Listening to the Noise: Model Improvement on the Basis of Variation Patterns in Tweets

Hans van Halteren, Nelleke Oostdijk
Radboud University Nijmegen, Dept. of Linguistics / CLST
P.O. Box 9103, NL-6500HD Nijmegen, The Netherlands
E-mail: hvh@let.ru.nl, N.Oostdijk@let.ru.nl

Abstract

In this paper, we take the view that the wide diversity in the language (use) found on Twitter can be explained by the fact that language use varies between users and from one use situation to another: what users are tweeting about and to what audience will influence the choices users make. We propose to model the language use of Twitter tribes, i.e. peer groups of users tweeting in different use situations. We argue that the use of tribal models can improve the modeling of the substantial variation present in Twitter (and other social media), and that the resulting models can be used in the normalization of text for NLP tasks. In our discussion of variation at the linguistic levels of orthography, spelling, and syntax, we give numerous examples of various types of variation, and indicate how tribal models could help process text in which such variation occurs. All examples are derived from our own experience with the Dutch part of Twitter, for which we could draw on a multi-billion word dataset.

Keywords: Twitter, social media, language variation, metadata induction, language modeling, spelling, syntax

1. Introduction

In many fields, data from the social media are judged to have enormous potential for research. At the same time, social media data are generally quite different from data originating from the traditional media. In many of the contexts of social media communication, the authors/users do not appear to feel bound to adhere to the norms that have been set for the standard language and deviate from these norms in their use of orthography, spelling, and/or syntax. Most of these deviations are intentional. In fact, they tend to follow conventions upheld within the authors' peer groups. This means that if we manage to identify the peer groups in question, we are able to model the variation to a large degree. This in turn leads to (a) better recognition of the factual information being transmitted as well as (b) information about the authors and their communicative goals as encoded in the variation.

In this position paper, we look at a specific type of social media data, namely text on the Dutch part of Twitter.[1] Now, in tweets we find a number of special communicative devices, either unique to Twitter or shared with other social media. Emoticons can be used to represent emotional content effectively, discussed topics can be marked with hashtags, other authors can be addressed or mentioned by quoting their username preceded by an at sign, and URLs can link to tweet-external additional content. How these devices are used by various groups is also an interesting subject of study. However, in this paper, we will ignore these devices and focus exclusively on linguistic objects already known in traditional text types.

In the following sections, we first explain our viewpoint in more detail (Section 2), after which we zoom in on separate linguistic levels, viz. orthography (Section 3), spelling (Section 4), and syntax (Section 5). Finally, we return to the overall picture for conclusions and a vision of the future (Section 6).

2. Twitter tribes

In recent years, numerous studies have been directed at the mining of social media data for various purposes. In most of these studies, the observation is made that texts in the social media are quite unlike texts published in traditional media and it is not uncommon to find that texts are being characterized as "noisy" and/or "to be corrected" (e.g. Kaufman and Kalita, 2010; Han and Baldwin, 2011). These days there is a lively research area both investigating the extent of the problem (e.g. Baldwin et al., 2013; Baldwin and Li, 2015) and committed to the attempt to extract as much information as possible despite the level of noise, using various methods as we can see e.g. in the report on a 2015 shared task on text normalisation for Twitter (Baldwin et al., 2015). The activity of the field is witnessed by the presence of a multitude of workshops, such as W-NUT, SocialNLP, NLPIT, and NormSoMe.[2]

While our work clearly falls in this research area, and also concerns the improvement of mining of social media text, our primary mining activities are linguistic studies, such as those into the linguistic variation in the social media (van Halteren & Oostdijk, 2015). Our perspective leads us to approach the task from a direction which is rather different from what appears to be mainstream, but which is closer to what we find in approaches adopted for example in Bryden et al. (2013)[3]

[1] That is, tweets by users communicating primarily in the Dutch language, most of whom are of Dutch or Flemish origin. The TwiNL dataset on which we draw is already collected in such a way that only a few percent of non-Dutch tweets remain (Tjong Kim Sang and van den Bosch, 2013), and we have managed to reduce that percentage to well under 0.5% (van Halteren, 2015).

[2] It is outside the scope of this paper to give a full inventory of such work here, and we limit ourselves to some examples. A good starting point for a deeper literature study would be the proceedings of the mentioned workshops.

[3] In fact, we adopted the term *Twitter tribes* as suggested by Jason Rodrigues in a Guardian blog about Bryden et al.'s work (http://www.theguardian.com/news/datablog/2013/mar/15/twitter-users-tribes-language-analysis-tweets).

and Eisenstein et al. (2014).

So far, in our research we have focused on Twitter as here data are available to us in large quantities (Tjong Kim Sang and van den Bosch, 2013). For any given research project, we first need to create a balanced corpus with reliable metadata. As all tweets carry a time stamp and many tweets are tagged for geolocation, efforts are mostly directed towards collecting additional metadata, viz. pertaining to the author and the use situation (topic, purposive role). As it turns out, author characteristics (gender, age) can to a fair degree be induced from the authors' language use (van Halteren & Speerstra, 2014; van Halteren, 2015). The same is true for the use situation. A special case here is the use of hashtags as a kind of user markup to indicate the main topic(s) of a tweet explicitly. However, this use of hashtags is mainly related to a specific type of Twitter discussion and is far less used in more personal tweets, i.e. the majority of tweets. This means that for topics too, we have to look at the contents, here topic-related words, rather than to the explicit metadata.

Now, on the one hand, for proper modeling of language variation and, on the other, for facilitating all mining tasks, we too want to identify some kind of "normal form" for social media text, both to be able to generalize away from individual forms and to be able to use available NLP tools. However, we feel that deriving such a normal form from the observed form can be informed by knowledge about the individual author, the peer group addressed, and the communicative goal. In the context of Twitter, we defined a "Twitter tribe" to be "a set of authors who share specific characteristics, discussing a set of related topics, in specific use situations"[4]. Such Twitter tribes can range from very narrowly focused, e.g. a specific community of twelve students discussing public transport, to very widely focused, e.g. all youngsters discussing any kind of topic. At both these levels of focus, we have found that there are measurable differences between tribes.[5] [6] As for the use situations, our investigations have so far been limited, especially since the use of Twitter is in itself already rather restricting the range of situations. However, we did observe that the language use in tweets carrying a hashtag, i.e. tweets aimed at a larger audience, conforms more to the standard language (van Halteren & Oostdijk, 2014).

Having only just proved the validity of the Twitter tribe concept, we did not yet implement a full tribal recognition engine, nor did we apply tribal models to text normalization. This means that as yet we cannot measure the potential quality improvement for any given task. However, we can give an overview of the types of variation we observed in our investigations,[7] and sketch how tribe modeling could be used to harness this variation.

3. Orthography

In traditional text types, we are used to a very specific markup system, with spacing separating words, punctuation indicating larger structural units, and capitalization fulfilling both lexical and structural functions. For most professional Twitter feeds, as well as many discussions by older users, we see that this markup is generally used in the standard fashion.[8]

Elsewhere, orthography appears to be far more random. Capitalization is generally ignored, or at most used to stress words. One reason may be that text input is not done with a standard keyboard with a simultaneous upper case key, but with some touch screen input method which toggles between separate upper and lower case keyboards. A similar situation exists for punctuation. Given the additional effort, and the fact that not using this standard markup does not seem to affect the interpretability of the message, many authors apparently decide just not to use the standard, as a side effect freeing capitalization for expressing stress. The effect when examining random tweet samples is that the use of capitals and punctuation appears almost random. Ideally, we should construct a tribe model, preferably modulated by a usage model for each individual character input method.[9] This, however, is still future work. For now, we are limited to trying to recognize that a specific user does not adhere to the traditional standards or conventions, and then (for this user) just assume that this component of the information in the message is not available.

For spacing, the input method does not seem to be the problem, as the space bar is almost always available.[10] Still, spacing as well is often different from

[4] We already have indications that the language use also changes over time. However, we decided not to include this factor in the definition of the tribe. We intend to study differences over time as a separate dimension, and view it as the evolution of each tribal language.

[5] As for narrowly focused tribes, we have investigated communities of authors (4-50 members) being in frequent contact, discussing the topic areas of school work, public transport, football, politics, and personal grooming (i.e. care for one's appearance, not to be confused with grooming in the internet predator sense). When comparing n-gram counts in which topic dependent words have been masked, there are (on average) significant differences in language use, both between discussions of different topics within each community, and between discussions of the same topic within different communities (van Halteren & Oostdijk, Submitted).

[6] As for widely focused tribes, we have shown differences in language use between the young and the old (van Halteren, 2015), as well as between men and women (van Halteren & Speerstra, 2014).

[7] Given that each investigation was quite extensive, we can only provide a summary in this paper, and will have to restrict ourselves to directing the reader to the individual publications for more details.

[8] With some exceptions, such as information feeds like job agencies and dating bureaus, which employ a more field-like structure in which spacing is sometimes omitted.

[9] Which input method is used can most often be deduced from the metadata in the Twitter JSON.

[10] The exception here might be voice input, and input method errors there should lead to more than just spacing problems.

the norm. We see both blanks left out and added where they are not needed. We investigated the extent of this phenomenon by manually annotating 1,000 tweets, randomly sampled from a year of tweets. In these tweets, we found some 300 cases of variant spacing in about 200 tweets. In most cases (60% of all variants), the variation was adjacent to punctuation. For processing, additional blanks in such contexts (36%) are completely unproblematic. Leaving out blanks where they are expected next to punctuation may sometimes lead to (mild forms of) ambiguity, e.g. where emoticons flow together with normal punctuation or where a word-period-word sequence might be mistaken for a URL, but generally this does not cause serious problems for processing.

More interesting are those cases where only words are adjacent to the variant spacing. In most cases where two or more words are merged (be it just glued together or fused more extensively), we found this is done deliberately (24%), possibly as a shortening mechanism. This is supported by the fact that there is quite some regularity here. We see that blanks are deemed superfluous within common bigrams, and that in many of these cases we see the formation of clitics (3%). In later investigations, we observed that even though cliticization occurs frequently the authors do seem to avoid ambiguity. As an example, *dat is* ('that is') can be shortened to *das*, and *dat ik* ('that I') to *dak*. Both of these shortened words are in the lexicon as an existing noun. *das* is both "badger" and "scarf" or "tie"; *dak* is "roof". Now, the alternative interpretations of *das* are needed much less frequently (in the case of scarf also because of more often used alternatives) than those of *dak*, and this difference is reflected in the usage of the shortened forms: if we examine the forms which are closest in terms of context vectors (using a window of two tokens left and two tokens right; cf. van Halteren, In prep.), *das* gives us a top-5 with *da's*, *dat's*, *dats*, *datis*, and *dass*, proving active use of the clitic, but *dak* gives us the top-5 *dakkie* (vernacular diminutive of *dak*), *balkon* ('balcony'), *dakje* (official diminutive of *dak*), *plafond* ('ceiling'), and *aanrecht* ('sink'), showing the clitic here is apparently shunned.

Such deliberate spacing variations can be lexicalized in the language use of specific tribes, leading to a situation much like that for spelling (Section 4). As an example, in one user community, we observed that the combination *maar ja* (lit. 'but yes', i.e. 'but well') was practically always written without a space; interestingly, the initial form *maarja* was over time more and more replaced by the even shorter *mja*.

In other cases of spacing variation between words (11%), we assume that the user is ignorant of the norm for spacing, e.g. when components of separable verbs are adjacent, or with compounds (which in Dutch should be written as a single word). Other categories of words where variant spacing is found include names, archaic forms, and words containing prefixes. In ignorance-related cases, the variation is typical for the user, but it sometimes propagates through conversations.

Finally there are cases (2%) where we did not identify any (apparent) regular system underlying variant spacing, and which might therefore just be typos.

Even though the majority of spacing variants appear to be resolvable, we think that here lies the hardest problem for proper processing, especially if one intends to use the traditional NLP architecture where tokenization is addressed in a separate preprocessing step.

4. Spelling

Regarding the spelling used by Twitter users, a random selection of tweets also tends to give the impression of almost random noise. However, if we investigate the data more extensively, and apply some classification, we start seeing patterns.[11]

As with orthography, there are large numbers of tweets, produced in a professional context or in the context of serious discussions between adults, where spelling usually conforms to the accepted norms for written language. Virtually all spelling deviations here are caused by typos; only in very few cases users appear to opt for a form of creative spelling. In some contexts, we do see extensive use of foreign words, but these too tend to follow standard spelling and topic-specific lexicons could be created. Alternative spellings are mostly found with younger and/or less educated users. But here too, we have the impression that each group of users mostly uses its own lexical and morphological conventions, picking mechanisms from the repertoire we describe below. Once we have determined what tribe we are dealing with, we can select the corresponding lexicons and rules for processing.

As already mentioned, there appears to be a fixed repertoire of mechanisms to vary spelling. However, before the discussion of this repertoire, we will first exemplify the level of variation with the word *school* ('school'), which we investigated when working on various techniques for modeling spelling variation. Table 1 shows the most frequent forms derived for school with a word form clustering algorithm using form relation information based on both contextual similarity and edit distance calculated with the Viterstein algorithm (van Halteren & Oostdijk, 2012). Figure 1 shows the forms that were only suggested for a single text instance to be connected to the same cluster. Apart from the forms shown, there were many more, leading to a total cluster of 507 forms. It should be noted that these 507 do appear to contain some false positives. In Table 1, we see the plural form *scholen*, as well as some other nouns with similar spelling, such as *schoot* ('lap').[12]

[11] For more quantitative information, and a description and evaluation of an early technique for spelling normalization for Dutch tweets, see van Halteren & Oostdijk (2012).

[12] Although *schol* is also a kind of fish ('plaice'), we do not think this should be counted as a false positive, given the distribution of discussion topics on Twitter.

8585 schooll	1643 sgoool	740 sgl	383 schoooooool	187 schooooooool
6245 schooool	1637 schooooool	637 schoel	340 chool	179 schooooll
5468 sgol	1403 sschool	627 sgooll	277 schoop	169 sqool
5412 schol	1119 sxhool	549 scholl	276 skoool	161 scho
4926 shool	1011 schoolll	542 svhool	269 dchool	161 schoiol
3964 schoool	988 achool	529 schoot	260 schoolo	160 schoolx
3955 schoooool	981 scool	514 shcool	245 schooolll	159 schoola
3644 schhool	964 scholen	500 schoorl	231 schoor	156 sgoowl
2451 schhol	891 scchool	448 schoowl	220 schoolt	150 schiol
2410 schoooll	866 sgool	437 scgool	219 schoof	147 schhoool
2345 schook	768 schoolk	393 skool	214 schoolie	145 schiool
2323 schoo	754 sjool	389 schooo	204 schok	143 schoolen

Table 1. Most frequent spelling variants for ‘school’, as suggested by a system built on the principles explained by van Halteren & Oostdijk (2012). The numbers represent the number of instances of the form for which the system suggested the normalized form ‘school’.

achol aschol dcholl dnsschool echschool eschool hagol higchool higschool hughschool oschool pschool rschool sachoool sccchoool scchok scchooool scchoot scghoool scgoll schaol schgoool schhhooooll schhlool schhok schhoo schhoolo schhoooollll schhooooon schill schjooll schlll schlol schlool schoeonen scholk scholll schollol schollos schooa schoog schoohol schoohoon schoohooon schookll schoolh schoolkl schooloe schoolof schoolollololollooollolllo schoolp schoolschool schoolse schooltl schoolzl schoolzn schoont schooohl schooola schoooohooool schoooolen schoooollll schooooohooool schooooolen schoooooollllll schooooooon schooooooooon schooooooooooooooool schooooooooooooooooooool schooop schoow schorel schorn schosol schotel schuool scoll scoolh scooool sggoo sghhoog sgiool sgoil sgoof sgoohool sgookl sgoolc sgooloo sgooollk sgooon sgoooooool sgpool sgvool shcooool shoooool sichool siol sjoooool skoooooooooool sochool sohool sschhool sschoooon sschoooooool sschoot ssssschool svhooll sxcholll sxool vschool wegschool

Figure 1. Spelling variants suggested for only one instance in our data set by a system built on the principles explained by van Halteren & Oostdijk (2012).

In Figure 1, there are more false positives, mostly similarly spelled forms that have been attracted to the cluster by the relative frequency of school (e.g. *shoohooon* is more likely *schoon* (‘clean’), and specific types of school (e.g. *higschool* is probably meant to be ‘highschool’). All in all, the precision appears to be very high.[13] Table 2 shows the twenty most similar forms in terms of context vectors based on a window of two tokens left and two tokens right, and using a larger data set than in the one used in the previous study (van Halteren, In prep.). We see mostly the same variants, but now in a different order, namely similarity instead of frequency. The order appears to distinguish between intentional variants, such as *sgool* and *schooooool*, which appear to be slightly more distant in context from *school*, and typos, such as *shcool* and *schook*, which are found in much the same contexts as *school*. Notably missing in the top-50 is *sgl*, but closer inspection shows that this is because in 2013 and 2014, there was an extensive discussion about a financial fraud by the director of an institute called SGL, which had repercussions for the measurements underlying Table 2, but not Table 1 as that reflects data up to 2012. Notably added in Table 2 are the forms *scorro/skola* and their variants. These are street language words for school and should therefore be seen as synonyms rather than spelling variants.

One of the more noticeable mechanisms for variation is actually used to attach additional information to the words themselves, namely repetition of individual characters or strings of characters. Such repetition signifies stress, and is a written kind of prosody. When repeating longer substrings, stressed words do become more prone to typos, but given the regular repeating pattern, resolving typos should be relatively easy. Repetition is productive rather than lexicalized, but can be handled as a morphological process, as demonstrated with the Viterstein algorithm (van Halteren & Oostdijk, 2012).

There are a number of other conscious variation mechanisms. First, we see various methods of shortening the text. Shortening is possible, for example, by clipping forms, e.g. *eig* for *eigenlijk* (‘in fact’), replacing the full form by an acronym, e.g. *pww* for *proefwerkweek* (‘exam week’), vowel deletion, e.g. *gwn* for *gewoon* (‘just’), or using rebus-like forms, e.g. *w8* for *wacht* (‘wait’).

[13] Obviously the data set is far too big to measure recall.

0.7270 shool	0.6848 achool	0.6487 schoowl	0.6015 schoolll	0.5798 schooooool
0.7082 sschool	0.6828 schhol	0.6485 scholl	0.6014 sqool	0.5738 scoro
0.7018 schhool	0.6808 schoolk	0.6483 schol	0.6012 skoele	0.5618 schooolll
0.6993 scchool	0.6798 scgool	0.6464 sgol	0.5985 schooool	0.5614 schooooooool
0.6952 shcool	0.6679 schooo	0.6308 schooll	0.5983 scorro	0.5591 schooooooool
0.6929 scjool	0.6657 schoool	0.6259 schhoool	0.5981 sgooll	0.5588 schooooll
0.6929 svhool	0.6624 schoo	0.6201 skola	0.5940 skorro	0.5575 scola
0.6905 schiol	0.6596 schoop	0.6104 sgoool	0.5876 schoooll	0.5447 scorroo
0.6875 schook	0.6580 sgool	0.6055 skoool	0.5825 schoooool	0.5440 scoroo
0.6855 sxhool	0.6488 schoolie	0.6046 skolla	0.5820 schoollll	0.5413 scho

Table 2. Most similar forms to the word form 'school', as calculated on the basis of all instances of each form with a text window of two tokens left and two tokens right (van Halteren, in prep.). The numbers represent cosines between the context vectors of school and of the form in question, with the vector dimensions being PMIs between the word form and the context.

Many shortened forms are already quite lexicalized. Shortening is mainly meant for efficiency, but the exact type of shortening used is often indicative of a (confirmed or desired) group membership of the user. Again, generally, users avoid ambiguity, but such avoidance is in the context of the tribe communicated with, and shortened forms may well have other meanings in other contexts, implying that modeling shortening mechanisms, including lexicon formation of lexicalized forms, should be done within the contexts in question.

Another frequent conscious variation is phonetic writing. Here, we also see effects mirroring reduction in speech, in Dutch e.g. *n*-deletions, so that it too can sometimes serve as a shortening mechanism. Phonetic writing is even more an indication of tribe membership and/or user characteristics like the regional background of the user, and can therefore be used to much effect in processing, in the sense that selection of the proper tribe model is more likely to be successful.

There are also instances where spelling variation resulting in deviation from the standard norm is unintentional, and which are traditionally grouped as spelling errors. Here we should distinguish between typographical errors, i.e. errors caused by mismanipulation of the input device, and orthographical[14] errors, i.e. errors caused by lack of knowledge of the correct spelling.[15] How to model typographical errors has been studied extensively, but mostly for traditional text types. The degree to which these errors can be modeled in the Twitter context depends on how regular they are for a specific user, and on the input device used. We may be able to recognize which input device has been used on the basis of the Twitter metadata, or possibly by other effects in spelling and orthography, which could facilitate the recognition of the intended word. For example, there is a higher likelihood of substitution of characters by an adjacent character on the keyboard (e.g. *schook* instead of *school*), but the usefulness of this observation depends on whether a keyboard is used at all, the keyboard layout, and the key selection method.[16] Orthographical errors are more user related, and are often similar to phonetic writing. Here it is the recognition that the user belongs to a specific tribe that can help identify the intended word.

5. Syntax

Considering the previous sections, one might expect the use of syntax in the more professional and "serious" tweets to conform to the norms for standard Dutch, and a more chaotic throwing together of words by the more adventurous users. However, this is in fact unlikely. After all, a reader can be expected to cope with a bit of variation in spelling and orthography, and the author can probably judge what is still comprehensible. To come up with a syntactic structure which is non-standard, but still able to convey the intended message to one's readers is much more difficult, which means that most users can be expected to simply choose (consciously or subconsciously) from their available standard repertoire of syntactic structures.

This assumption is confirmed by an investigation of sets of tweets representing various discussion topics (Oostdijk & van Halteren, 2016). In four topic areas, we took eight related hashtags and, for each hashtag, investigated a random sample of 100 tweets.[17] We (manually) split each tweet into parse units and annotated each parse unit for its syntactic category, e.g. full declarative sentence, elliptic declarative sentence, interrogative sentence, noun phrase, etc., and then examined the distribution of these categories.

[14] This is the term traditionally used in research on spelling errors. Note that our use of the term 'orthography' in this paper is different.

[15] Related are errors against morphology, such as erroneous past participle formation, which we will not analyse here.

[16] An additional complication here is caused by the fact that many of the possible input methods for tweets contain 'user friendly' components adjusting words to what they should be according to the method's statistics, and that users most often do not invest in correcting unwanted adjustments. In such cases, it will be much harder to use knowledge of the input method to reconstruct what the user meant.

[17] For a more detailed analysis, and quantitative information, see the already mentioned Oostdijk & van Halteren (2016).

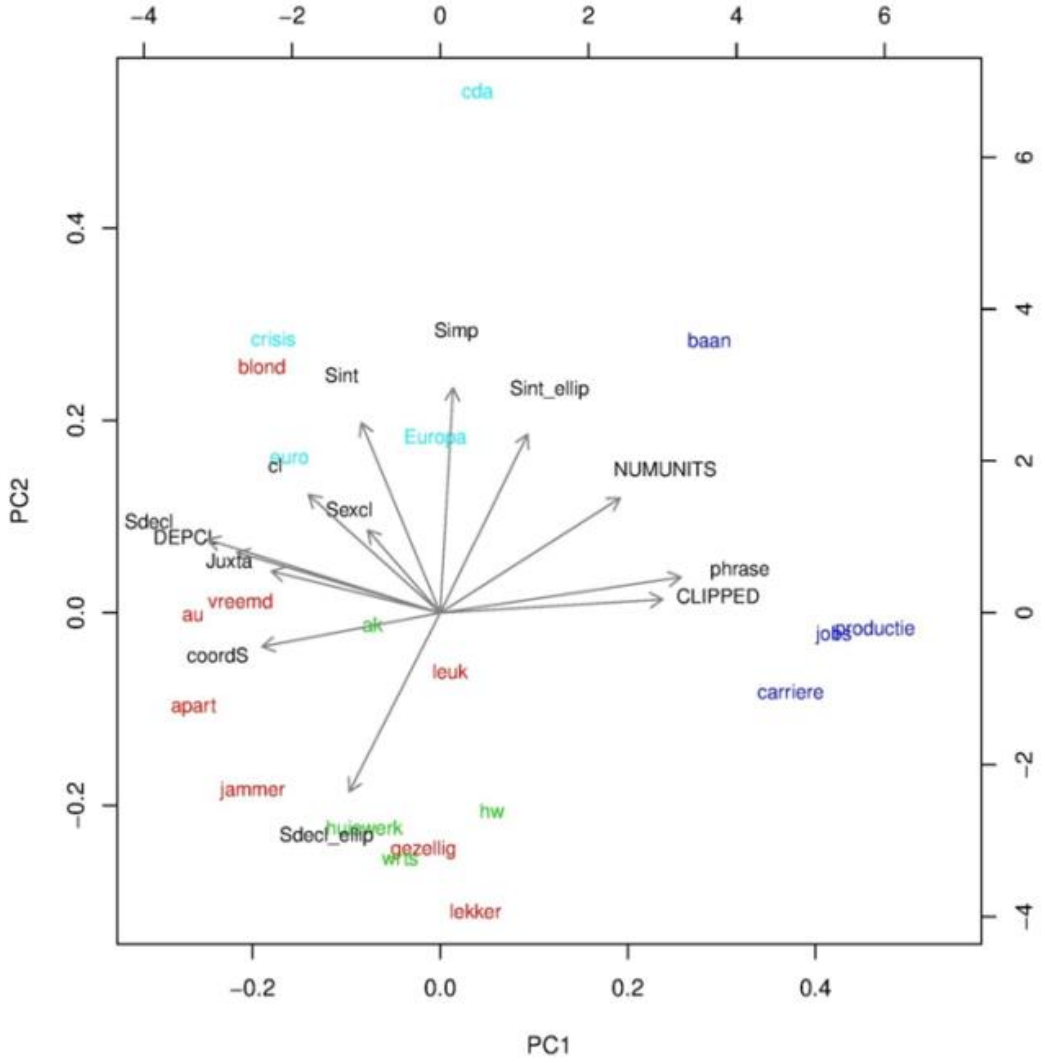

Figure 2: Biplot showing the placement of investigated hashtags and individual annotations in relation to the first two principal components. The hashtags are marked with the cluster colours: 'school' (green), 'employment' (dark blue), 'politics' (light blue), and 'appreciation' (red).

A quick impression of the difference between the clusters can be gleaned from Figure 2, which shows a principal component analysis on the basis of the frequencies of the various parse unit annotations (Oostdijk & van Halteren, 2016).

On the "serious" side of Twitter, we looked at tweets about politics, with hashtags referring to e.g. political parties and political issues. Here, we indeed found mostly full sentences, following standard syntax. To see the extent of variation elsewhere, we targeted tweets where we expected the most severe variation, namely tweets about school, with hashtags referring to e.g. homework and school subjects. Here, we saw a very high frequency of elliptic structures, but most all (over 95%) of the structures encountered were taken from the standard repertoire.[18] The third cluster targeted another extreme, namely the job market, with hashtags referring to e.g. vacancies and career development. Here, we saw a more telegram-like style of communication, trying to pack as much information as possible into the limited space by foregoing sentence structures and replacing them by (sometimes long) sequences of phrases. The phrases, though, followed a standard structure. The final cluster, called "appreciation", was built around hashtags consisting of adjectives expressing an opinion.

In Figure 2, we show the result of a principal component analysis based the frequencies in which the various annotations were assigned in tweets with the various hashtags (Oostdijk & van Halteren, 2016). On the horizontal axis (PC1), we see the distinction between normal clausal structure and phrase stringing, with "employment" favouring the latter and all three other

[18] We do not know whether this observation can be generalized to all tweets, as only tweets with hashtags were included here. We have seen in previous research that tweets without hashtags are more irregular in the sense that they contain more OOV-words (van Halteren & Oostdijk, 2014). We do not know (yet) whether their syntax is also more irregular.

clusters favouring the former. On the vertical axis (PC2), we see the differences in applying the clausal structure, with "politics" mostly adhering to full structures, and apparently also more use of interrogatives and imperatives, and "school" showing much more ellipsis. "appreciation" is spread out over PC2, but is clearly in the clausal camp on PC1. All in all, there are clear differences between topic clusters, but there is also substantial variation within the clusters, implying that widely focused tribal models should already help processing, but that more narrowly focused ones can improve the modeling quality even further.

It would seem that the syntax of tweets can be modeled using much the same methods as for traditional text, at least once the variation in the lower levels of analysis (orthography and spelling; see above) has been accounted for. When using probabilistic methods, however, we would do well to derive probabilities per tribe. Furthermore, such probabilities might also serve to recognize which model should be used for a specific tweet or conversation.

There is one additional complication in the area of syntactic analysis. In some cases,[19] especially when information is forwarded, the text may be clipped, usually marked with an ellipsis sign (…) and a URL. These cases are therefore easy to recognize, but the clipped text is irretrievably lost.[20]

6. Conclusion

In the previous sections, we looked at the wide (and frequent) linguistic variation in the language use on Twitter. Most of this we judge to be intentional, and to be related to the conventions used in the peer group the author belongs to, or would like to belong to, in specific types of communication about specific topics (i.e. what we call *Twitter tribes*). Another source of variation is the author's idiolect, sometimes with clear influences from his/her sociolect. Finally, variation may be caused by mismanipulation of the input device.

All three of these causes are such that we can expect the variation to show a substantial amount of regularity, which means that it can be modeled and that the derived models can be employed in a noisy channel model approach to the normalization of tweets. For various linguistic levels, we have shown the most important processes that constitute the noisy channel. We judge that they can indeed be modeled.

Obviously, we are not the first to suggest a noisy channel model approach. Traditional approaches to (contextual) spelling correction tend to think in terms of noisy channel models (e.g. Dutta et al., 2015) and there is also already experience with applying statistical machine translation techniques for text normalization (e.g. Limsopatham and Collier, 2015). However, we think that this approach is vulnerable because of the heterogeneity of Twitter, and stands to benefit from modeling the patterned variation we see in the language use of tribes.

Taken to its extreme, our proposal would imply that we need to train billions of individual models, which includes finding sufficient training data for each of them. However, as far as we can see, there are gradual rather than radical differences when comparing closely related tribes. We therefore propose to build models for clusters of tribes (which in principle are by themselves also tribes) and use weighted combinations when operating the noisy channel model.

In the near future, we aim to test our proposal. We intend to implement a system that can identify the appropriate tribes (characteristics of author, topic and use situation) for a tweet. In parallel, we will complete our system for linking variant spellings of a word form to a consensus form. [21] Once these are in place, we can evaluate whether tribal modeling indeed outperforms global modeling.

References

Baldwin, T. (Timothy), Cook, P., Lui, M., Mackinlay, A. & Wang, L. (2013). How Noisy Social Media Text, How Diffrnt Social Media Sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*. Nagoya, Japan, 2013. Pages 356–364.

Baldwin, T. (Timothy), de Marneffe, M., Han, B., Kim, Y., Ritter, A. & Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text (WNUT 2015)*, Beijing, China.

Baldwin, T. (Tyler) & Li, Y. (2015). An In-depth Analysis of the Effect of Text Normalization in Social Media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado. Pages 420-429.

Bryden, J., Funk, S. & Jansen, V. (2013). Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science*, 2013, 2:3.

Dutta, S., Saha, T., Banerjee, S., & Naskar, S.K. (2015). Text normalization in code-mixed social media text. In *Proceedings the IEEE 2nd International Conference om Recent Trends in Information Systems*, 9-11 July 2015, Kolkata, India. Pages 378-382.

Eisenstein, J., O'Connor, B., Smith, N. & Xing, P. (2014). Diffusion of lexical change in social media. *PLOS-ONE*, 9, 11 2014.

Han, B., & Baldwin, T. (Timothy) (2011). Lexical normalisation of short text messages: Makn Sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

[19] In our sample discussed here, as much as 7% of the tweets.

[20] At least within the tweet. It may be present at the URL mentioned, but recovery in such cases is outside the scope of this paper.

[21] The choice for spelling is pragmatic rather than optimal for our goal. We expect that the highest quality gain can be reached in syntax, but a syntactic analysis system is far more difficult to build, if this is even possible as long as the spelling variation is not resolved.

Portland, Oregon, June 19-24, 2011. Pages 368–378.
Kaufman, M., & Kalita, J. (2010). Syntactic normalization of Twitter messages. In *Proceedings of the International conference on natural language processing*, Kharagpur, India.
Limsopatham, N. & Collier, N. (2015). Adapting Phrase-based Machine Translation to Normalise Medical Terms in Social Media Messages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
Oostdijk, N. & van Halteren, H. (2016). Twitter Tribal Languages. In: *Handbook of Twitter for Research* (Proceedings CTR2015).
Tjong Kim Sang, E. & van den Bosch, A. (2013). Dealing with Big Data: The case of Twitter. *CLIN Journal* Vol. 3, 121-134.
van Halteren, H. (2015). Metadata Induction on a Dutch Twitter Corpus: Initial phases. *Computational Linguistics in the Netherlands Journal*, Vol. 5. 37-48.
van Halteren, H. (in prep). Word similarity in Dutch tweets. To be submitted to *Computational Linguistics in the Netherlands Journal*, Vol. 6.
van Halteren, H. & Oostdijk, N. (2012) Towards Identifying Normal Forms for Various Word Form Spellings on Twitter, *Computational Linguistics in the Netherlands Journal*, Vol. 2. 2-22.
van Halteren, H. & Oostdijk, N. (2014). Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens. *Journal for Language Technology and Computational Linguistics*, Vol 29(2). 97-123.
van Halteren, H. & Oostdijk, N. (2015). Word Distributions in Dutch Tweets. *Tijdschrift voor Nederlandse Taal- en Letterkunde*. 2015/3. 189-226.
van Halteren, H. & Oostdijk, N. (Submitted). Twitter language model differences between topics and between communities. Submitted to ACL2016.
van Halteren, H. & Speerstra, N. (2014). Gender Recognition on Dutch Tweets. *Computational Linguistics in the Netherlands Journal*, Vol. 4. 171-190.