Hans van Halteren, Nelleke Oostdijk

# Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens

**Abstract**

In this paper, we attempt to estimate which proportion of the word tokens in Dutch tweets are not covered by standard resources and can therefore be expected to cause problems for standard NLP applications. We fully annotated and analysed a small pilot corpus. We also used the corpus to calibrate automatic estimation procedures for proportions of non-word tokens and of out-of-vocabulary words, after which we applied these procedures to about 2 billion Dutch tweets. We find that the proportion of possibly problematic tokens is so high (e.g. an estimate of 15% of the words being problematic in the full tweet collection, and the annotated sample with death-threat-related tweets showing problematic words in three out of four tweets) that any NLP application designed/created for standard Dutch can be expected to be seriously hampered in its processing. We suggest a few approaches to alleviate the problem, but none of them will solve the problem completely.

## 1 Introduction

With the advent of the social media, communication has changed drastically. Where before the public mass media (radio, television, newspapers) were dominated by a relatively small group of communication professionals, the social media have provided a platform for the masses through which they can voice their experiences and opinions and interact with other users. Nowadays the volumes of user-generated content that are produced on a day to day basis exceed by far whatever expectations existed when the first services were launched.[1] Access to the social media is unrestricted and generally services are widely and freely available. Communication is fast and as such extremely suitable for spontaneous, almost instant communication.

User-generated data have attracted the attention not only of linguists and communication experts, but also businesses, governmental and non-governmental organizations. The data are being exploited for a wide range of purposes, from linguistics and communication research to developing marketing strategies or evaluating policy issues. Linguists and communication experts have taken an interest particularly in the conversational data that are available from chats and discussions lists. Where the chat data are typically real time (private) conversations between a small number of people, with discussion lists the communication is usually more public but slower, and therefore often also more edited. In recent years, the interest for user-generated data has received an immense boost as Twitter was adopted by the masses. Twitter combines more instant communication and public availability which make it a very valuable source also for information mining.

One of the problems with user-generated data is that users take the liberty of expressing themselves as they see fit, without necessarily adhering to spelling conventions, grammati-

[1] In 2007, the fledgling Twitter boasted a meagre 5,000 tweets per day (Weil 2010). In 2013, this has grown to 500 million tweets per day (Krikorian 2013).

cal rules, etc. As a result we find that texts display a great deal of variability as regards typography, orthography, syntax, semantics, and discourse. The variability has several dimensions. The variability may be

- medium-related, that is, the medium may impose limitations on the length of the texts, while authors may also experience that they are under pressure of time (e.g. in chats) as their interlocutors may claim their turn prematurely while they are yet to finish. Moreover, there may be an effect of the (im)possibilities of the text entry mechanism used.
- author-related, that is, the language use of each author is characterized by its own idiosyncracies
- use-related, that is, depending on whether texts are used for professional or social use they adhere more or less to more widely accepted conventions or standards. Thus it appears that news feeds and government communications are quite conventional in the language they use, whereas texts exchanged between pals especially by sms or whatsapp but also on Twitter may be almost incomprehensible to people outside their peer group.

Clearly, it would be naive to think that all deviations from the norm set by the language used in the conventional media are errors. Some obviously are, but in many other cases the author made a deliberate choice to use some variant form.[2] As a result, we find that processing user-generated data is severely hampered as standard tools such as tokenizers, part-of-speech taggers, lemmatizers, morphological and syntactic parsers, and named entity recognizers cannot handle the variability very well. Han and Baldwin (2011: 369) report that they "found Twitter data to have an unsurprisingly long tail of OOV [out-of-vocabulary; HvH/NO] words, suggesting that conventional supervised learning will not perform well due to data sparsity. Additionally, many ill-formed words are ambiguous, and require context to disambiguate." Thus the variability is found to be prohibitive when it comes to successfully using applications such as text-to-speech systems (for example when wanting to have a text message read aloud), search and retrieval systems, and machine translation systems. In previous research we have found that in the n-gram based recognition of threatening tweets modeling spelling variation alone increased recall with a further 2.7-5.8% (Oostdijk and van Halteren 2013).

In this paper the research question we attempt to answer is: What proportion of the word tokens occurring in Dutch tweets is not covered in lexical resources designed/created for standard Dutch, either because tokens are not included at all or because tokens are (unrelated) homographs of the tokens listed there.[3] Related questions here are (a) is it possible to estimate the proportion automatically on the basis of a sample? and (b) to what extent does the proportion vary between various authors and topics? The question underlying our main

[2] See also the comprehensive study by Tagg (2009) who on the basis of a corpus of sms texts investigates the many different strategies that people use when the medium imposes severe restrictions on the length of a text and text entry is hindered by the entry device.

[3] Examples include clitics, such as *dak* (normally "roof") for *dat ik* ("that I") or *int* (normally "collects") for *in het* ("in the"), and the abbreviation *eik* (normally "oak") for *eindelijk* ("finally").

research question and the motivation for undertaking the research described here is that we want to know whether or not, and if so, to what extent the variability in Dutch tweets hinders automatic processing

Starting-point for our approach is the collection and manual annotation of a pilot corpus of tweets in which a small number of specific hashtags are represented. The pilot corpus is annotated for different types of token, both words and non-words, after which the annotated corpus can be used for a detailed investigation of the variability at this scale. Furthermore, we use the annotation as a benchmark for an automatic estimation procedure with which much larger amounts of tweets can be processed. After conformation of the validity, we apply the estimation procedure to a large – almost 2 billion tweets – collection of Dutch Tweets.

The structure of the paper is as follows. First, in Section 2 we introduce the pilot corpus that constitutes the experimental material that we use as a basis for obtaining estimates of the proportion of problematic tokens. Next, we describe the annotation of experimental material (Section 3). In Section 4, we discuss our findings as regards the different types of word and non-word tokens in the manually annotated data. The automatic estimation of the proportion of problematic cases is the topic of Section 5, while the estimates for the whole collection of tweets considered in this paper are given in Section 6 together with an analysis of our findings. We conclude this paper with a brief summary and our plans for future work.

## 2 Selection of experimental material

Most work on out-of-vocabulary words in tweets has been done on the basis of type lists (e.g. Han and Baldwin 2011; Sidarenka et al. 2013). However, such lists lack vital information. They do not show us how the words are distributed over the tweets, so that we cannot estimate which percentage of the tweets is affected, or whether there are differences between users and/or topics. They also do not show the context, so that we cannot know whether an in-vocabulary word is in fact known or whether the form is merely a homograph of another word in a lexicon used by some NLP application. For a proper investigation, we will have to look not at type lists, but at the underlying tweets. We will start with a modestly sized pilot corpus, so that a full manual annotation is feasible. For this corpus, we have selected ten hashtags, intended to provide a reasonable spread in topics and language use. Obviously, no exact predictions can be made about the language use for any topic, as even the most professional topics will occasionally attract emotional or humorous comments.

- **#aardbevingen (A)**[4] Tweets carrying this hashtag discuss earthquakes in the province of Groningen, that are caused by extracting gas from below the surface. The tweets are expected to be mostly official communiqués and attacks by interest groups, and therefore rather clean language.
- **#doodsbedreiging (D)** Tweets with this hashtag contain (death) threats and reactions to them. They tend to be very emotional and regularly contain street language.
- **#file (F)** These are tweets concerning traffic jams, generally people reporting on new traffic jams and their reaction. The level of emotion is less high than might be expec-

[4] We will be using single letter abbreviations for the various hash tags in tables and figures.

ted, possibly because one is used to being in traffic jams. A special type of token here are the various names of cities and roads.

- **#houdoe (H)** This hashtag does not refer to any specific topic, as *houdoe* is a dialect word for goodbye. We included this hashtag in order to find uses of dialect in tweets.[5]
- **#irri (I)** This hashtag like #houdoe, #jaloers and #omg does not refer to a spefic topic. *irri* is a short form for Dutch *irritant* (English: irritating). Given this hashtag, we expect high emotion levels with concomitant effects on language use.
- **#jaloers (J)** With *jaloers* meaning 'jealous', we again expect some emotion, although less strong than with *#irri*.
- **#miljoenenjacht (M)** Tweets carrying this hashtag relate to a Dutch tv game show. These tweets are expected to be mostly from the average user rather than (semi-) professional authors.
- **#ns (N)** *NS* is short for *Nederlandse Spoorwegen* (Dutch Rail). Tweets with this hashtag will regularly contain official communiqués, but also quite a lot of train traveler reports and, sometimes vehement, complaints. A special type of token here is formed by the various names of train stations and routes.
- **#omg (O)** *OMG* is short for "Oh, my God". Here we expect quite emotional tweets.
- **#syrie (S)** Tweets with this hashtag discuss the situation in Syria. These tweets mostly contain reports and comments, and they regularly refer to and quote from foreign media. Although the tweets have been marked as Dutch, we find a surprising number of tweets here in a foreign language, mostly French.

For each of these hashtags, we randomly sampled the tweets in the data collection available from the Dutch eScience Centre (Tjong Kim Sang and van den Bosch 2013) with a date stamp from January 1, 2011 to June 30, 2013. Sampling for each hashtag continued until at least 1,000 word tokens (see below) contained in Dutch tweets were found.[6]

## 3 Annotation of experimental material

In order to get a more precise overview of the (word) tokens in tweets that are found to be problematic for standard tools, we have manually annotated all tweets in our pilot corpus. The text was first tokenized automatically, after which we marked all word and hash tokens in Dutch-language tweets which were either out of vocabulary with regard to the OpenTaal[7] word list or that did occur in the list but only because the variant form just happened to be a

[5] If we wanted to find a lot of dialect, we should have selected tweets from the province of Limburg. However, Limburg dialects are often closer to being another language entirely.

[6] With the annotators deciding whether tweets were in Dutch and how many word tokens were present..

[7] 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.

homograph of an item in the list (e.g. the token *na* used as the preposition *naar* instead of the preposition *na*). Where necessary, we corrected the automatic tokenization.

We tokenized all text samples with our own specialized tokenizer for tweets.[8] Apart from normal tokens like words, numbers and dates, it is also able to recognize a wide variety of emoticons. 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. URLs and email addresses are not completely covered. The tokenizer counts on clear markers for these, e.g. http, www or one of a number of domain names for URLs. Assuming that any sequence including periods is likely to be a URL proves unwise, given that spacing between normal words is often irregular. And actually checking the existence of a proposed URL is computationally infeasible for the amount of text we intend to process. On the current sample, the tokenizer performs adequately, except that it still misses a number of emoticons, e.g. *:$* and *o.o*. In addition, the samples include one missed URL, *dlvr.it/10hgmg*. Finally, for words and hashtags, the tokenizer assigns a classification INVOC (in-vocabulary) or OOV (out of vocabulary), by looking up the token, decapitalized and stripped of diacritics, in the abovementioned word list (similarly normalized).

When annotating, we also found several tweets which were written completely in a different language than Dutch. The eScience Twitter corpus was collected by searching for tweets with any of a number of probably Dutch words, after which a character n-gram language filter was applied (Tjong Kim Sang and van den Bosch 2013). For older sections of the corpus, only tweets clearly marked as Dutch are included. Later sections include more tweets, together with an indication of the language proposed by the filter. Where this is another language, such as French or English, or where this is UNKNOWN, we automatically exclude the tweet from our sample. However, there is also a marker *notdutch*, which we find for both foreign language tweets, e.g. English, and for Dutch language tweets containing multiple non-Dutch tokens. For the manual annotation, we remove tweets entirely or mainly written in a foreign language by hand.[9]

The tokenizer distinguishes between the following types of token:

- **<word>** A normal word, as can be expected to be found in a dictionary. Apart from letters, a word may include digits (e.g. *A4*) and punctuation (e.g. *dag-/nachtlicht*).
- **<rt>** The sequence RT, used in tweets to indicate a retweet.
- **<num>** A numerical token. This includes numbers, but also dates, times, phone numbers etc.
- **<hash>** A hashtag as prescribed by Twitter.
- **<@>** A Twitter user name which is addressed, marked as such, in the tweet. The name of the author of the tweet is not annotated.

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

[9] In the automatic estimation procedure, we will also exclude notdutch, leaving only tweets that are likely to be Dutch.

- **<url>** An included URL.
- **<eaddr>** An included email address.
- **<emo>** An emoticon which is built including symbols. Emoticons built completely from letters, e.g. *xd*, are tokenized as <word>.
- **<symb>** All other symbols or symbol sequences. Often, but not always, symbols are punctuation marks.

In addition there are a few minor types for rare special cases. We only annotated <word> and <hash>. In this paper, though, we will focus only on the tokens of type <word>.

For the annotation of the <word> and <hash> tokens we applied the following markers for OOV tokens and INVOC tokens where these were used in a non-standard manner:

- **Missing.** The token is a correctly spelled word, but proves to be absent from the OpenTaal word list. We subclassify these as neologisms (missing-neo) or traditionally known words (missing-trad). Many of the neologisms are related to the new media, e.g. *facebookaccount*, *smst*, *tweet*, and *appt*.
- **Diminutive.** The token is a correctly spelled diminutive form of an INVOC word, but the form is OOV, e.g. *drinkmaatje*, *zenuwtrekje*.
- **Compound.** The token is a correctly spelled compound, which is not present in the list, e.g. *verkeershel*, *hamsterwangen*, and *schildpaddennek*
- **Hyphenation.** The token is spelled correctly, except for the hyphenation which is incorrect, e.g. *leeg-halen* and *bom-aanslag*.
- **Complex.** The token contains punctuation for some special effect, e.g. *huis/leerwerk* which combines *huiswerk* and *leerwerk*.
- **Proper name.** The token forms, by itself or in combination with adjacent tokens, a proper name. We will discuss these separately below.
- **Spelling.** The token is spelled in a non-standard manner. This could be due to a typographical error (e.g. *funcitoneert* instead of *functioneert*), but could also be intentional (e.g. *regenbooog* instead of *regenboog*, in a tweet mimicking the lyrics of a song 'Vlieg met me mee naar de regenboog!'). We subclassify these into lexicalized spelling variants (spelling-lex; e.g. *me* instead of *mijn* for the first person possessive pronoun) and productive ones (spelling-prod; e.g. *zukkels* instead of *sukkels* or *wilt* rather than *wil*).
- **Abbreviation.** As spelling, except that the variant spelling is clearly meant to shorten the word. Again, we differentiate between lexicalized (abbreviation-lex; e.g. *tv* for *television* and *ff* for *even*) and productive (abbreviation-prod; e.g. *is* for *eens* and *gewn* for *gewoon*) instances.
- **Dialect.** The token is a dialectal form. Here we distinguish between out-of-vocabulary forms (dialect-oov; e.g. *houdoe* is a dialect word originating from Brabants, one of the dialects spoken in the south of the Netherlands) and forms which are confused with a word in the standard word list (dialect-conf; e.g. *ons* as dialectal possessive form where standard Dutch would have *onze*).
- **Street language.** As dialect, except that this is the "street dialect" rather than a regional dialect. Again we distinguish street_language-oov (e.g. *wollah* and

*djoeken*) and street_language-conf (e.g. *kantelen*, street language for 'to kill', where in standard Dutch its meaning is 'to turn over').

- **Foreign.** A word from another language, which we consider not to be lexicalized yet in Dutch, e.g. *party, ciao, jihad*.
- **Interjection.** The token is an interjection. These tokens are often variants of invocabulary interjections (e.g. *hahaha*), with sometimes extreme repetition to stress the degree of emotion.
- **Emoticon.** The token is a recognized short letter combination expressing some emotion, e.g. *xd*.
- **Formula.** A non-linguistic combination, often containing measurements, e.g. *1u30* for an hour and a half.
- **Part of multiword.** The token forms a multi-token expression together with one or more adjacent tokens, and at least one of the tokens in the expression is not present in the word list, e.g. *in feite*, where *feite* is absent from the list (both tokens are marked).
- **Clitic.** The token is a concatenated, and usually shortened, combination of a pronoun and a verb, e.g. *kzal* for *ik zal*.
- **Merge.** The token is another concatenated combination of words. We distinguish between instances where the concatenation is used to form a hashtag (merge-hash; e.g. *#zieligpersoon*) and other concatenations produced by the author (merge-aut; e.g. *ofzo* for *of zo*).
- **Split.** The token is the beginning of a sequence of tokens which together form a word. We subclassify as to the reason for splitting. Just as for merge, we see splits for reasons of hashtag formation (split-hash; e.g. *trein #storing* in which the word *treinstoring* has been split to be able to use the hashtag *#storing*) and other author produced splits (split-aut; e.g. *ex politici, stop gezet*). Only when it is clear that the split was not intended do we mark it as such (split-typo; e.g. *moete n* for *moeten*).[10]
- **Insplit.** The token is a follow-on of the preceding split-token.
- **Clipped.** The token has been clipped because the text was cut off by Twitter or by a retweeting author, e.g. *probl, werel, reizig*.
- **Unknown.** The annotators are unable to determine the intended form and meaning of the token and can therefore not assign it one of the above classes.

Proper names form a rather special class of tokens.[11] They are usually only partly covered in lexical resources and therefore one can expect that in any text a proportion of OOV words can be explained in terms of proper names. In an NLP context proper names are often handled not by relying on a lexicon, but by some heuristics or separate module dedicated to their identification. In Dutch as in many other languages, proper names can often be recognized

[10] The reason for the indication aut for merge and split is that there are also splits caused by tokenization errors (split-tok; e.g. *o. o* for the emoticon *o.o*); however, these are corrected during annotation and will therefore no longer be found in the frequency counts below. Note that merge-tok does not currently occur, but this could change if a future tokenizer attempts to recognize multi-token units.

[11] We include under this class also derived forms such as adjectives, e.g. *Turkse*.

because they are capitalized. In Twitter, unfortunately, capitalization is often not regular. As a result, the identification of proper names is even more problematic than with more conventional text types. In order to be able to estimate the size of the problem, we differentiate between proper names written with (cap) or without (decap) a starting capital letter. However, there is another complication. It may be that a proper name, stripped of case and diacritics, coincides with another word which is in the list, and is therefore confused with it, e.g. minister *Kamp* is not in the word list, but the common noun *kamp* is. All in all, we distinguish seven cases:

- **Oov-cap**. The stripped form is not in the stripped list and the form was capitalized in the tweet, e.g. *Kerry*, and might therefore be recognized as a proper name.
- **Oov-decap**. The stripped form is not in the stripped list and the form was not capitalized in the tweet, e.g. *kilkowski*, and would most likely be processed incorrectly.
- **Inlex-decap**. The stripped form occurs only in the stripped list as proper name, and it can therefore be recognized as such. However, the form is not capitalized, e.g. beatrix.
- **Inlex-capdia**. The stripped form occurs in the stripped list as proper name. The form is capitalized, but deviates in the use of diacritics and/or use of capitals with regard to the standard spelling, e.g. *SYRIE* instead of *Syrië*, and *PVDA* instead of *PvdA*.
- **Inlex-decapdia**. The stripped form occurs only in the stripped list as proper name. The form is not capitalized, and deviates in the use of diacritics and/or use of capitals with regard to the standard spelling, e.g. australie instead of Australië, and *ipod* instead of *iPod*.
- **Conf-cap**. The stripped form occurs (also) in the stripped list due to a non-proper-name word. However, the form is capitalized, e.g. *Ban* (in *Ban Ki-moon*), where *ban* ("ban") is in the list as a common noun.
- **Conf-decap**. The stripped form occurs (also) in the stripped list due to a non-proper-name word. This problem is aggravated by the form being written in the tweet without capitalization, e.g. *robben* (instead of *Robben*), where *robben* ("seals") is in the list as a common noun.

The annotation process is not yet completely streamlined. Currently, we have all tokenized samples in Excel, with the rows each pertaining to one token and the columns to the various fields of information. When annotating, we regularly switch between the original order, so that we have a good view of the context, and sorting on specific column combinations, so that we can consistently annotate specific types of tokens. If we would ever want to annotate more material, an instruction manual for the annotators would obviously be useful.

## 4 Estimates from the annotated material

The pilot corpus that we compiled, after tokenization, comprises 14,783 tokens. A breakdown of the tokens into the different types of non-words and words is shown in Table 1. The

spread in the total proportion of non-word tokens is sizable. #doodsbedreiging has the highest with 35.6% non-word tokens. This is not surprising seeing the high number of retweets (mostly by people commenting on the tweet, either by an involved person to bluff back in the threat discussion or as an outsider to complain about the awfulness of this kind of tweet), each leading to an <rt>, sometimes a <symb> (the colon) and an <@>, plus often an <@> for the threatened person. At the low end, we find 20.2% for #irri, which is more surprising. Apparently, irritation is mostly uttered in words and is shouted to the world rather than addressed to specific people. In general, the differences we find are unexpected. The emotional groups (#irri, #jaloers and #omg), e.g., are very different, which suggests that each emotion has its own means of expression. The only grouping where we do find similarity is the two transport hashtags, #ns and #file. Here only the number of hashes shows a real difference; if we examine these, we see that #file has many more hashtags for the location and the reason for the traffic jam, where #ns tends to just list this information in the text. If we look at similar proportions of specific types of tokens, we often see unexpected rather than expected combinations. Take <symb>, where the highest numbers are seen for #doodsbedreiging (210 <symb>) and #aardbevingen (205 <symb>). Death threats and discussions of earthquakes are hardly comparable topics. On further examination, differences come out: in #doodsbedreiging the high symbol count is again due to the retweets which are expressed with a colon, whereas in #aardbevingen we see more proper punctuation than for most other hashtags and relatively many quotes. Just as for traditional text types, we can conclude that each domain has its peculiarities.

**Table 1.** Frequency and distribution of the different types of token in the various samples. The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

| | A | D | F | H | I | J | M | N | O | S |
|---|---|---|---|---|---|---|---|---|---|---|
| <@> | 56 | 122 | 32 | 12 | 18 | 101 | 18 | 26 | 43 | 47 |
| <emo> | 5 | 0 | 5 | 11 | 8 | 19 | 9 | 4 | 5 | 0 |
| <hash> | 181 | 110 | 220 | 249 | 127 | 149 | 159 | 138 | 173 | 142 |
| <num> | 17 | 18 | 27 | 14 | 7 | 18 | 28 | 33 | 16 | 14 |
| <rt> | 33 | 83 | 10 | 1 | 2 | 7 | 12 | 6 | 16 | 28 |
| <symb> | 205 | 210 | 162 | 114 | 92 | 160 | 141 | 192 | 163 | 190 |
| <url> | 26 | 9 | 17 | 3 | 0 | 4 | 4 | 12 | 9 | 19 |
| <word> | 1007 | 1003 | 1008 | 1002 | 1010 | 1016 | 1005 | 1001 | 1014 | 1007 |
| Percentage non-word tokens/token | 34.2 | 35.6 | 32.0 | 28.8 | 20.2 | 31.1 | 27.0 | 29.2 | 29.6 | 30.4 |

When we look at the proportion of OOV words in the different samples (Table 2), we find a better distinction between our preconceived groups. The emotional hashtags all show a high OOV proportion (#irri 10.8%, #jaloers 11.0% and #omg 10.9%). They are joined by #houdoe (11.1%) which is not necessarily emotional, but the familiar greeting does indicate a more personal involvement, so that more informal language should not be unexpected. Lower proportions are found for the more factive and/or newsrelated hashtags (#syrie 3.9%, #aardbevingen 4.3%, #ns 4.5% and #file 5.6%). That #miljoenenjacht (7.5%) is somewhere

in between is also to be expected, as part of the tweets reflect personal, sometimes emotional reactions to what is happening in the tv show. Only #doodsbedreiging with a rather low proportion of 9.0% is somewhat surprising, but of course this sample also includes tweets with less emotional commentary on the threats.

**Table 2.** Frequency and distribution over various samples of the problematic (PROBLEM) word tokens, i.e. the out-of-vocabulary (OOV) and of in-vocabulary word tokens that are used in an alternative way (INVOC-ALT), i.e. other than the use foreseen for their entry in the word list. The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

| | A | D | F | H | I | J | M | N | O | S |
|---|---|---|---|---|---|---|---|---|---|---|
| OOV | 43 | 90 | 56 | 111 | 109 | 112 | 75 | 45 | 110 | 39 |
| Percentage OOV/word token | 4.3 | 9.0 | 5.6 | 11.1 | 10.8 | 11.0 | 7.5 | 4.5 | 10.9 | 3.9 |
| INVOC-ALT | 34 | 51 | 40 | 58 | 34 | 41 | 27 | 49 | 46 | 17 |
| Percentage INVOC-ALT/word token | 3.4 | 5.1 | 4.0 | 5.8 | 3.4 | 4.0 | 2.7 | 4.9 | 4.5 | 1.7 |
| PROBLEM | 77 | 141 | 96 | 169 | 143 | 153 | 102 | 94 | 156 | 56 |
| Percentage PROBLEM/word token | 7.7 | 14.1 | 9.5 | 16.9 | 14.2 | 15.1 | 10.2 | 9.4 | 15.4 | 5.6 |

When we consider the proportion of tweets that contain one or more OOV word tokens, we also see (Table 3) a sizeable variance, from 31.5% for #syrie to 65.5% for #doodsbedreiging. Furthermore, the proportion of affected tweets is not necessarily correlated with the proportion of affected words. #houdoe, that shows the highest proportion of OOV words (11.1%) has an OOV tweet proportion of only 33.3%, the second lowest. This implies that #houdoe is a mix of reasonably clean tweets and tweets that contain a lot of OOV words.

**Table 3.** Proportion of problematic tweets in the various samples: overall (PROBLEM TWEET/tweet), containing OOV word tokens (OOV TWEET) or containing in-vocabulary words used in an alternative way (INVOC-TW). The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

| | A | D | F | H | I | J | M | N | O | S |
|---|---|---|---|---|---|---|---|---|---|---|
| Percentage OOV-TWEET/tweet | 36.6 | 65.5 | 45.2 | 33.0 | 61.2 | 57.3 | 45.6 | 35.1 | 57.8 | 31.5 |
| Percentage INVOC-TWEET/tweet | 26.8 | 35.7 | 29.0 | 26.1 | 27.2 | 23.3 | 16.8 | 41.9 | 31.0 | 16.4 |
| Percentage PROBLEM-TWEET/tweet | 48.8 | 75.0 | 55.9 | 47.9 | 72.8 | 65.0 | 53.6 | 58.1 | 70.7 | 37.0 |

Continuing on to the in-vocabulary word tokens that are used in an alternative way to the one foreseen for their entry in the word list (INVOC-ALT), we see (Table 2 and Table 3) fairly high proportions (1.7% to 5.8% for the words and 16.4% to 41.9% for the affected tweets), especially considering that these tokens are hardly ever recognized as problematic because the focus is generally on OOV words. On average, there are about half as many INVOC-ALT words as there are OOV words, but the two proportions are not correlated (correlation factor of only 0.465).

If we look at both types of problematic words together, we see proportions ranging from 5.6% (#syrie) to 16.9% (#houdoe) at the word level (Table 2) and 37.0% (#syrie) to 75.0% (#doodsbedreiging) at the tweet level (Tabel 3). The most important finding of the annotation and its analysis may well be that tools that have to rely on their built in lexicon will have trouble with at least one in three tweets, for more extreme topics even three in four. We think this can justly be called a serious problem.

In order to judge how easy it might be to solve this problem, we need to look at the individual types of problematic words (see Tables A and B in the appendix for a detailed overview of the frequency and distribution of the various types of problematic word tokens). The easiest solution would be to add frequently occurring problematic words to the lexicon. These would mostly be the words that can be considered to be lexicalized (either in general or at least on Twitter), i.e. spelling-lex and abbreviation-lex, together with the often social-media-related neologisms (missing-neo). These three groups together make up about one third of the problematic cases. This means that this simple solution considerably alleviates the problem, but by no means solves it completely. A next partial solution might be to add pattern matching techniques to recognize specific classes of productive tokens. Emoticons are an example of this, but then we are currently considering only the words. Here, only (most of) the OOV interjections form such a class. Assuming we could recognize all interjections, this would resolve about one tenth of the problematic cases, not impressive but still worthwhile. A final substantial class of problematic words is formed by the proper names, about one fourth of the problematic cases. Here we would have to adapt existing named entity recognition techniques to the kind of text found on Twitter. This will certainly not be easy, as the two main information sources for NER are both corrupted. As for capitalization, only about half of the problematic proper name tokens are written with a capital letter. As for context, the system would have to be able to cope with spelling variation as well as deviant syntax in the surrounding text. As for other types of problematic cases, no easy solutions come to mind.

## 5 Automatic estimates of proportions of problematic tokens

The manual annotation of data for spelling variation is rather labour intensive. Obviously, if we want to investigate whether and how the proportion of spelling variants varies per user or per topic, and we need a sufficiently large amount of data to be able to draw conclusions, manual annotation is out of the question. We will therefore have to look to automatic means to estimate such proportions.[12]

So far we have considered two types of problematic cases. First, there are the out-of-vocabulary words. These are in principle relatively easy to find. We only need proper tokenization and then lexicon lookup. However, tokenization, as mentioned above, is not flawless. Furthermore, there is the problem of the presence of non-Dutch tweets in the material. Still, we will show below that we can automatically derive adequate estimates for OOV

[12] Note that, in this paper, we are not concerned with finding the actual problematic cases, but merely in estimating how many there are as a proportion of the total number of tokens, which means that a comparison at the level of an overall number is sufficient and that we do not have to use e.g. precision and recall to measure that we are finding the correct cases.

words. The in-vocabulary words with alternative uses, however, are much more problematic. Attempting to find them will involve language models, be it grammars or n-gram statistics, and although we are working on this (van Halteren and Oostdijk 2012), this work is not yet at a stage where we could sensibly attempt this task. Also, the frequency of these words is not strongly correlated with the frequency of the OOV words: looking at our ten samples, the two show a Pearson product-moment correlation of only about .46 Therefore, extrapolation from the OOV frequency is not possible. For now, we have decided to concentrate on estimating the proportion of OOV words.

There is yet a third source of possible problems for NLP applications. Such applications are designed to work with words and punctuation, possibly including numbers and abbreviations. All the other types of tokens that we encounter in tweets will severely hamper applications such as the already mentioned text-to-speech and translation systems. For this reason, we will also estimate the proportion of non-word tokens for various tweet types.

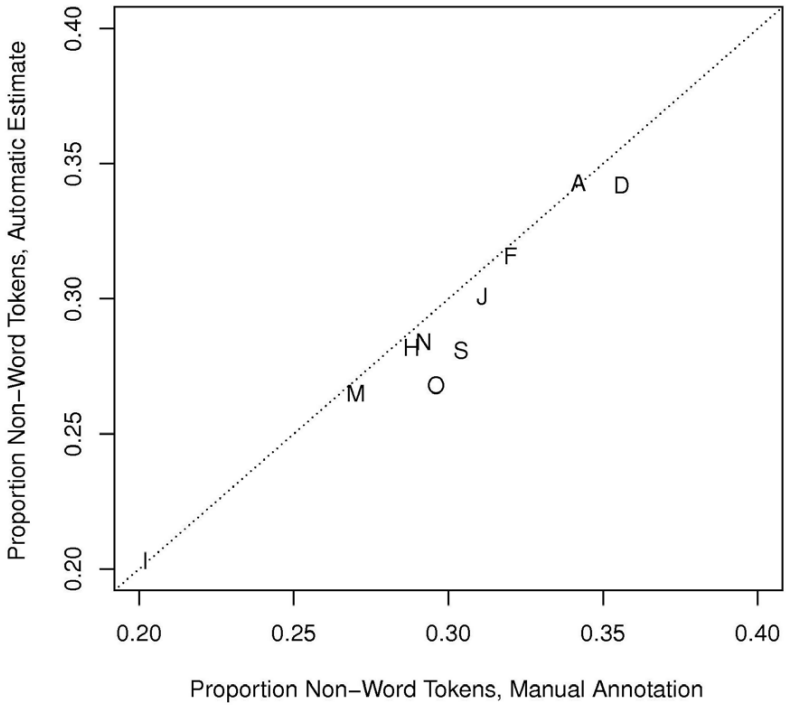


Figure 1. Benchmarking the automatic estimate of the proportion of non-word tokens on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

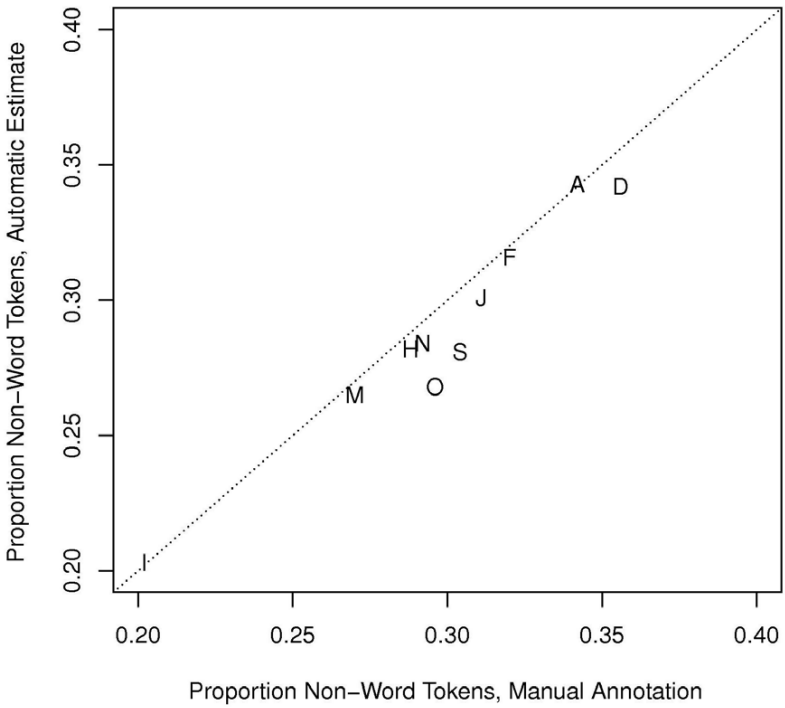


Figure 2. Benchmarking the automatic estimate of the proportion of non-word tokens on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

For both automatic estimates we will do just what we described above: tokenize and look up the tokens in the lexicon. We will ignore the fact that the tokenizer is known to make occasional mistakes. We do try to compensate for the presence of non-Dutch tweets in the material. All tweets marked either for another language or as UNKNOWN or notdutch are left out.

We tested the estimation process on the manually annotated material, starting from the raw rather than from the manually corrected version. Figures 1 to 3 show the comparison between the estimates derived from the manually annotated data and those derived automatically.

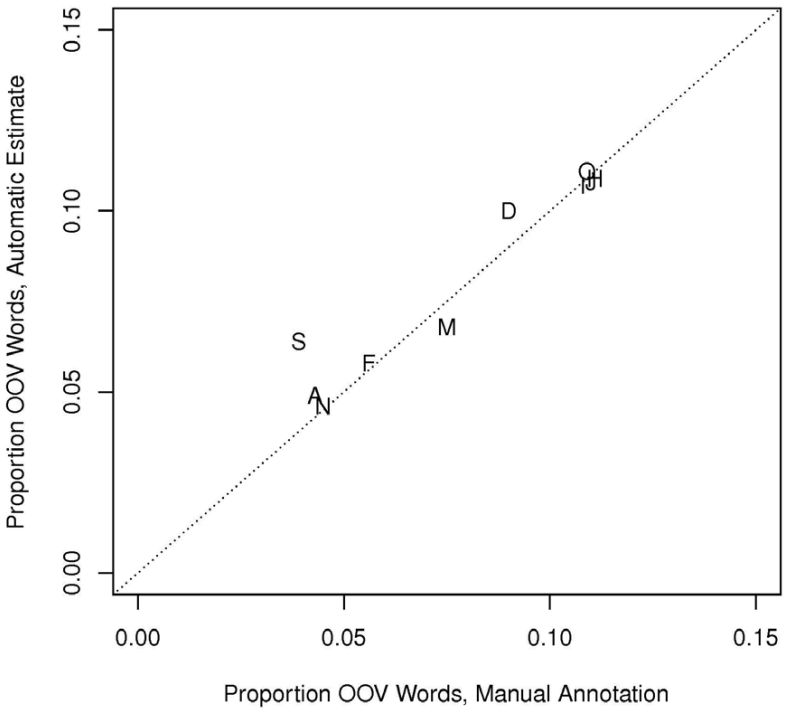


Figure 3. Benchmarking the automatic estimate of the proportion of OOV words on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

For the proportion of non-word tokens (Figure 1), the estimate is clearly adequate. This is confirmed by a correlation of .974 (confidence interval .890-.994). The same can be said for the proportion of OOV words within all words (Figure 2), although it has a slightly lower correlation of .959 (confidence interval .833-.991). The procedure is less effective when it tries to estimate which proportion of the tweets contain OOV words (Figure 3). Given the lower correlation of only .828 (confidence interval .415-.958), and with the errors concentrated in specific samples, we will refrain from using this estimate in the investigations below.

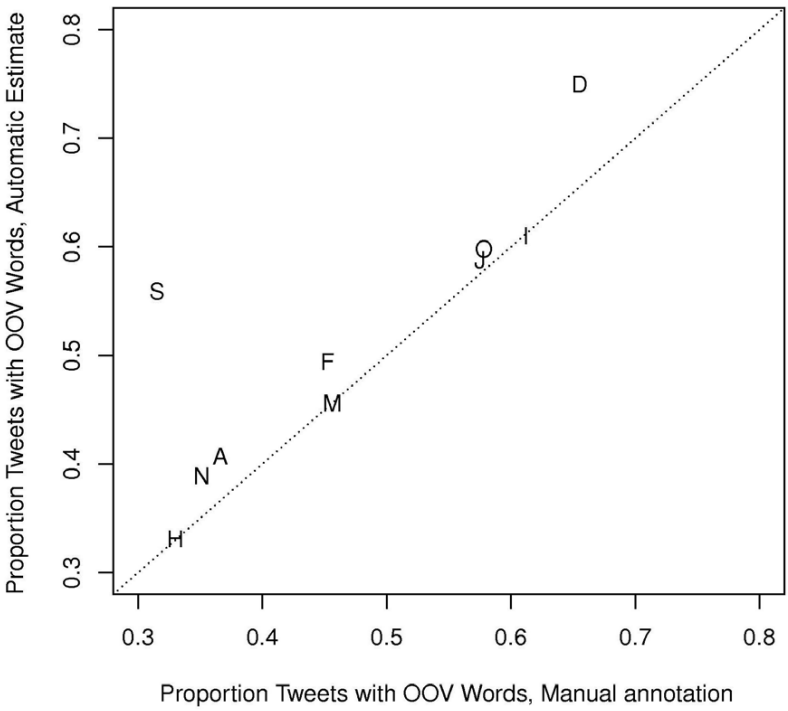


Figure 4. Benchmarking the automatic estimate of the proportion of tweets including OOV words on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

## 6 Automatic estimates for the whole tweet collection

We applied the estimation procedure described above to all tweets in the eScience Centre Twitter corpus with date stamps between January 1, 2011 and June 30, 2013. In all, almost 2 billion tweets marked with the language indication 'dutch' were processed, comprising about 23 billion tokens of which 17.5 billion are words (76.6%). Overall, the procedure finds about 1.8 billion OOV words (10.2%).

If we look at individual users, we see quite some variation. In Figures 4 to 9, we show measurements for all 1.7 million users producing at least 1,000 words in the given time period.[13] The full lines indicate the measurements when taken over all tweets, and the dashed lines when taken over all tweets produced by users who did not reach the word thres-

[13] The maximum number of words was not restricted and the full production of each user is being measured.

hold (i.e. the less active users). In all, there were more than 38 million active user names during this period, which is remarkably high, given that the Netherlands and Flanders together have only about 24 million inhabitants.

Starting with the OOV words (Figure 4), we first observe a large main cluster ranging from close to 0% up to about 30% OOV, which is likely to be the core Dutch speaking Twitter population. Higher up, especially around 50% and 60%, we find secondary clusters. Looking at the user names involved, we get the impression that these clusters at least partly stem from foreign tweets erroneously marked as Dutch. This does of course affect the overall estimate somewhat, but not too drastically we expect, as only about 1% of the users shows a proportion of OOV higher than 30%, and they contribute only 0.2% of the examined words. Furthermore, we also find Dutch users in the higher regions, with OOV proportions well over 90% for several contact ad feeds which consist of URLs, hashtags and compacted information fields. At the other end of the spectrum, we find users who manage to produce tens of thousands of words without any recognized OOV word. An examination here shows various automatic text generation systems, varying from ads linking to websites, to a solar power driven work of art reporting whether it is awake. Looking at the other two plots (Figures 5 and 6), we see that the proportion of non-word tokens varies more predictably, mainly between 0% and 40%. In Figure 6, the secondary clusters again show up, but mostly on the OOV dimension, so that it would seem that the proportion of non-word tokens is similar, although possibly a bit higher, for the other languages involved.

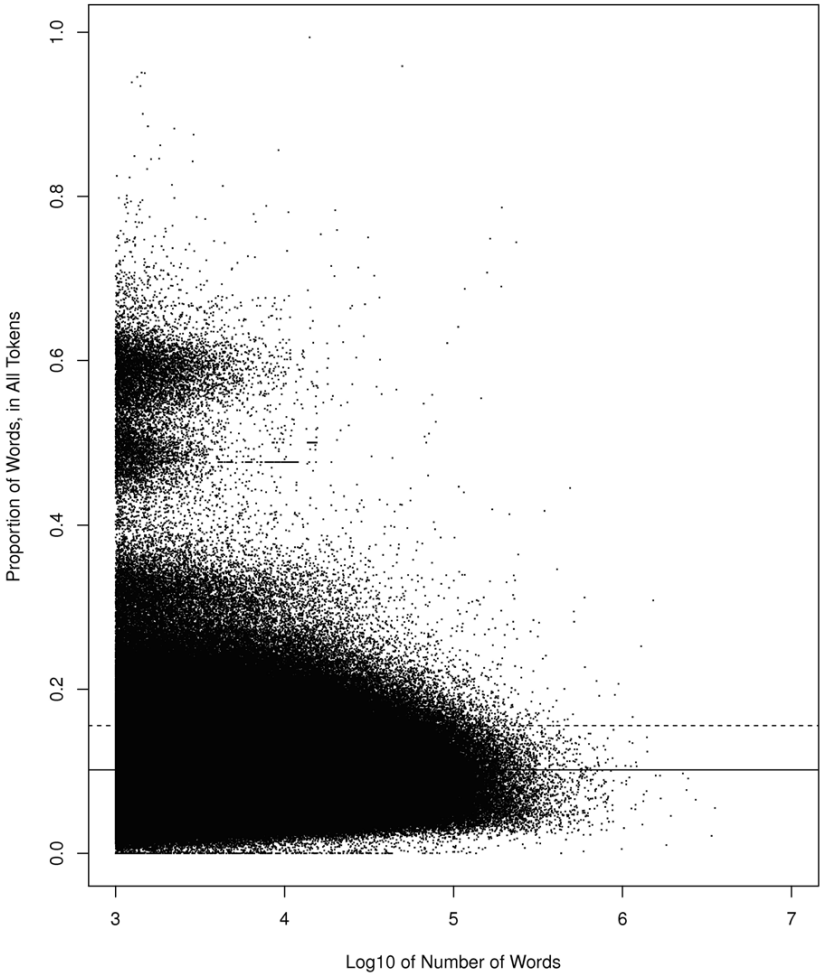


Figure 5. The estimated proportion of OOV words as a function of the produced number of words for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

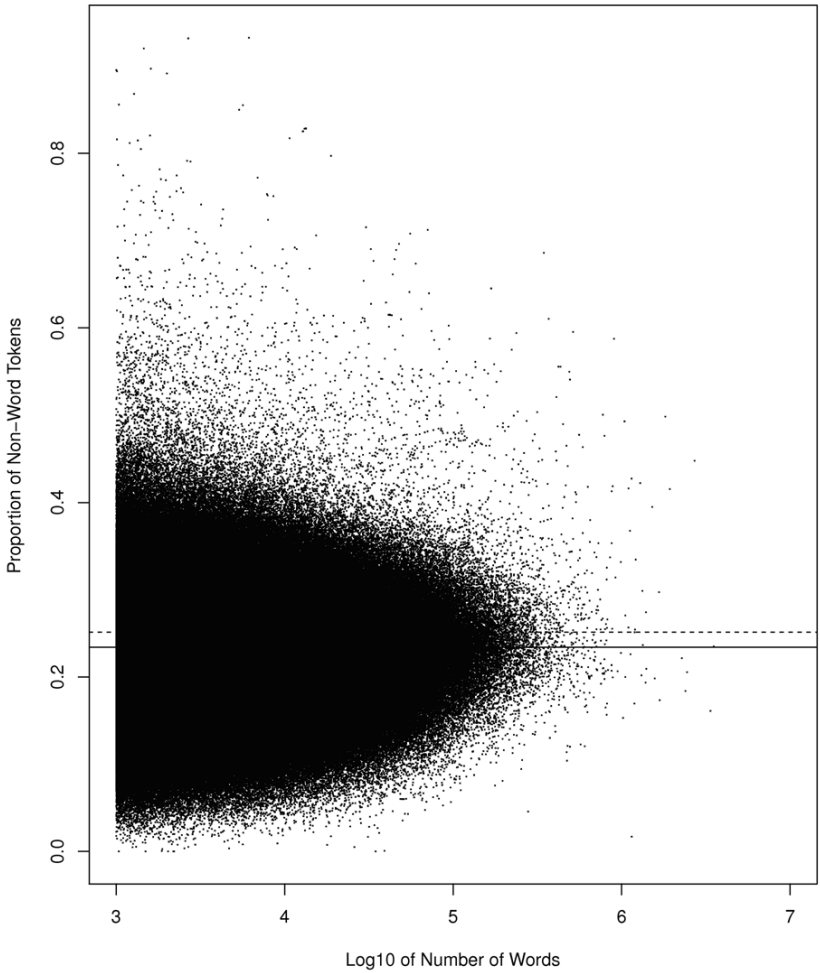


Figure 6. The estimated proportion of non-word tokens as a function of the produced number of words for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

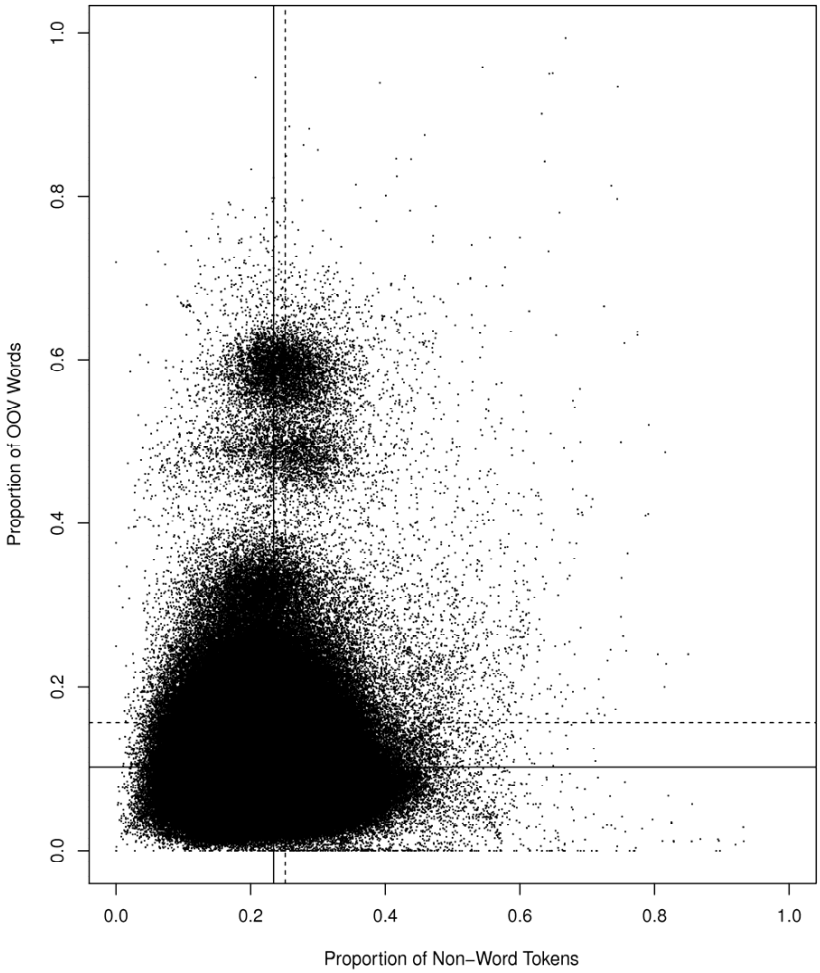


Figure 7. The estimated proportion of OOV words as a function of the estimated proportion of non-word tokens for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

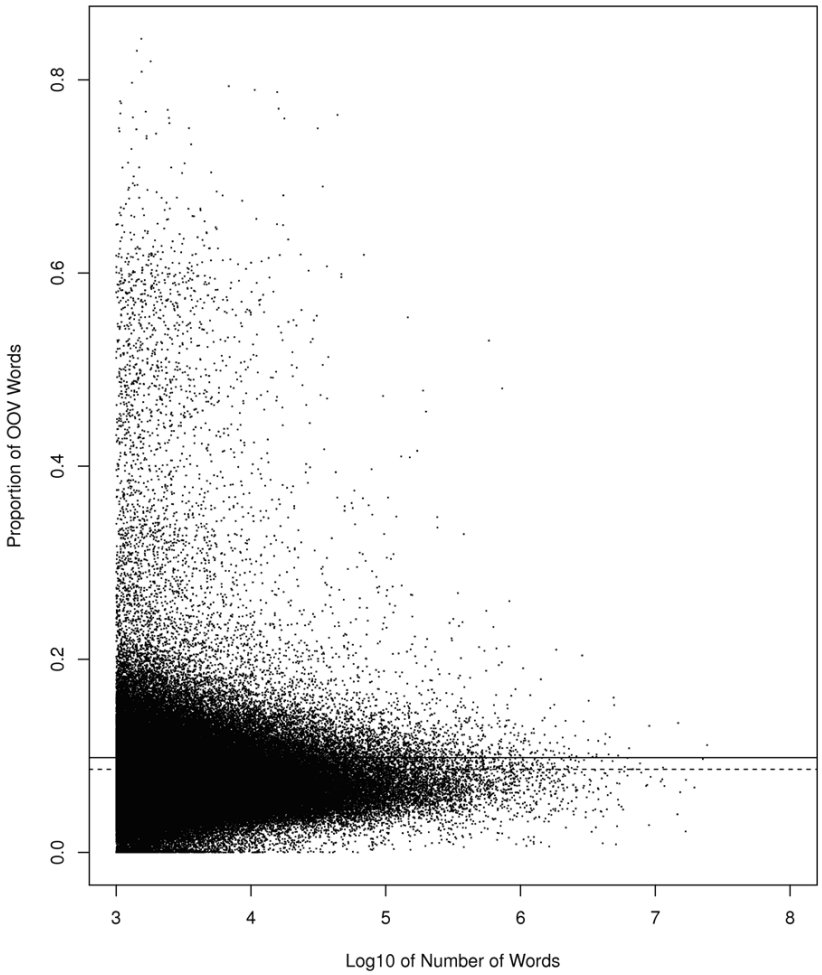


Figure 8. The estimated proportion of OOV words as a function of the produced number of words for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

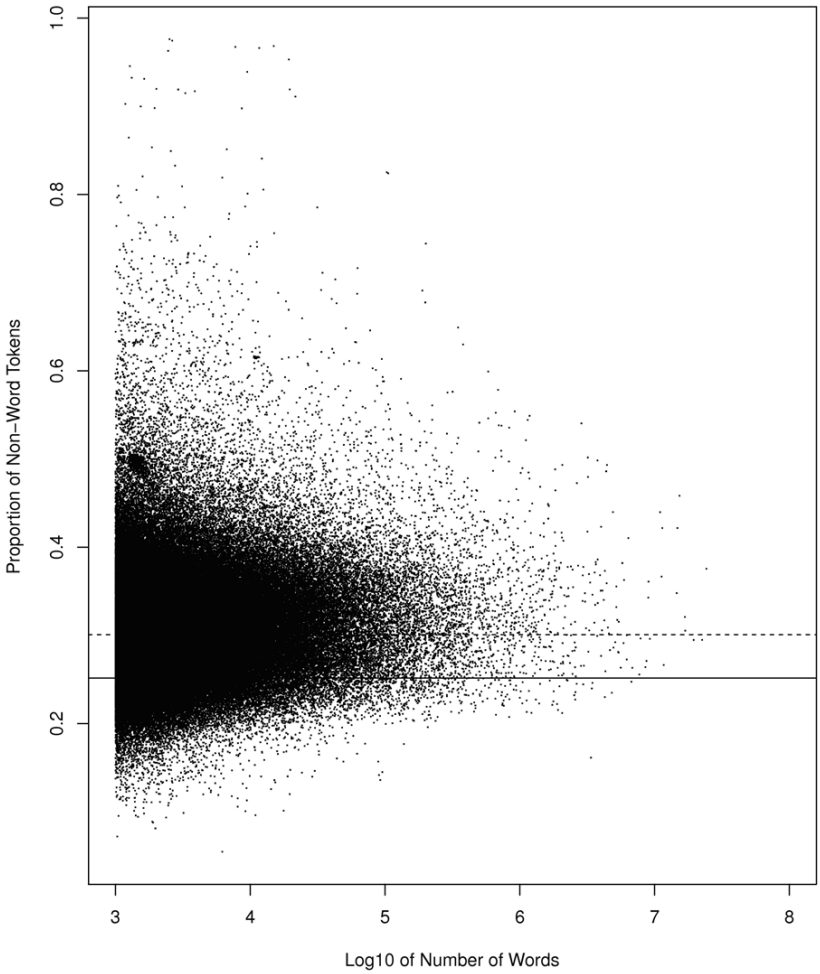


Figure 9. The estimated proportion of non-word tokens words as a function of the produced number of words for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

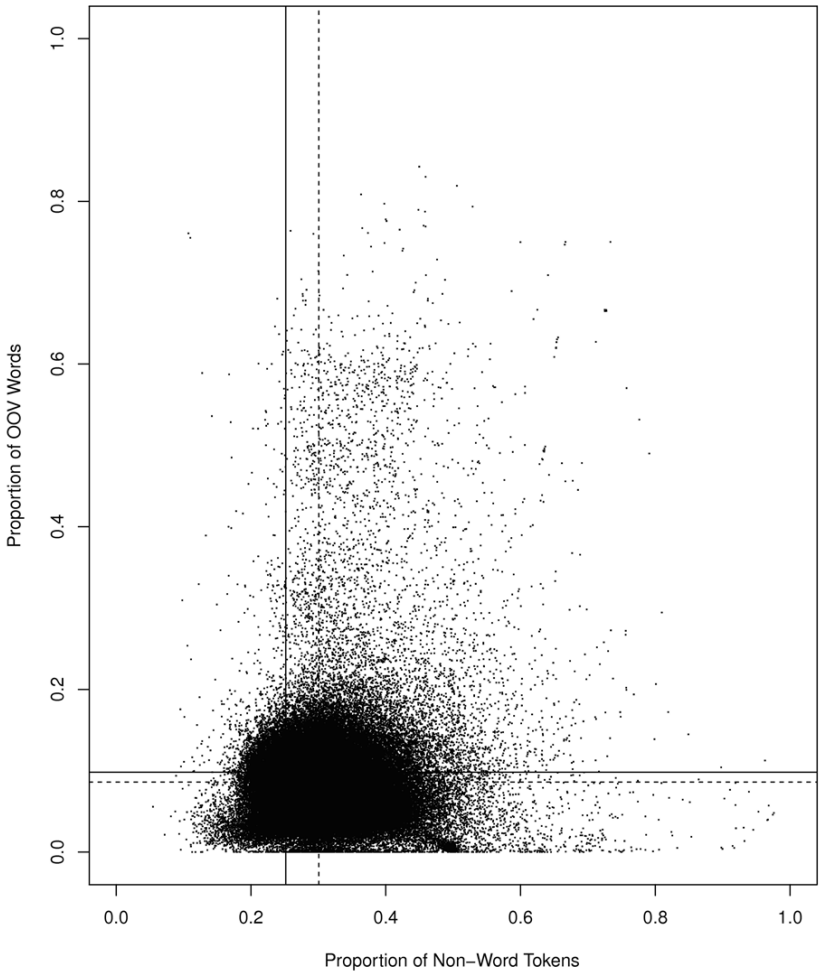


Figure 10. The estimated proportion of OOV words as a function of the estimated proportion of non-word tokens for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

We repeated this investigation for hashtags. Of the approximately 23 million hashtags used in the time period in question, only about 200,000 produced at least 1,000 words. If we examine the same plots as for users (Figures 7 to 9), we again see most activity in the main cluster. For OOV words (Figure 7), the cluster now ranges upto only about 20%. This would imply that tweets with hashtags on average contain fewer OOV words. We also get this impression from the position of the line for the overall estimate. In order to confirm our impression, we separately measured the OOV words for tweets with and without hashtags. Those with a hashtag (2.7 billion words) showed a clearly lower OOV count of 8.1% than those without a hashtag (14.8 billion words) at 10.5%. We do not see any secondary clusters in this plot, but there is a large sparse cloud extending up to around 80% OOV, which on closer examination is dominated by German hashtags. The influence of the overall estimate should be even lower than for the users, as the hashtags with a higher than 20% OOV comprise 1.7% of the hashtags, but contribute less than 0.1% of the words. As for non-word tokens (Figure 8), we see a similar spread, except that the cluster moved up slightly. The average for the tweets with hashtags is 30.1% versus 22.0% for those without hashtags. This difference is easily explained as it is probably completely accounted for by the presence of the hashtag required. The combination plot (Figure 9) does show an interesting small secondary cluster around 50% Non-Word and 0% OOV. This turns out to be caused by a single spam feed, each time asking whether you are looking for a specific specialist profession (hashtag) in a specific location (hashtag) and giving a URL.

## 7 Conclusion

In this paper, we investigated the proportion of tokens in tweets which might cause problems for automatic processing. We were able to look in detail at a pilot corpus containing for each of ten selected hashtags a random selection of tweets containing around 1,000 words (Sections 2 to 4). We also automatically estimated the proportion of non-word tokens and OOV words on almost 2 billion Dutch tweets (Section 6), after having shown that the automatic estimation procedure is adequate for these two measurements (Section 5).

In our annotated samples, we see that the proportion of non-word tokens ranges from 20% to 36%. The proportion of OOV words ranges from 4% to 11%, whereas forms judged to be in vocabulary because they are homographs of listed words range from 2% to 6%. Especially the latter class calls for our attention, as these tokens tend to go undetected at first but are bound to have a negative effect when attempting to automatically process texts with standard resources. In all, there are problematic words in 37% to 75% of the tweets in the examined ten samples. In an automatic investigation of all tweets, we see that, for the bulk of tweet types, the proportion of non-word tokens ranges between 0% and about 40%, with an average of around 23%, and of OOV words between 0% and about 30%, with an average of around 10%. In the automatic investigation, we did not attempt to investigate the proportion of confused words and the proportion of tweets with problematic cases; however, if the average of about half as many INVOC-ALT words as OOV holds for the whole collection, there should be about 15% of problematic words in total. The differences between the measurements for annotated and automatically processed material are consistent with our fin-

dings that tweets with hashtags generally have a higher proportion of non-word tokens, but a lower proportion of OOV words.

We found pronounced differences between tweet samples focusing on different topics. In the annotated material, there is a clear gap between the OOV proportions for the hashtag with a higher expected emotion load and those with a lower one. We are somewhat surprised that the tweets related to Dutch Rail are found to be in the non-emotional cluster, but then there is a large number of official tweets among them. The difference due to emotional load is not visible in the automatic estimates, but this may well be because of the plot denseness caused by showing measurements for 200,000 hashtags. For non-word tokens, neither analysis shows a clear pattern. There is also extreme variation between users, as can be expected, but we observe no recognizable clusters. Outside the main bulk of users and hashtags, we find several deviant types of tweets, such as spam. Also, we find foreign language tweets, many of them German, even though the language filter marked them as Dutch. Although these do not appear to seriously impact our main findings, we would like to investigate how the material can be cleaned up in this respect.

Where the automatic estimates only provide overall percentages, the analysis of the annotated material gives us more insight in the types and distribution of problematic tokens. The main conclusion from our analysis is that at least some of the problems can be solved with relatively simple means, e.g. addition of lexicalized items to the lexicon (about one in three of the problematic cases) or patterns matching techniques (about one in ten), or possibly more involved ones, e.g. adaptation of named entity recognition to the Twitter environment (about one in four).

With these findings, we can now also address our underlying research questions. Given the observed proportions of problematic cases, it would seem unwise in most cases to attempt to process the bulk of Dutch tweets with NLP tools developed for standard Dutch. Especially the proportion of tweets in which problems arise is too high for this. However, it might be possible to process tweets authored by specific (types of) users or containing specific hashtags. And there are clear approaches to alleviate the problem by adjusting the tools. On the other hand, we should remember that we have only addressed the lexicon in this paper, and that other problems such as deviant syntax are also present.

Now that it has been confirmed that Dutch tweets need special means for proper processing, we will continue our work in this area. Apart from the work on spelling variation (van Halteren and Oostdijk 2012), this is likely to include improved language filters so that we can better focus on Dutch. Also, we are inspired to initiate a deeper investigation into classes of users and topics, as it appears that the best approach to processing might well vary per class, and that, for most types of research, it is advantageous to focus on specific types of tweets, e.g. those either more or less emotional. Furthermore, for most types of research one probably would like to remove spam tweets. Finally, seeing the proportion of problematic cases we found, and seeing for example the surprising finding that tweets with hashtags are somehow different from those without, it might also be interesting to investigate how these facts could have affected previous or existing research.

## Bibliography

Han, B. and T. Baldwin (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In Proceedings of the 49th Annual Meeting of the Assoc. of Computational Linguistics. Portland, Oregon, June 19-24. ACL. 368-378.

Krikorian, R. (2013). New Tweets per second record, and how! Twitter Offical Blog. August 16, 2013.

Oostdijk, N. and H. van Halteren (2013). N-gram-based Recognition of Threatening Tweets. A. Gelbukh (ed.) CICLing 2013, Part II, LNCS7817, pp. 183-196. Springer-Verlag Berlin/Heidelberg.

Small, H., K. Kasianovitz, R. Blanford and I. Celaya (2012). What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data. The International Journal of Digital Curation. Vol. 7(1): 174-197.

Sidarenka, U., T. Scheffler and M. Stede (2013). Rule-based normalization of German Twitter messages. Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology.Sept. 25-27, 2013, Darmstadt, Germany. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/gscl2013/workshops/sidarenka_scheffler_stede.pdf

Tagg, C. (2009). A Corpus Linguistics Study of SMS Text Messaging. http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

Tjong Kim Sang, E. and A. van den Bosch (2013). Dealing with Big Data: The case of Twitter. CLIN Journal Vol. 3:121-134.

van Halteren, H. and N. Oostdijk (2012). Towards Identifying Normal Forms for Various Word Form Spellings on Twitter. CLIN Journal Vol. 2: 2-22.

Weil, K. (2010). Measuring Tweets. Twitter Offical Blog. Februari 22, 2010.

## Appendix

In Tables A and B a more detailed overview is given of our findings in the pilot corpus as regards the frequency and distribution of the different types of out-of-vocabulary words (OOV; Table A) and in-vocabulary words that are used in an alternative way, i.e. other than the use foreseen for their entry in the word list (INVOC-ALT; Table B).

The different samples in the pilot corpus are referred to by means of capital letters as follows: A=#aardbevingen; D=#doodsbedreiging; F=#file; H=#houdoe; I=#irri; J=#jaloers; M=#miljoenenjacht; N=#ns; O=#omg; S=#syrie. Each of the hashtags has been described briefly in Section 2. For a description of the different types of OOV and INVOC-ALT word tokens, we refer to Section 3.

**Table A.** Detailed overview of the frequency and distribution of the different types of out-of-vocabulary words. The single letter column headingcs represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

| | A | D | F | H | I | J | M | N | O | S |
|---|---|---|---|---|---|---|---|---|---|---|
| Total OOV | 43 | 90 | 56 | 111 | 109 | 112 | 75 | 45 | 110 | 39 |
| abbreviation-lex | 13 | 21 | 5 | 24 | 21 | 15 | 16 | 10 | 18 | 10 |
| abbreviation-prod | 2 | 3 | 1 | 6 | 4 | 4 | 1 | 1 | 2 | 1 |
| clipped | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 1 |
| clitic | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 0 |
| complex | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 7 | 1 | 0 |
| compound | 1 | 5 | 5 | 3 | 3 | 2 | 2 | 2 | 1 | 3 |
| dialect-oov | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| diminutive | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| emoticon | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 3 | 0 |
| foreign | 5 | 4 | 4 | 15 | 2 | 10 | 6 | 3 | 11 | 6 |
| formula | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| hyphenation | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| interjection | 1 | 6 | 7 | 3 | 10 | 36 | 21 | 1 | 19 | 0 |
| merge-aut | 1 | 3 | 3 | 7 | 10 | 10 | 5 | 2 | 7 | 0 |
| missing-neo | 2 | 3 | 1 | 1 | 6 | 3 | 4 | 2 | 8 | 2 |
| missing-trad | 1 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| part_of_multi-word | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| proper_name-oov-cap | 11 | 13 | 17 | 1 | 4 | 8 | 11 | 16 | 5 | 19 |
| proper_name-oov-decap | 1 | 9 | 8 | 12 | 9 | 9 | 9 | 4 | 3 | 8 |
| spelling-lex | 1 | 9 | 2 | 5 | 9 | 5 | 2 | 2 | 6 | 0 |
| spelling-prod | 3 | 10 | 4 | 30 | 28 | 15 | 7 | 5 | 31 | 1 |
| split-aut | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 |
| split-typo | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| street_language-oov | 0 | 7 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| unkn | 0 | 0 | 0 | 5 | 1 | 2 | 0 | 0 | 3 | 0 |

**Table B.** Detailed overview of the frequency and distribution of the different types of in-vocabulary words. The single letter column headingcs represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

| | A | D | F | H | I | J | M | N | O | S |
|---|---|---|---|---|---|---|---|---|---|---|
| Total INVOC-ALT | 34 | 51 | 40 | 58 | 34 | 41 | 27 | 49 | 46 | 17 |
| abbreviation-lex | 6 | 13 | 15 | 15 | 11 | 8 | 6 | 29 | 10 | 6 |
| abbreviation-prod | 2 | 1 | 0 | 4 | 2 | 0 | 0 | 0 | 2 | 1 |
| clipped | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| clitic | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 0 | 0 |
| compound | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| dialect-conf | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| foreign | 2 | 0 | 4 | 2 | 0 | 6 | 2 | 0 | 3 | 2 |
| interjection | 0 | 2 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 |
| merge-aut | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| part_of_multi-word | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| proper_name-inlex-decap | 1 | 6 | 4 | 6 | 4 | 8 | 1 | 3 | 5 | 0 |
| proper-name-conf-cap | 14 | 3 | 2 | 2 | 1 | 0 | 2 | 4 | 6 | 2 |
| proper_name-conf-decap | 1 | 9 | 8 | 5 | 3 | 4 | 3 | 2 | 1 | 0 |
| proper_name-inlex-capdia | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| proper_name-inlex-decapdia | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| spelling-lex | 0 | 9 | 2 | 9 | 4 | 5 | 5 | 0 | 5 | 0 |
| spelling-prod | 0 | 2 | 2 | 7 | 5 | 5 | 5 | 6 | 11 | 2 |
| split-aut | 4 | 4 | 2 | 2 | 1 | 1 | 2 | 2 | 0 | 1 |
| split-hash | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| street_language-conf | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| unkn | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 0 |