, we will vary this number. We based the sampling on relative frequencies per hashtag dataset, so that the samples would not be biased by the difference in size between the datasets. The virtual sampling was done by down- or up-scaling of the actually observed frequencies, depending on how many actual observations were present.[51]

We then used the number of observations of the word within each dataset in this virtual sample as the frequency of that word with the corresponding hashtag. We divided the various possible frequencies (from 0 to 100,000) into frequency bands, namely 0 (band 0), 1 (band 1), 2-3 (band 2), 4-7 (band 3), etc., each new band starting with the next power of 2. With a maximum frequency of 100,000, the highest frequency band is 17. Now, for each frequency band, we counted how many datasets there are in which the word has a frequency in that band. Examples of such frequency distributions are given in the main text (Fig. 2).

[49] These were harvested because the name of the character Zou is a homograph of the Dutch auxiliary verb *zou*. These tweets were marked as **nondutch**, but as stated above, we initially kept these in our set.

[50] We use the term observation instead of occurrence, since, as described in Footnote 15, a single occurrence in the original tweet collection of a tweet with multiple hahstags may lead to inclusions in multiple hashtag datasets,

[51] This is possible because we only used frequencies and not the underlying tweets.

11.2 Step 2: Selecting a reference point

Since the distribution for the most generally applicable words is clearly different from the theoretical extreme distribution in which a word has exactly the same relative frequency in each dataset (cf. Fig. 2 in the main text), we decided not to measure the U-score against this extreme distribution, but rather against the distributions of a number of prototypical generally applicable actual words. There is a need to use more than one such prototypical word, as we can expect some variability between these words as well. However, the selected words must be generally applicable beyond doubt. We therefore examined all words for which we had at least 100,000 observations in our data collection and took a sample of 100,000 of their observations. We then selected those words for which the sample contained at least 2 observations in each of the 1,000 datasets. As it turned out, there were 19 such words, viz. *aan* ('to'), *als* ('as'), *de* ('the'), *dit* ('this'), *een* ('a'), *en* ('and'), *goed* ('good'), *in* ('in'), *is* ('is'), *keer* ('times'), *met* ('with'), *niet ('not'), of* ('or'), *ook* ('also'), *op* ('on'), *tijd* ('time'), *van* ('of'), *wil* ('will'), *zeker* ('sure').

11.3 Step 3: Measuring distance to the reference point

Once we had determined our basis of prototypical distributions, we started measuring all words with at least 1,000 observations in our data collection. For each such word, we again generated a virtual distribution of 100,000 observations over the frequency bands. We then measured the similarity of this distribution to each of the 19 prototypical word distributions, using Jensen-Shannon Divergence, a popular method for comparing probability distributions (Lin 1991).[52] Next, we derived a single divergence score from the 19 results by taking the third lowest value.[53] However, this score has no intuitive meaning and varies if we change the size of the virtual distributions.[54] For this reason, we apply a normalization by taking the z-scores[55] with regard to the mean and standard deviations of all words observed at least 100,000 times.[56]

11.4 Step 4: Thresholding against the minimum number of required observations

The quality of the measurement is dependent on the original number of observations for each word: the lower the frequency of a word, the worse the approximation of the extrapolated virtual sample of 100,000 observations will be to a real sample of 100,000 observations.

[52] For the moment, we judged the Jensen-Shannon Divergence the best method for measuring similarity here. However, it has at least one disadvantage, namely that it treats the various frequency bands as nominal rather than ordinal variables. In future work, we may reassess the choice of similarity measure.

[53] Since there is variation in distribution even within the group of prototypical words, it is unlikely that a word is similar to all of them. Therefore, we do not want to take some kind of average over all prototypical words. On the other hand, a close distance to only a single prototypical word is not good enough in our view. Therefore, we decided to use the distance to the third closest prototypical word.

[54] The measurement, although varying in absolute numbers, proves stable at different virtual sample sizes. The correlation between the measurements for all words at sample size 1,000 and size 100,000 is as high as 0.997. We decided to use 100,000, as we expect the best resolution at this size.

[55] The z-score is the number of standard deviations the measurement differs from the mean. For normally distributed measurements, 95% of the data points are expected to have a z-score in the range -2 to 2. Even though linguistic data is often not distributed normally, the use of the z-score has still proven fruitful.

[56] These being the ones for which we can count on statistically reliable measurements.

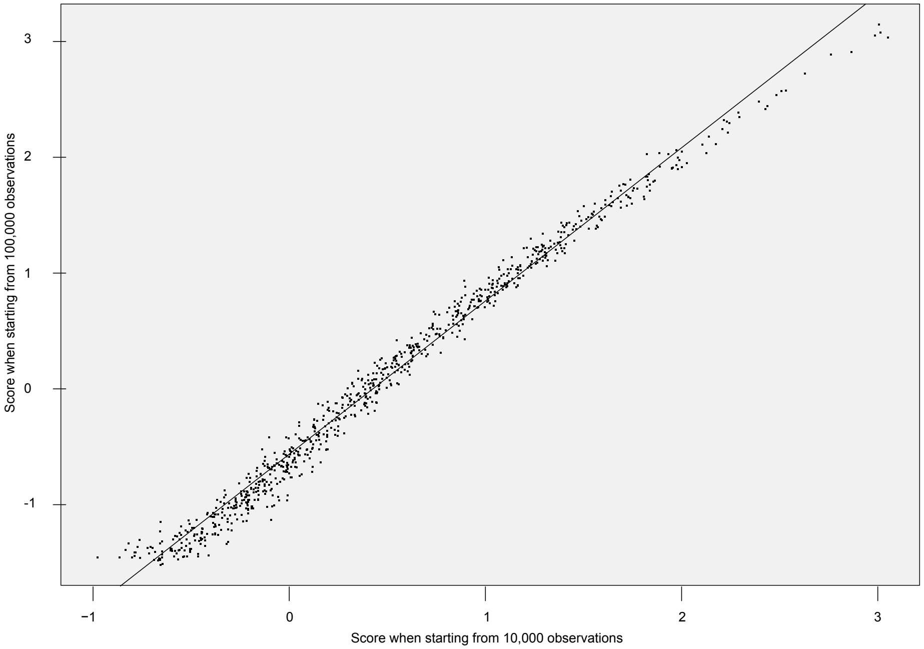
We estimated the speed at which the quality degrades by creating random subsamples of the samples for all 840 words for which we have at least 100,000 actual observations, this time bootstrapping on the observations themselves rather than using relative frequencies (Efron & Tibshirani 1993). We first created subsamples of 100,000, 30,000, 10,000, 3,000, and 1,000 observations. For these subsamples, we measured the abovementioned z-score, and compared this to the measured z-score when using all original observations. We found Pearson correlations of 0.999 (100,000), 0.997 (30,000), 0.992 (10,000), 0.959 (3,000), and 0.853 (1,000) respectively. On the basis of this measurement, we decided to continue only with those words for which we had at least 10,000 actual observations.

11.5 Step 5: Correcting for original sample size

Firgure 8 shows the relation between the z-score as measured when using subsamples of 10,000 and 100,000 words. Although the correlation is high (0.992), the fitted line shows that the values are not equal (slope 1.323, intercept -0.564; derived with linear regression, to be exact the function lm in R (R Development Core Team 2008)). Clearly, we will need to correct the z-score whenever the original sample has fewer than 100,000 words. To determine the correction parameters, we generated 10 subsamples at 10,000, 20,000, 30,000, …, 100,000, and again used lm to find the slope and intercept. We then fitted functions to these values with nls (non-linear least-squares estimate; again in R). Setting x at the log10 of the original sample size minus 3 (so that 10,000 maps to 1 and 100,000 to 2), we find that

- the slope value can be modeled with $x^{-2.6716} \bullet 0.3798 + 0.9428$
  (residual sum of squares 0.00012)

Fig. 8 The influence exerted on the measurement (in terms of z-score) by the number of observations on which the measurement is based



- the intercept value can be modeled with $x^{-2.439} \bullet -0.693 + 0.129$ (residual sum of squares 0.000054)

With these models, we corrected the z-score for each word, using the exact number of original observations.[57]

11.6 Step 6: Mapping to the desired range

The final step in the calculation was the transformation by means of which the measurements were fitted to the intended range of values. The most widely applicable words should receive the value 1, and the most restricted words the value 0. For the former, we used the most widely applicable words that we actually observed, namely *aan* and *op* with a corrected z-score of around -1.5, as we preferred actual words over an unattainable ideal here, and we could assume that truly ubiquitous words ought to be observed in a data collection as large as ours. For the latter, we could not assume the same, and we therefore calculated the corrected z-score that would be assigned to a word observed 10,000 times, with all observations with one and only one frequent hashtag. This score turned out to be around 5.2. So, for the final step, we applied a linear transformation, resulting in *aan* and *op* receiving a U-score of 1 and the virtual completely restricted word receiving a U-score of 0.

Appendix III Comparison of U-score with dispersion measures

As stated in the description of the U-score, we intended to benchmark word distributions not to a mathematical ideal distribution with equal relative frequencies in all samples, but to the known distributions of actual words which were clearly ubiquitous and which occurred very frequently in our data collection. This meant that existing word dispersion measures could not be used, as they tend to benchmark against exactly this mathematical ideal (Gries 2008).[58] Another reason for concern was that the relation between what we mean by *ubiquity* and what previous work apparently means by *dispersion* is not clear. Therefore, we implemented a new measure, aimed exactly at what we intended.

However, now the new measure has been developed, it is interesting to compare it to a few of the wider known existing measures for dispersion. On the basis of the discussion by Gries (2008), we selected three existing measures:

- Juilland et al.'s D (1971), which is a long-time well-appreciated measure designed specifically for word dispersion. It is calculated for a word by first determining the coefficient of variation *vc* (the standard deviation of its sample frequencies divided by the mean), dividing that by the square root of the number of samples minus 1, and subtracting the result from 1. The measure produces values in the range 0 to 1, with perfect dispersion represented by the value 1.

[57] These models are still a rather rough approximation, especially for the more extreme values, when the number of observations of a word approaches 10,000. For such words, we estimate that for 95% the calculated U-score will deviate less than 0.05; for the remaining 5%, this should not be more than 0.1.

[58] Some measures try to introduce additional sensitivity by modeling the distances between the occurrences of a word within each sample (Washtell 2007). However, as our data consists of unordered tweets rather than coherent text, distance-based measures cannot be applied.

- Gries' DP (2008), which he introduced to address problems encountered for many measures proposed earlier. It is calculated for a word by summing the differences between the percentage of the word's occurrences in every sample and the percentage of the total corpus size covered by that sample, and dividing the sum by 2 in order to let the results range from 0 to 1, with perfect dispersion represented by the value 0.
- Inverse document frequency (IDF; Spärck Jones 1972), which is much-used in natural language processing (NLP) circles. It is calculated for a word by dividing the total number of samples by the number of samples containing the word, and then scaling by taking a logarithm (normally base 2). The results range from 0 to infinity (in principle; for any given collection, the maximum is the log of the number of samples), with perfect dispersion represented by the value 0.

We implemented these measures and applied them to the full data we had at our disposal for the 5,520 words for which we determined the U-score.[59] Given D's sensitivity to differences in sample size, we used relative frequencies rather than absolute ones to calculate *vc*. DP is similarly affected, on purpose, by such differences; we therefore calculated DP twice, once using the actual different-size samples (DP$_{\text{abs}}$), and once using the same virtual equal-size samples we used for the U-score (DP$_{\text{rel}}$). All four measurements are included in a data file, which can be found at http://cls.ru.nl/staff/hvhalteren/VanHalteren_Oostdijk_TNTL2015_Data.csv.

TABLE 3 Correlation (Pearson's r) of the various dispersion measures on 5,520 measured words

	D	*DP$_{abs}$*	*DP$_{rel}$*	*IDF*
U-score	0.651	-0.967	-0.986	-0.601
D		-0.684	-0.712	-0.632
DP$_{\text{abs}}$			0.977	0.627
DP$_{\text{rel}}$				0.620

TABLE 4 Rank correlation (Spearman's ρ) of the various dispersion measures on 5,520 measured words

	D	*DP$_{abs}$*	*DP$_{rel}$*	*IDF*
U-score	0.813	-0.973	-0.986	-0.800
D		-0.836	-0.877	-0.681
DP$_{\text{abs}}$			0.978	0.806
DP$_{\text{rel}}$				0.828

The mutual correlations of the measures are shown in Tables 3 and 4. We did not correct for direction, so that some correlations are negative. As there may be non-linear relations between the measures, rank correlations might give a better indication. On the other hand, in our case the measures are not used for ranking, but to place a word at some point on a scale from 0 to 1, and the actual values therefore do matter.

[59] With 'full data' we mean all observations actually made, not down- or upscaled to virtual samples of e.g. 100,000 observations.

Fig. 9 Relation between the U-score and Juilland et al.'s D

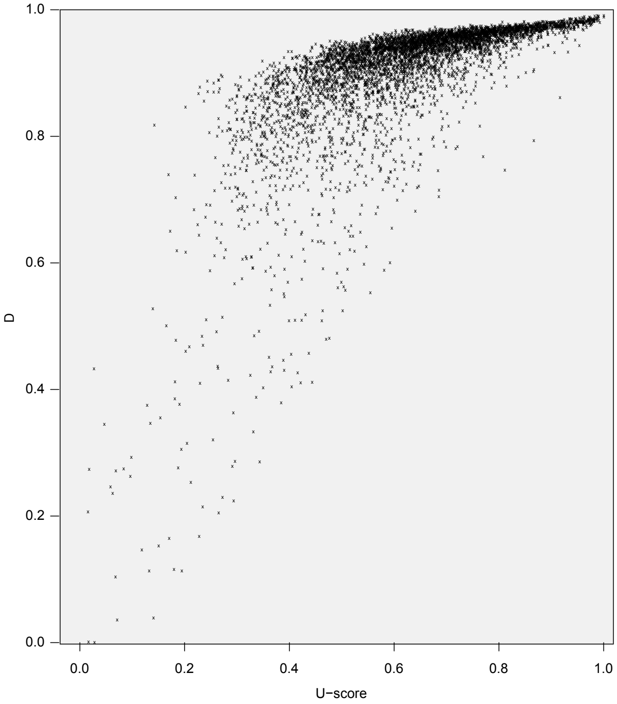
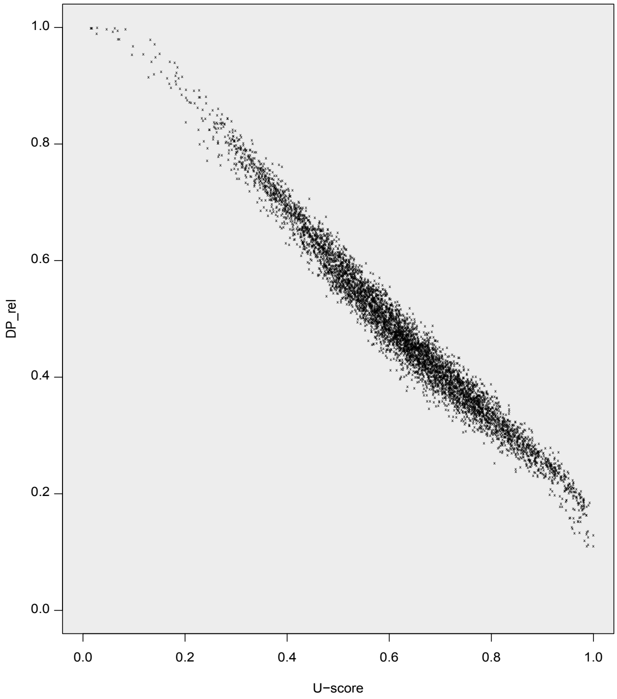


Fig. 10 The relation between the U-score and DP_{rel}



Given that IDF only measures presence in datasets, and not frequencies, it does surprisingly well, especially on the rank correlation, but clearly not well enough for our purpose. D is more sensitive, but also seems to be measuring something different than the U-score and DP. If we look at Figure 9, we see a better correlation for the more dispersed words, but ever larger discrepancies with the U-score as words get more concentrated. Both versions of DP are very close to the U-score, which is unexpected, seeing that DP does not take into account the shape of the distribution graph, but merely compares to a fully unbiased distribution. It would seem that the way in which the shape varies per word is restricted, so that the two measurements tend to lead to a similar ranking.

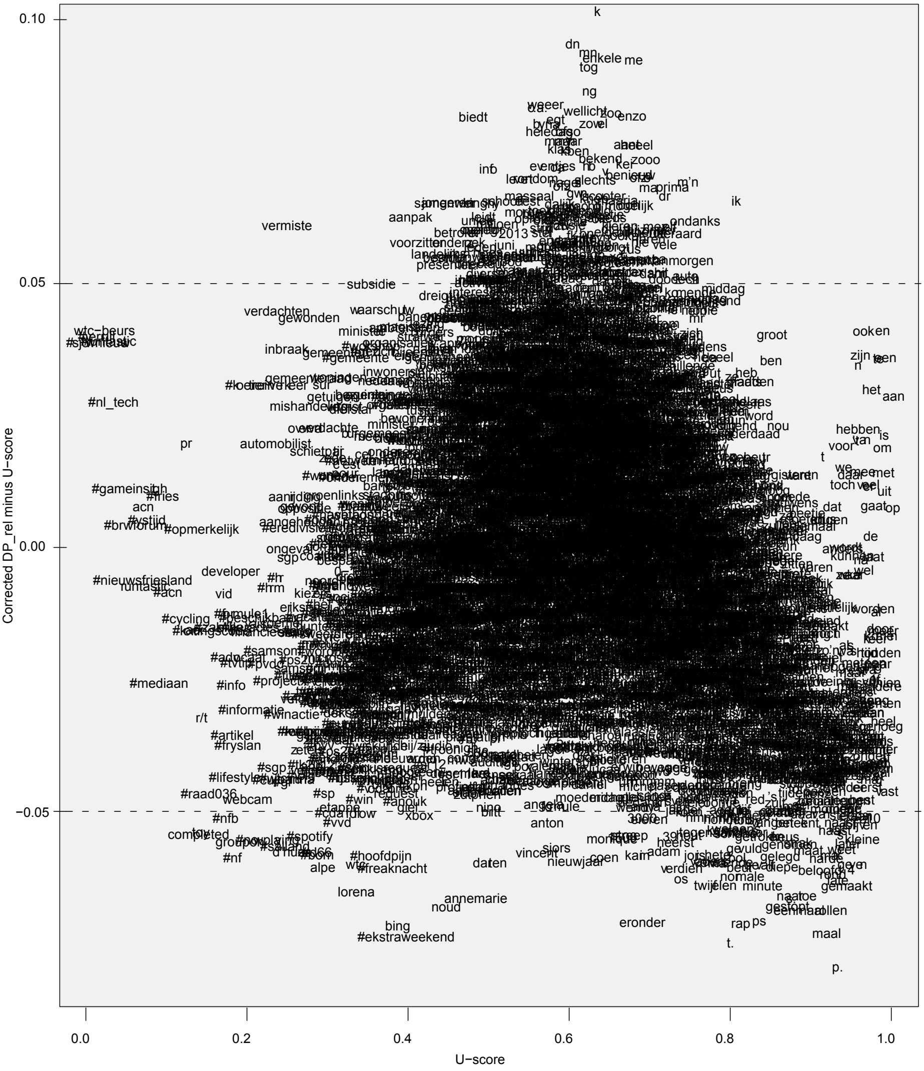
Also unexpected is the (relatively) low correlation between DP$_{\text{abs}}$ and DP$_{\text{rel}}$. Gries meant for DP to take into account the differences in sample sizes, but this leads to excessive effects in our dataset, where the differences are sometimes enormous (cf. Fig. 1). The word for which DP$_{\text{abs}}$ and DP$_{\text{rel}}$ differ most is *iemand* ('somebody'). The difference of +0.14 (DP$_{\text{abs}}$ higher than DP$_{\text{rel}}$) is caused completely by two highly frequent hashtags, *#dtv* with +0.08 and *#durftevragen* with +0.06 (both 'dare to ask'). We also see large differences for a number of first names. Why this might be, is exemplified by *mike*. Its difference of +0.11 is caused by its high frequency in *#gtst* (a soap opera; 5.5M words), leading to +0.14, which is counteracted by -0.03 from *#rtl7darts* (TV program; 260K words) and -0.01 from *#fcutrecht* (soccer club; 720K words). Clearly, the huge datasets have too much impact on DP$_{\text{abs}}$. Given that we want to give equal weight to each hashtag, DP$_{\text{rel}}$ is obviously the better choice for our experiment theoretically, and it is satisfactory to see that DP$_{\text{rel}}$ also correlates better with the U-score in actual practice.

With this correlation, DP$_{\text{rel}}$ might be considered as an acceptable approximation for the U-score, and in some cases preferable, because it is much easier to calculate. However, before accepting it as such, we should examine where the differences between DP$_{\text{rel}}$ and U-score lie. The original scores for DP$_{\text{rel}}$ and U-score are shown in Figure 10. This plot shows the difference in range and direction, but is not clear enough to examine details. For this, we first recast DP$_{\text{rel}}$ as an approximation of the U-score, by the formula DP$_{\text{relcor}}$ = -1.094•DP$_{\text{rel}}$ + 1.148, with the parameters again calculated with lm in R. For the result, we show the difference with the U-score as a function of the U-score (Fig. 11). For the majority of words, the difference is smaller than 0.05, which for our current purpose would be acceptable. However, there is also a substantial number of words for which the difference is larger.

The main reason for the larger differences can be found in the formula with which the two measures are calculated, the sum of differences in percentage for DP and the Jenson-Shannon distance between distributions for the U-score. For DP$_{\text{rel}}$, each sample percentage is compared to 0.001, which would be the expected percentage for perfect dispersion. As the frequency bands follow the powers of 2, each next band has double the percentage. For bands lower than 7, the penalty is therefore 1-2^{-n} times 0.001, with the special case of 0.001 at 0. Even for really low bands, the penalty can never exceed 0.001. For bands higher than 7, however, the penalty is 2^{n}-1 times 0.001, which increases fast. In general, DP considers mostly overuse of words in samples, and much less underuse in others, whereas the U-score considers the whole distribution curve equally.

The effect of the large penalties at higher bands can be seen for, e.g., the proper name *annemarie*, which we see in the center of the lower regions of Figure 11, meaning that the DP$_{\text{rel}}$ approximation of the U-score is too low. In our data, 63% of the occurrences of *annemarie* are found with the hashtag *#bzv* (connected the extremely popular reality show Boer zoekt Vrouw), meaning that this single spike alone leads to a DP$_{\text{rel}}$ of 0.26, while all other hashtags are practically irrelevant for the calculation. Now, if the goal is to measure dispersion, heavily penalizing high spikes is perfectly defendable. We, however, aim at a measure for ubiquity, and ubiquity is not ruled out by the presence of high frequency band spikes.

Fig. 11 The difference between DP$_{\text{relcor}}$ and the U-score, plotted against the U-score itself



In other words, for words which show very high concentrations for one or more hashtags, the DP$_{\text{rel}}$ is not a good approximation of the U-score.

The effect of smaller penalties in the lower bands, on the other hand, can be exemplified with the word *k* (shortened form of *ik*, 'I'), which is found at the very top in Figure 11, meaning that the DP$_{\text{rel}}$ approximation of the U-score is too high. Here we see that the U-score itself is mostly low because the word is penalized for occurring rarely or not at all with a substantial number of hashtags. The shape of the curve for bands up to 7 account for three times more penalty points that the bands from 7 on. DP$_{\text{rel}}$ on the other hand penalizes much less in the lower bands. Now, this region is crucial when considering ubiquity, and there is obviously a problem, but this problem is harder to pinpoint. The complication is that lower concentrations in some hashtags imply higher concentrations elsewhere, and apparently the underpenalization in the lower bands is compensated by penalization in the higher bands for most words, as the overall correlation between DP$_{rel}$ and the U-score is high. However, it is clear that DP$_{rel}$ leads to different results for quite a number of words in the center region of the U-score, such as *k*, where the compensation is not enough.[60]

In summary, we see reasonable to high correlations between the U-score and existing measures for dispersion, but even for the most correlated measure (DP$_{rel}$), there are clear differences with the U-score. Because of these differences, we feel that the U-score is more appropriate for our target of measurement, ubiquity, which we need to answer our research question. Furthermore, although we feel that ubiquity is not the same as dispersion, the U-score is clearly related to dispersion measures, and may also be useful in other research contexts where word distributions are of interest. A further discussion of this, however, is outside the scope of this paper.

Bibliography

Baayen 2001 – R.H. Baayen, *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers, 2001.

Denham & Lobeck 2009 – K. Denham & A. Lobeck, *Linguistics for Everyone: An Introduction*. Cengage Learning, 2009.

Efron & Tibshirani 1993 – B. Efron & T. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, 1993.

Elbers & Wijnen 1989 – L. Elbers & F. Wijnen, ‘Kennis, vaardigheid en “performance” in de taal/spraakontwikkeling: differentiatie van inhoudswoorden en functie-woorden’. In: J.F. Matter and J.M. van der Geest (eds.), *Lexicon en Taalverwerving, toegepaste taalwetenschap in artikelen 34.* Amsterdam: VU Uitgeverij, 1989.

Fagan & Gençay 2011 – S. Fagan & R. Gençay, An Introduction to Textual Econometrics. In A. Ullah and D. Giles (eds.), *Handbook of Empirical Economics and Finance*. Boca Raton: Chapman & Hall/CRC, 2011, p. 133-154.

van Gelderen 2005 – E. van Gelderen, Function Words. In: P. Strazny (ed.), *Encyclopedia of Linguistics*. New York: Fitzroy Dearborn, 2005.

Gries 2008 – S. Gries, ‘Dispersions and adjusted frequencies in corpora’. In: *International Journal of Corpus Linguistics* 13 (2011) 4, p. 403-437.

Haeseryn et al. 1997 – W. Haeseryn, K. Romijn, G. Geerts, J. de Rooij, & M.C. van den Toorn, *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk. Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn, 1997.

Haspelmath 2001 – M. Haspelmath, ‘Word classes/parts of speech’. In: P. Baltes & N. Smelser (eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Pergamon, 2001, p. 16538-16545.

Juilland et al. 1970 – D. Juilland, D. Brodin, & C. Davidovitch, *Frequency Dictionary of French Words*. The Hague-Paris: Mouton, 1970.

Lin 1991 – J. Lin, ‘Divergence measures based on the Shannon entropy’. In: *IEEE Transactions on Information Theory* 37 (1991) 1, p. 145-151.

Manning & Schütze 1999 – C. Manning & H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

Pearson 1895 – K. Pearson, ‘Notes on regression and inheritance in the case of two parents’. In: *Proceedings of the Royal Society of London* 58 (1895), p. 240-242.

Quirk et al. 1985 – R. Quirk, S. Greenbaum, G. Leech & J. Svartvik, *A Comprehensive Grammar of the English Language*. London: Longman, 1985.

[60] If, despite these differences, we would want to use DP$_{rel}$ as an alternative for the U-score, it too should be adjusted for the number of observations (cf. Step 5 in Appendix II). The deviation for words observed only 10,000 times is less pronounced than with the U-score (cf. Fig. 8), with a slope of only 1.113 and an intercept of -0.082, but is still significant and leads to a noticeable undervaluation of less dispersed words with lower frequencies.

R Development Core Team 2008 – R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008.

van der Sijs 2002 – N. van der Sijs, *Chronologisch woordenboek. De ouderdom en herkomst van onze woorden en betekenissen*. Tweede druk. Amsterdam/Antwerpen: Veen, 2002.

Spärck Jones 1972 – K. Spärck Jones, 'A statistical interpretation of term specificity and its application in retrieval'. In: *Journal of Documentation* 28 (1972), p. 11-21.

Tjong Kim Sang & van den Bosch 2013 – E. Tjong Kim Sang & A. van den Bosch, 'Dealing with Big Data: The case of Twitter'. In: *CLIN Journal* 3 (2013), p. 121-134.

Washtell 2007 – J. Washtell, *Co-Dispersion by Nearest Neighbour: Adapting a Spatial Statistic for the Development of Domain-Independent Language Tools and Metrics*. MSc Thesis, University of Leeds, 2007.

van Wijk & Kempen 1979 – C. van Wijk & G. Kempen, *Functiewoorden: een inventarisatie voor het Nederlands.* Nijmegen: Vakgroep Functieleer, Psychologisch Lab. Kath. Universiteit, 1979.

Zipf 1932 – G. Zipf, *Selective Studies and the Principles of Relative Frequency in Language*. Boston: Cambridge University Press, 1932.

Zipf 1935 – G. Zipf, *The Psycho-Biology of Language*. Boston: Houghton Mifflin, 1935.

Zipf 1949 – G. Zipf, *Human Behaviour and the Principle of the Least Effort. An Introduction to Human Ecology*. Reading MA: Addison-Wesley, 1949.

Adressen van de auteurs

CLS, Dept. of Linguistics
Radboud University Nijmegen
hvh@let.ru.nl
n.oostdijk@let.ru.nl