Computational Linguistics in the Netherlands Journal 8 (2018) 138–162 Submitted 06/2018; Published 12/2018

Tweet geography. Tweet Based Mapping of Dialect Features in Dutch Limburg.

Hans van Halteren B.VANHALTEREN@LET.RU.NL
Roeland van Hout R.VANHOUT@LET.RU.NL
Romy Roumans ROMY.ROUMANS@STUDENT.RU.NL

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Abstract

We investigated whether tweets can be used to map dialect features (such as pronunciation or lexis) in the Dutch province of Limburg and, if so, how the resulting maps can be interpreted. We developed a mapping procedure based on the relative frequency of dialect variants of individual Twitter users and the relative frequencies of their geographically neighbouring Twitter users. We evaluated this procedure by comparing the geographical locations of written dialect variants retrieved from Twitter with the isoglosses and dialect regions known from dialectology. The results show that Twitter can indeed be a good source for dialect studies, when applied with some caution, to track new patterns of dialect variation caused by dialect shift and loss, internal migration within Limburg and the immigration of non-dialect speakers. Next, we compared, for the same Twitter data, this knowledge-rich approach (known dialect variants) to a knowledge-poor approach (letter trigrams). Here we found that trigram counts show strong correlational overlap with dialect variant counts, but the exact relation between the two needs further study.

1. Introduction

A high number of inhabitants of the Dutch province of speak a Limburg dialect.[1] A survey in 2016 of the Limburg newspaper *Dagblad De Limburger* reported that 72% of the Limburg inhabitants are able to speak dialect. This amount is far higher than what is reported for other Dutch dialect regions. Another interesting aspect of the Limburg dialect area is the variety in local dialects. The map in Figure 1 shows six areas (Keulen et al. 2007). The two northern areas belong to the so-called Kleverland dialect area. The other areas can be subsumed as being South Lower Franconian, with the exception of the deep south-eastern Ripuarian region.

In the same survey, people in Limburg reported that they use dialect in written form on the social media: 56% on Whatsapp, 50% on Facebook and 27% on Twitter. These percentages are high, given that dialect is primarily a spoken and not a written language. Moreover, most people are not trained in any form of dialect spelling and normally they use forms of spontaneous dialect spelling based on Dutch orthography. They do not use the diacritics that are abundantly present in official dialect spelling systems, based on the so-called Veldeke spelling system 2003 (http://www.limburgsedialecten.nl/download/spelling2003.pdf).

The percentages from the survey suggest that it might be feasible to mine social media for information on dialect use, as the dialectal pronunciation is also visible in the written form. If this were not the case, the only dialect features which could be observed would be lexical or syntactic in nature. Furthermore, only variants which are in alternation can be reliably measured and not subject to variation due to the content of the texts. With pronunciation features also visible in the text, and visible pronunciation variant automatically in alternation with the standard form, we have a much wider scope for mapping the dialects.

1. We only looked at Dutch Limburg. We left out the province of Belgian Limburg for two reasons. Dialect use is falling down much more rapidly in Belgian Limburg and the number of tweets from Flanders in our dataset is far lower than from the Netherlands.

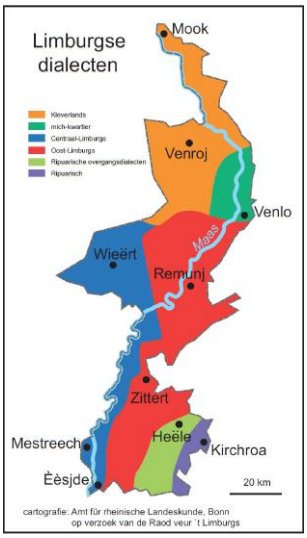

Figure 1: Map of the Limburgian dialects

If we are indeed able to measure the use of dialect features in social media, we can derive data that are much more detailed and represent many more informants than what is collected in more traditional dialect surveys and that we might get a better view on more recent geographical and sociolinguistic patterns in the distribution of Limburg dialects. In this paper we describe how we investigated a specific social media platform, Twitter. We explored whether it can indeed be mined successfully for dialect data, by investigating regional patterns and comparing these patterns to isoglosses and dialect regions known from dialectology and by comparing it to population developments in the Limburg area.

Furthermore, we examined whether dialect-related information can be detected on Twitter without previous knowledge about dialectal forms, by measuring and mapping the relative use of letter trigrams, and then correlating the outcome of this knowledge-poor approach to the mappings we obtained from the former, knowledge-rich approach. We expected mixed results here. Trigram frequencies may vary with the region either because they contain some spelling variant linked to dialect pronunciation or because they are linked to (normal) words which are used more often in specific areas (and potentially for both reasons at the same time). Whether trigrams can bring a useful contribution to our goals depends on the relative strength of the two causes and their distributions.

Other studies have used Twitter data to study geographical patterns of lexical variation and change. Rahimi et al. (2017) and Eisenstein et al. (2014) developed methods to map the use of lexical innovations in Twitter. Grieve (2013) presents a statistical comparison of common patterns of regional phonetic and lexical variation in American English based on the results of two previous dialect studies. His approach is more relevant for us because he focused on existing dialects. He used a general method to make dialect maps that are based on different sets of locations comparable, the so-called *kriging* interpolation technique.

Although some of the previous studies already made use of Twitter, we still see our viability test of such use as our main focus. We do extend this focus with two areas of special attention, being the comparison between the knowledge-rich and knowledge-poor approaches and a more fine-grained geographical representation. The latter also forced us to introduce innovative smoothing and plotting techniques. Finally, our methodology is somewhat tighter than (most of) the previous studies, especially where it concerns data clean-up and (for the knowledge-rich approach) a strict

Number of tweets in period	Number of users
50-99	448
100-499	1,925
500-999	1,113
1,000-4,999	2,759
5,000-9,999	622
10,000+	334

Table 1: Tweet volumes of 7,201 Limburgian Twitter users, linked to a location by means of GPS codes in their tweet metadata. The first column lists the range of the volume, in number of tweets, from January 2011 to March 2017.

adherence to measuring relative frequencies within alternations, which is a well-proven standard in sociolinguistics.

In the following sections, we first describe our experimental data (Section 2) and the mapping algorithm we used instead of kriging (Section 3). Next, we proceed to a comparison between the Twitter-derived data and existing knowledge about Limburg dialects (Section 4). We continue with a discussion of the viability of a knowledge-poor approach (Section 5). We close with a more general discussion and conclusion (Section 6).

2. Dialect Data

As the data source for our dialect data, we used Twitter. More exactly, we used data from the TwiNL data collection (Tjong Kim Sang and van den Bosch 2013) with time stamps from January 2011 to March 2017,[2] selecting users who produced at least fifty tweets in that period (about 4.8 million user accounts). From this data we wanted to select all tweets from users that lived in the Dutch province of Limburg. As we are hesitant to rely on user-supplied locations, we looked for tweets that had a GPS location code in their metadata (about 320 thousand accounts). In order for a user to be accepted as an inhabitant of Limburg in our experiment, we set several criteria.

First of all, two thirds of the GPS locations of a user's tweets had to be within one standard deviation of the mean location, both in longitude and in latitude. If this was the case, then a new mean was taken from all locations in this range, which had to be within the borders of Limburg. Our intention was to pinpoint the users' "home base", most probably their home except in cases of users who tweet predominantly at their work place. For determining a user's dialect, the current home base is not the perfect predictor of course, but in this way, we selected persons living and/or working in Limburg. We decided to apply a second filter, as we had noticed in previous experiments that some users only switched on their GPS during holidays and were assigned to the wrong region. To overcome this problem, we demanded a minimum number of ten tweets with GPS codes, as well as tweets within range of the home base in three different months of the year. We intend to revisit these procedures in the future, with a eye to replacing them by more advanced statistical methods, but for the current paper, they appear sufficient.

From the remaining 7,769 user accounts, we removed a further 568 because they did not tweet predominantly in Dutch according to a language identification procedure (van Halteren 2015) and/or manual inspection.

2. Although this appears to be quite a long period if one would like to repeat this approach for other areas, it is the volume of tweets and of dialect use that is important rather than the time span. Depending on the exact research question and the exact level of presence of the investigated dialect use, one will have to determine case by case which time period should be sampled.

Alternation	Regular expression	Number of users	
		Standard form	Variant
ik → ich	^ik$; ^ich$, ^ig$	7,126	2,775
mij → mich	^mij$; ^mich$, ^mig$	6,588	2,263
jou → dich	^jou$; ^dich$, ^dig$	6,132	2,948
s- → sj-	^s[lmnpt][aeiou]; ^sjk[lmnpt][aeiou]	7,198	1,509
aa → oa/ao	jaar[^aeiuo], jar[aeiuo], ^(j|g|n|st)a+, aan[^aeiuo], laa[pt]t*$, la[pt]e, haalt*$, hale, [dnw]aar$; j(oa\|ao)r[^aeiuo], j(oa\|ao)r[aeiuo], ^(j\|g\|n\|st)(oa\|ao)+$, (oa|ao)n[^aeiuo], l(oa|ao)[pt]t*$, l(oa\|ao)[pt]e, h(oa\|ao)lt*$, h(oa|ao)le, [dnw](oa|ao)r$	7,195	4,125/1,939
ij → ie	^mijn$, lijk$, ^blijf, ^blijv, tijd, kijk, krijg, kwijt, wijn; ^mien$, liek$, ^blief, ^bliev, tied, kiek, krieg, kwiet, wien	7,162	4,270
ui → uu	^uit, huis; ^uut, huus	7,117	1,276
ui → oe	^uit, huis; ^oe[ts], hoes	7,117	4,623
niet → nie	^niet$; ^nie$	7,096	3,312
niet → neet	^niet$; ^neet$	7,096	2,508
niet → nit	^niet$; ^nit$	7,096	509
niet → neit	^niet$; ^neit$	7,096	888
dat → da	^dat$; ^da$	7,092	3,377
dat → dea	^dat$; ^dea$, ^dae$	7,092	1,294
dat → det	^dat$; ^det$	7,092	1,592

Table 2: Alternations between standard Dutch word forms and dialect variants known to occur in specific Limburg dialect regions. There are three pronominal alternations, followed by five pronunciation-related alternations, and, finally, two frequent function words (*niet* is the negation "not"; *dat* has the same functions as the English word "that") with respectively four and three dialect alternations.

The result was a set of 7,201 Twitter users who tweeted predominantly in (potentially dialectal) Dutch and who could be linked with a reasonable degree of confidence to a location in Dutch Limburg. Table 1 gives an indication of the tweet volume of these users.

For each selected user, we counted the frequencies of various words and character n-grams. In the knowledge-rich approach, we focused on fifteen dialect alternations which are known from the literature on Limburg dialects (cf. Schrijnen 1920; Keulen et al. 2007; Bakker 2017). They are listed in Table 2, together with the regular expressions to find (reasonably unambiguous) instances. The left side of the alternations contains the standard Dutch form and the right the dialect variant. The table also shows how many of our 7,201 speakers used the forms in question. The number of Twitter users using the dialect variant is substantial. However, given that all standard forms are used by almost all the speakers, most of the dialect users must be alternating between dialect and standard language (code-mixing).

For each alternation we computed, per Twitter user, the relative frequency of the listed dialect word forms by dividing their frequency by the total frequency of the relevant listed words (standard forms plus dialect forms). In addition to investigating each alternation by itself, we also applied a principal component analysis with varimax rotation (Kaiser 1958; we used function `principal` from the package `psych` (Revelle 2018) in `R` (R Development Core Team 2008)) to the set of fifteen measurements and investigated the first three components (together explaining 75.4% of the variance). We selected three factors as the fourth factor had an eigenvalue of almost 1 (1.03) and, in addition,

there was a drop in eigenvalue between the third and fourth factor.[3] We rotated the three-factor solutions, applying varimax.

In the knowledge-poor approach, we avoided all previous knowledge (apart from the fact that pronunciation will mostly be shown in the written letters) and we simply counted the frequencies of all (in total 9,216) trigrams that consist of letters (mapped to lower case) and spaces (representing sequences of non-letter characters), and that are present in at least twenty tweets and produced by at least ten different users. We wanted to investigate whether these trigrams would also capture the variation fount to be related to the Limburg dialects. As the study of individual trigrams is unlikely to be fruitful, we decided to focus on the principal components of factor analysis, again after selecting the relevant number of components and varimax rotation. Here we used the first five components, even though they only explained 24.6% of the variance, as we were hesitant to apply varimax to more than five components. This choice was supported by the drop in eigenvalues after the fifth factor.[4] The number of factors with high eigenvalues is large because of the enormous number of trigrams. In the analysis below (Section 5) we do come back to individual trigrams, as examples to show and interpret their geographical pattern, and as an aid in interpreting the meaning of the five components by way of investigating the trigrams that have high component loadings.

3. Making Maps

Traditionally, dialect maps are plotted with symbols to indicate which locations use which dialect word or variant. Sometimes a dialect map gives isoglosses when two dialect variants are geographically clearly delimited, or dialect areas often but not always based on one or more specific isoglosses. More recently, dialectometric methods are being applied where larger numbers of dialect words or features are handled simultaneously to construct maps with dialect areas, using cluster analysis and multidimensional scaling. Examples are the SOMVIS multivariate mapping (e.g. Guo et al. 2005; Huang et al. 2016) and GabMap (www.gabmap.nl; Nerbonne et al. 2011). Both traditional and dialectometric maps are based on using a restricted set of fixed geographical locations. We need an alternative to handle 7,201 Twitter users all with their own location, spread over the whole province of Limburg, sometimes concentrated in cities but sometimes in much more isolated spots.

To handle this geographical diversity, we developed an alternative way of mapping that is user-based. Users in our set are each assigned their own location as their own plot point, coloured in accordance with the measured dialectal or trigram intensity at that location. Since even neighbouring users could be expected to differ greatly in their degree of dialect or trigram use, we needed to smooth the measurements along the geographical dimension. In most geographical information systems (GIS), this is done with an advanced method called *kriging* (Isaaks and Srivastava 1989; Wackernagel 2010). However, for our data this method was not optimal.[5] Instead we took, for each user, the measurements of the user in question and the 99 nearest neighbours. These 100 measurements were each weighted by distance to the user in question. If the distance to neighbour X was $d(X)$ and the distance to the furthest included neighbour was $d(max)$, the weight for neighbour X was set to $(1 - 0.99 * d(X)/d(max))$.[6] The weighted measurements were then averaged to yield the smoothed measurement for the user in question.

On the basis of these smoothed measurements, we first produced dialect variant strength maps. The colour of each point ranges along ten colours from blue (lowest frequency of variant use, often zero, which means only standard language forms used) to red (highest frequency of dialect variant use). By taking the extreme observed values as end points rather than 0% and 100% variant use,

3. The eigenvalues for the first five factors were 7.57, 2.66, 1.26, 1.03, and 0.88.
4. The eigenvalues for the first ten factors were 845.3, 508.4, 344.6, 300.5, 264.0, 207.6, 179.8, 169.5, 166.7, and 148.5.
5. See Appendix A for a discussion.
6. The multiplication factor of 0.99 is included to guarantee that even the furthest neighbour is included, with a weight of 0.01.

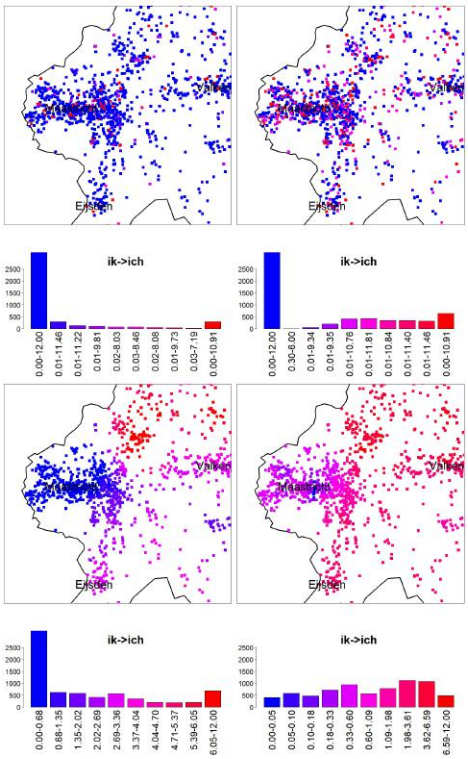

Figure 2: Example for our map making procedure: *ich* around Maastricht

and splitting the total range into ten equal blocks in the log scale, we achieved the clearest maps. We included a key for the interpretation of the colours in each map, in the form of a histogram for the various frequency bands. In Figure 2, we exemplify the map making procedure. In the top left pane, we show the raw percentage of the use for each individual user of *ich* within the *ich/ik* alternation for the area around Maastricht (but calculations and key are based on the whole of Limburg). We see that most speakers do not use *ich* and the ones who do use it, use it only part of the time. As a result, only the strongest cases are visible on the map. In the top right pane we base the colouring on a log scale. As a result, weaker use becomes more visible. To go from individuals to locations, we applied the smoothing procedure described above, leading to the two bottom panes. We can now recognize the spread of the usage of *ich* in both panes, but the clearer impression again comes from the right one, where the log scale was used. The advantage of using the log scale is also visible in the much more balanced histogram.

The fifteen final maps, i.e. with smoothing and log scale binning, for the whole of Limburg are shown with their keys in Figures 3 to 6.[7][8] The same strength mapping procedure was used for the principal components, using the factor scores of the users (generated by applying the function `predict` to the model and the smoothed user scores). However, the colour assignment is now done on the original scale rather than on the log scale, because of the different nature of the values:

7. In the main text and in tables, we refer by number to panes within figures. The leftmost pane is always number 1, and numbering continues to the right. Note that various figures have been turned sideways; in these figures the leftmost pane is shown bottommost on the page.
8. We feel that the maps are much clearer on the computer screen than on paper. Therefore we are providing them on our website. An index file with all URLs can be found at https://cls.ru.nl/staff/hvhalteren/clin2018/figure-index.txt

Alternation	Known isogloss	Twitter map
ik → ich	Uerdingen isogloss	Figure 3, pane 1
mij → mich	above Venlo, *mich* more northern than *ich*	Figure 3, pane 2
jou → dich	the same as for *mij→mich*	Figure 3, pane 3
s- → sj-	Panningen isogloss	Figure 3, pane 4

Table 3: Four well known Limburg dialect features

factor scores are normalized and do not have the skewed distributions with extreme values of the dialect alternations. Maps for the three first rotated components are shown in Figure 8. In the histograms accompanying these maps, we do not include legends for the bands, as the range for principal component values is not readily interpretable. Similarly, the maps for the first five principal components (after varimax rotation) of the measurements for the set of 9,216 trigram counts are shown in Figure 10.

4. Twitter as a Source for Dialect Data

In this section, we examine the viability of using Twitter to investigate the geography of dialect feature distributions. We start (Section 4.1) by comparing the maps for three lexical alternations and one phonological alternation against isoglosses that are generally accepted and widely supported by data from traditional dialect studies (Schrijnen 1920; Keulen et al. 2007; Bakker 2017). Then we examine eleven additional alternations known from earlier dialect studies (Section 4.2).

4.1 Verifying against four accepted isoglosses

We investigated four dialect forms analysed earlier by van Halteren and van Hout (2017): three pronominal forms with lexical variants plus the palatalisation of *s-* in word initial consonant clusters (see Table 3). Two famous Limburg isoglosses are involved: the Uerdingen isogloss (the dotted line in Figure 7, left pane; Keulen et al. 2007, page 171) and the Panningen isogloss (the red line in Figure 7, right pane; https://nl.wikipedia.org/wiki/Panninger_linie).

Our *ich* map (Figure 3, pane 1) shows several appealing results. The blue colours in the north of Limburg turn purple in the area of Venlo. This change in colour is exactly as the Uerdingen isogloss would predict. Weert and Nederweert show the use of the *ich* form, although Nederweert shows a stronger use of it. This might be explained by the frequency of dialect use, which probably is higher in Nederweert. In the area of Venlo we see geographical differentiation, which indicates that the dialect boundaries in Twitter have a more scattered and diffuse character potentially caused by mobility and migration effects. Furthermore, Venlo (*ik*) and Tegelen (*ich*) merged over time, which also caused dialectal agglomeration. Maastricht shows internal differentiation, with a blue spot that suggests a lower use of dialect by Twitter users in the center of Maastricht, potentially because of the presence there of various academic and commercial areas. The eastern mine area (Heerlen, Kerkrade, Landgraaf, Hoensbroek, Brunssum) displays some blue spots, a more negative outcome regarding dialect use than we expected. The use of dialects in this area seems to be less intensive than in the middle and other southern areas in Limburg. Overall, we can conclude that the Uerdingen isogloss can be retraced in the Limburgian Twitter users and their tweets, though more diffuse (mobility and migration of dialect speakers) and less outspoken.

In our *mich* map (Figure 3, pane 2) the same pattern emerges, except for the extension of the Venlo area to the north. The red colours north of Venlo nicely represent the so-called *mich* subarea (south of the full line in Figure 7, pane 1), with Horst (purple) not being part of this subarea. The maps of *mich* and *dich* (Figure 3, pane 3) are remarkably similar. These two forms largely have the same dialect distribution (see Bakker (2017) for a discussion of the *mich* and *dich* isoglosses north

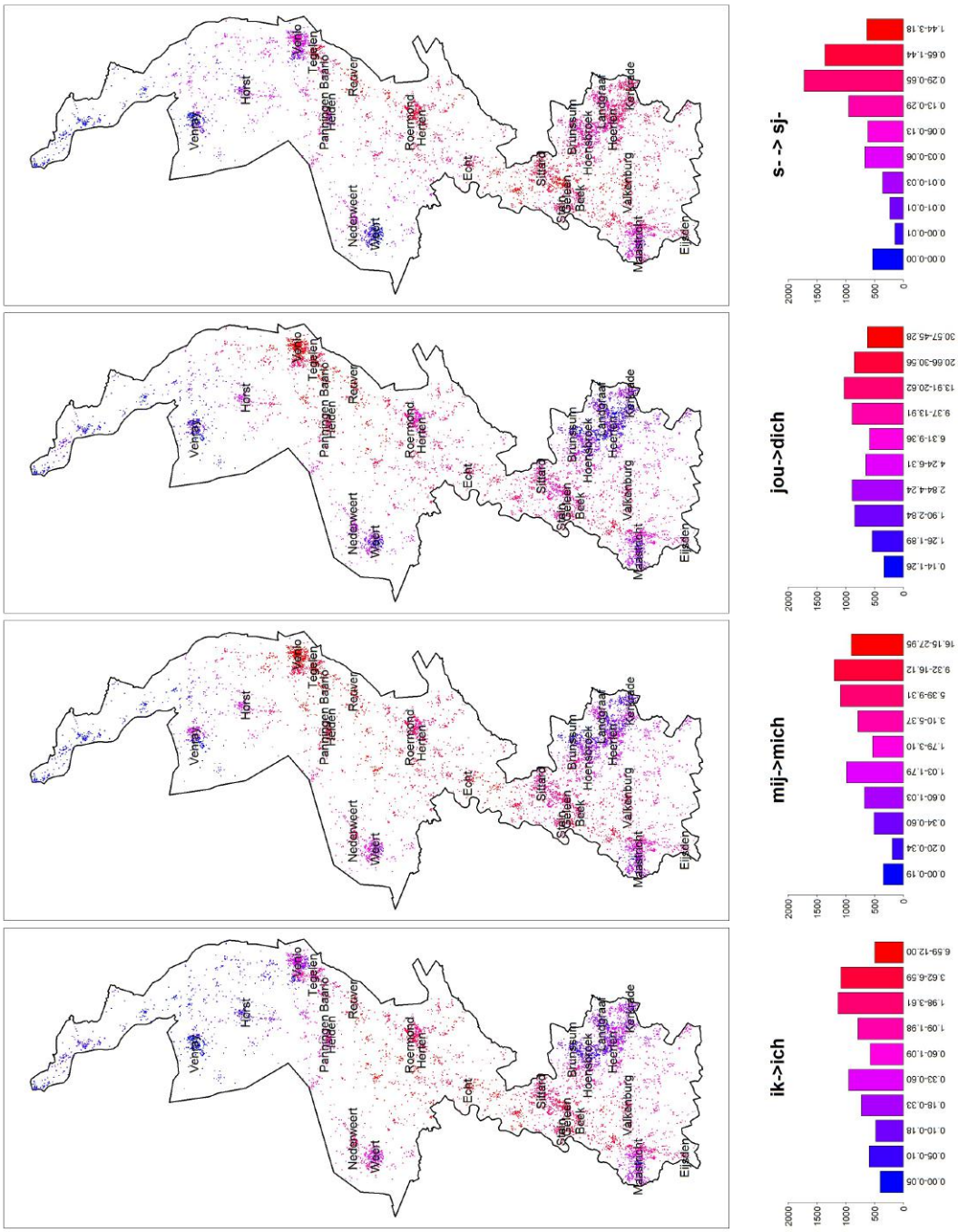

Figure 3: Alternation strength maps for *ik→ich*, *mij→mich*, *jou→dich* and *s-→sj-*

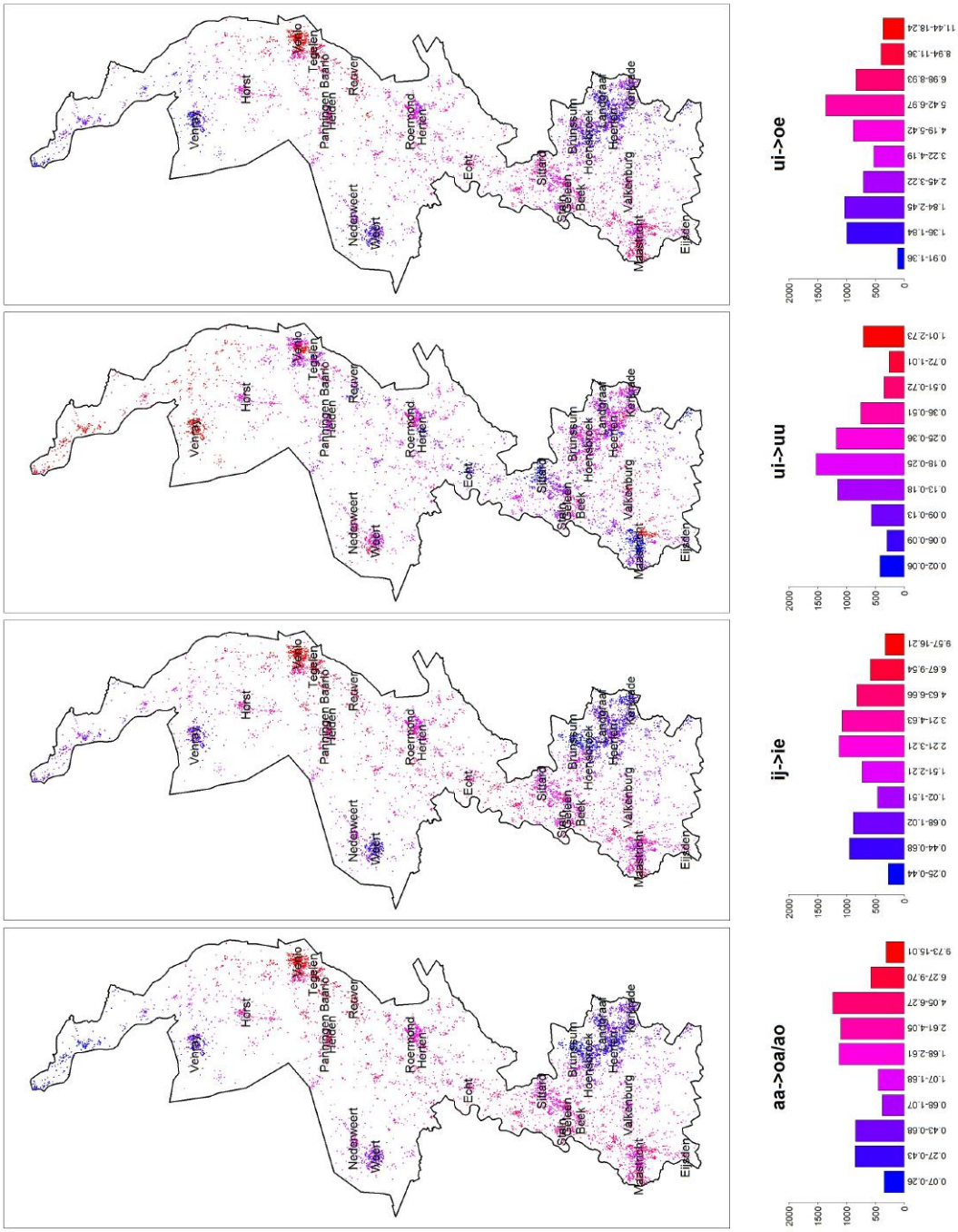

Figure 4: Alternation strength maps for *aa→oa/ao*, *ij→ie*, *ui→uu* and *ui→oe*

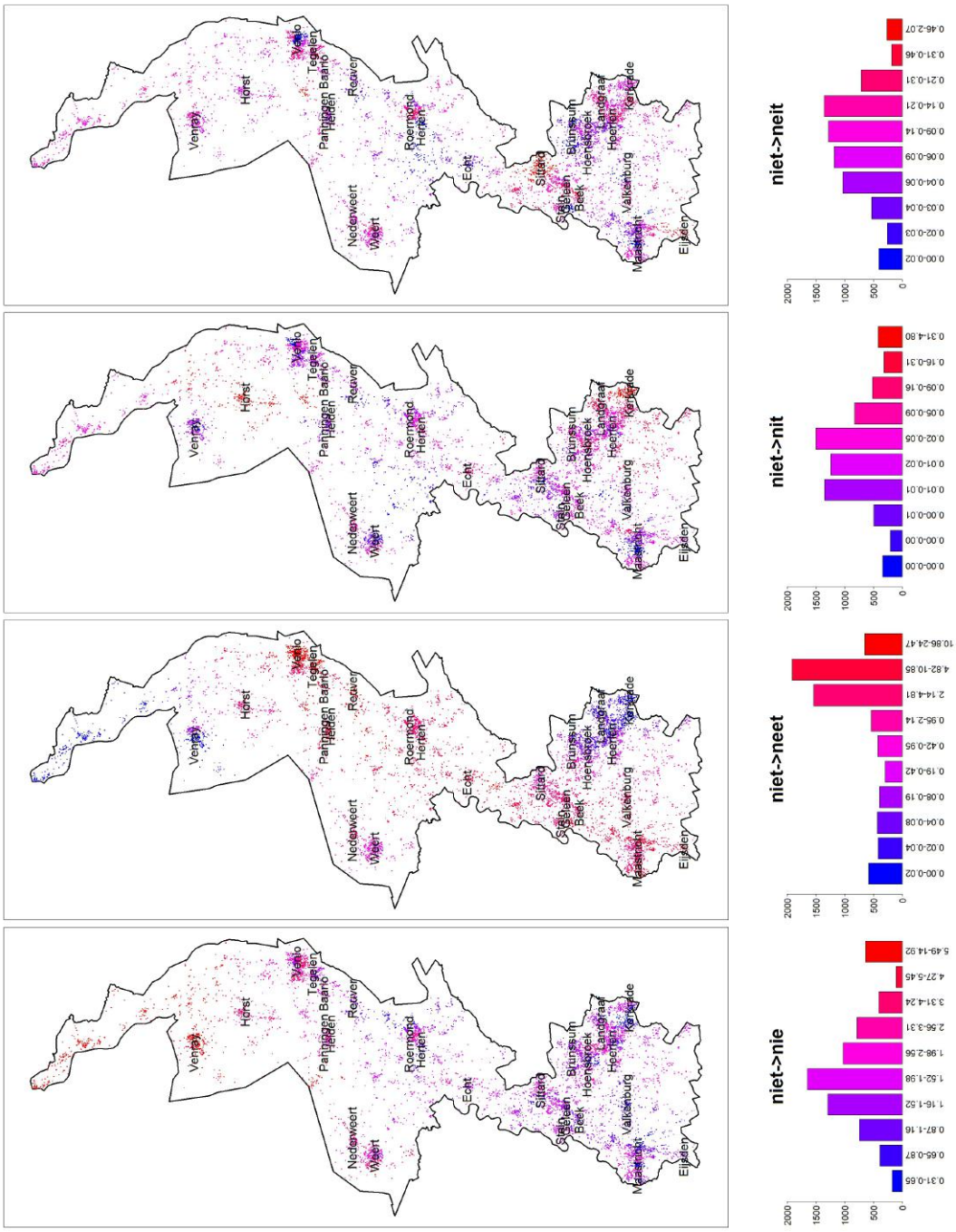

Figure 5: Alternation strength maps for *niet→nie*, *niet→neet*, *niet→nit* and *niet→neit*

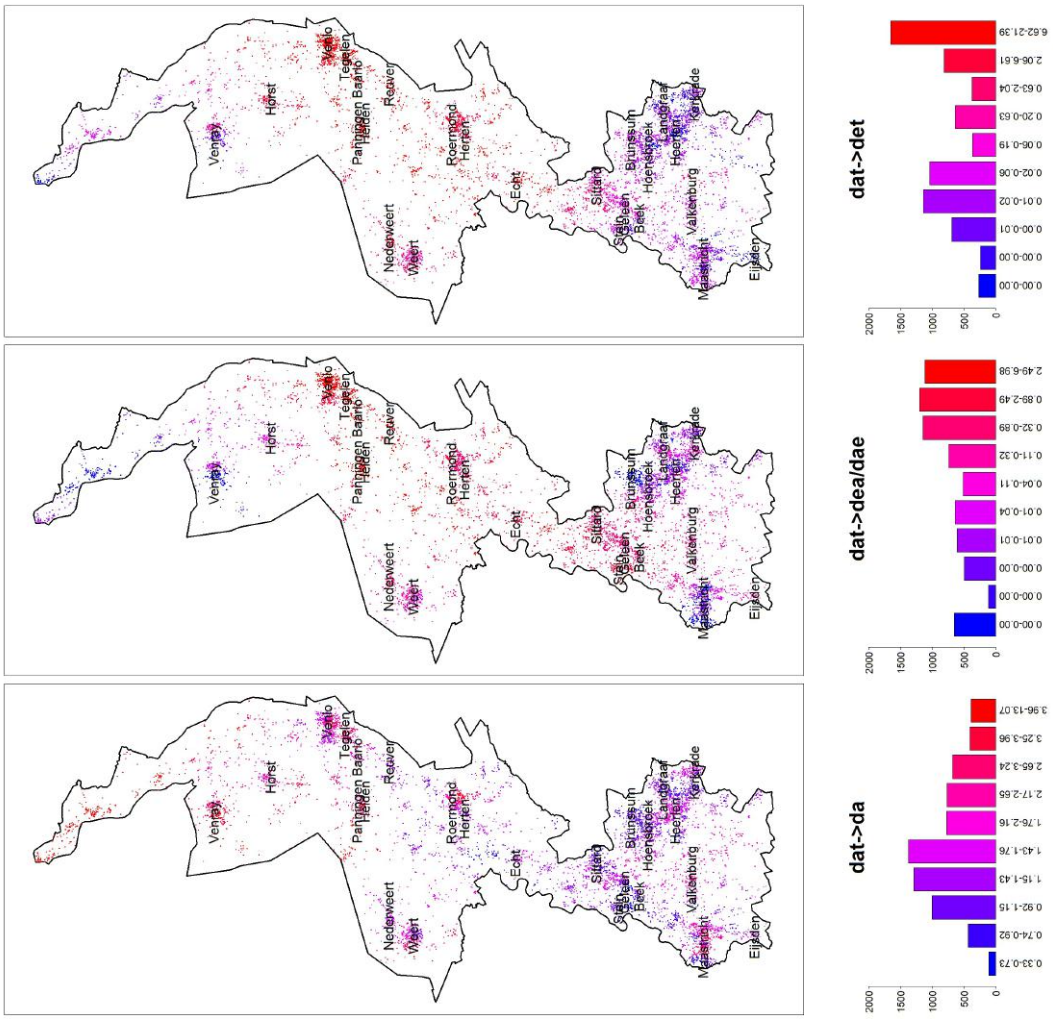

Figure 6: Alternation strength maps for *dat→da*, *dat→dea/dae* and *dat→det*

of Venlo). The most important differences can be found for Weert and Nederweert, as expected (for more details see Bakker (2017)). Remarkable is the expanse of blue in the eastern mine area, which once again seems to point out a low frequency of dialect use.

Our *sj-* map (Figure 3, pane 4) shows that the *sj-* use in Venlo and the area north of Venlo is very low. It gives a sharper distinction than the *ich* map. In Panningen, the colours become more red, but not in the areas around Weert and Nederweert. This corresponds nicely with the course of the Panningen isogloss (Figure 7, pane 2). The colouring of Maastricht, mostly blue to purple, also fits its location outside the *sj-* area. Other than for *ich*, *mich* and *dich* the eastern mine area now does participate in the dialect use, as expected. What we do not see is a small blue or purple coloured strip that should be visible along the river Maas (see Figure 7, pane 2), again perhaps because of on-going mobility and migration.

In general, we may conclude that the Panningen isogloss is more precisely reflected in the *sj-* map than the Uerdingen isogloss in the *ich* map. However, the contours of both isoglosses come

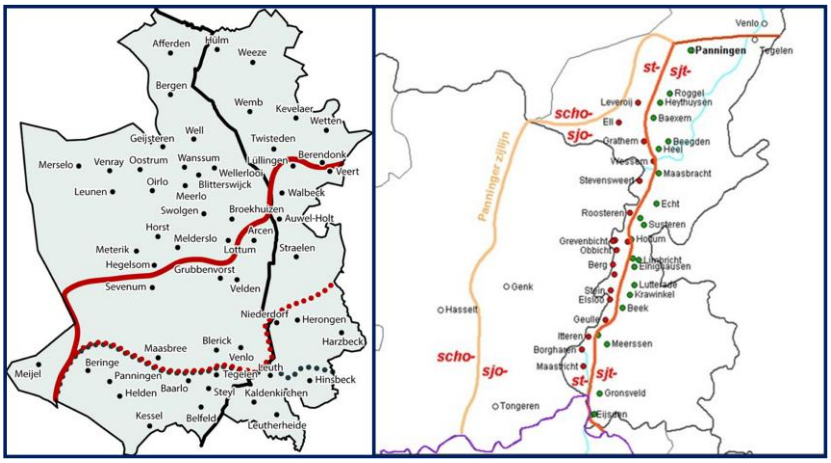

Figure 7: Maps of the Uerdingen and Panningen isoglosses

Alternation	Geography	Twitter map
aa → oa/ao	Whole of Limburg	Figure 4, pane 1
ij → ie	Whole of Limburg	Figure 4, pane 2
ui → uu	North of Limburg	Figure 4, pane 3
ui → oe	Middle and south of Limburg	Figure 4, pane 4
niet → nie	North of Limburg	Figure 5, pane 1
niet → neet	Middle and south of Limburg	Figure 5, pane 2
niet → nit	Eastern mine area	Figure 5, pane 3
niet → neit	Sittard	Figure 5, pane 4
dat → da	North of Limburg	Figure 6, pane 1
dat → dea/dae	Middle and south of Limburg	Figure 6, pane 2
dat → det	Middle and south of Limburg	Figure 6, pane 3

Table 4: Eleven additional Limburg dialect features. The second column lists which regions show the feature in question according to traditional dialectology.

back in our tweet-based geographical patterns, strengthening our thesis that Twitter can be used as a source for dialect studies.

4.2 Verifying the additional alternations

We selected eleven additional known dialectal alternations, listed in Table 4.

The first four alternations relate to vowel pairs that are systematically different between standard Dutch and the Limburg dialects. The *oa* and *ao*, variant spellings for the Dutch *aa*, apply to the whole Limburg dialect area, as the pronunciation of this vowel is more to the back and darker in the Limburg dialect area. We expected no distinctions in spontaneous dialect spellings between *oa* and *ao* and therefore compiled a joint frequency count. The map (Figure 4, pane 1) shows more dialectal variants starting above Venlo, going down south along the cities of Roermond and Sittard/Geleen towards Maastricht. The mostly blue parts are Venray and the very north of Limburg, Weert and the eastern mine area. We may see this alternation as a general marker for the degree of dialect use throughout Limburg.

The *ie* is the dialect counterpart of the Dutch diphthong *ij* in the whole of Limburg, although there is regional variation in the set of words involved. Again, this alternation should probably be seen as a global index of the degree of dialect use, given the amazing similarity of the *ie* map (Figure 4, pane 2) to the *oa/ao* map. The same distribution is found in the map of the *ui* → *oe* alternation (Figure 4, pane 4). Additional information is provided by the *uu* variant (Figure 4, pane 3). It is typical of the north as it ought to be, but there are additional red spots in Venlo, Weert, and Maastricht for which we do not have an explanation as yet.

The sensitivity of our approach to specific regional distributions is illustrated by the next four maps (Figure 5). The four alternations involved all pertain to the Dutch standard negative adverb *niet*. The four dialectal variants are typical of different dialect areas. The variant *nie* is a feature of the very north and *neet* is a feature that covers the remaining area, except for two subareas: the *nit* variant relates to the eastern mine area and *neit* is the variant of Sittard and surroundings. The maps confirm this result, though we have unexplained red spots on the *nit* maps in the north. The special position of Sittard nicely arises on the *neit* map, although we again have various unexplained red spots.

Finally, we have three variants of *dat*, which can be either a demonstrative or relative pronoun, or a subordinating conjunction. It would have been better if we could have distinguished its different functions, as e.g. Maastricht has the variant *dat* for the pronoun and *tot* for the conjunction. However, as we are lacking a POS tagger for (especially dialectal) tweets, we could only track the word forms. These forms do give different geographical patterns. The map of *da* (Figure 6, pane 1) shows that this form is more typical for the north. The *dae/dea* and *det* maps (Figure 6, panes 2 and 3) show that these variants are more typical for the rest of Limburg. As expected Maastricht remains mostly blue on these maps, although the center of Maastricht registers on the *da* map (for which we should investigate the full tweets as this might well be some other use of the form *da*).

Just like the maps for the four dialect features in Section 4.1, the maps for our eleven additional features give the general impression that they mostly mirror the expected geographical patterning. We hypothesize that the major differences can be explained by lower degrees of dialect use in specific areas, as supported by the fact that the maps for *oa/ao* and *ie* show an almost identical distribution.

4.3 Principal Component Analysis

From the fifteen individual measurements in the two previous subsections we derived three principal components, together explaining 75.4% of the variance, to which we applied a varimax rotation (Kaiser 1958). Maps for the three components after rotation can be found in Figure 8. Loadings of the fifteen measurements in the three components can be found in Table 5.

The two alternations that we hypothesized to indicate global dialect use have the highest loadings on the first component, Alt_C1. Their distribution is strengthened by six other alternations that do not include the very north of Limburg, a consequence of the fact that in that part the amount of dialect use is overall low as well. Looking at the larger towns we see a conspicuous pattern. Venlo is the most outspoken dialect center. The towns of Maastricht, Roermond and Sittard-Geleen contain purple areas, giving them an in-between amount of dialect use, although possibly restricted to specific localities. Clearly blue spots are Heerlen, Weert and Venray, pointing to a low amount of dialect use. So, Alt_C1 might be characterized as an indicator of overall dialect use.

The second component is related to two phenomena which are more typical for the north: the two alternations with t-deletion, *niet*→*nie* and *dat*→*da*, and the *ui*→*uu* variant. The Alt_C2 can hence be typified as the northern dialect area parameter.

The third component is a remarkable combination of the *ik*→*ich* variant (the Uerdingen isogloss) and the *s-* → *sj-* variant (the Panningen isogloss). The combination of these two variants excludes the Venlo area and assigns a special position to Maastricht that has no palatalization, but has the same colouring as the eastern mine area.

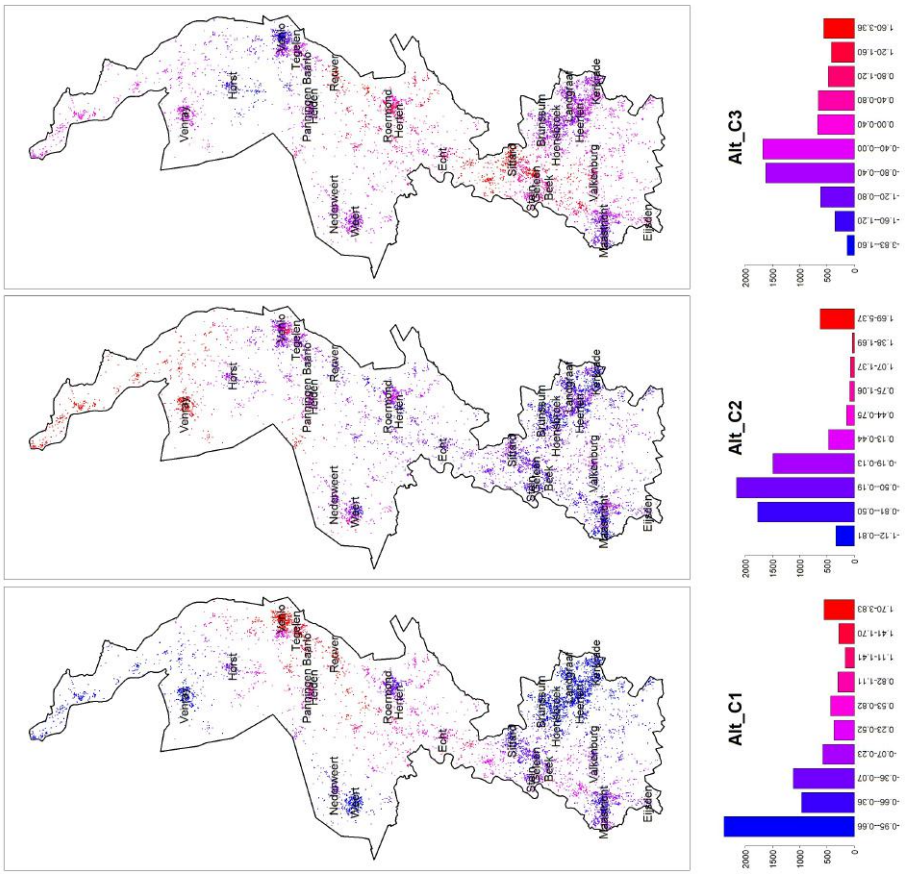

Figure 8: Strength maps for the first three components of a principal components analysis (with varimax) of the alternation measurements

The three components fit the fifteen measurements, giving high loadings for each of them. All components have an interpretable geographical distribution.

5. Knowledge-poor versus Knowledge-rich

Apart from specific word forms alternations, we also measured the frequency of all letter (or space) trigrams found in at least twenty tweets and produced by at least ten different users, 9,216 trigrams in all.

5.1 Trigrams

Character n-grams are used in various text classification tasks, having the substantial advantage of needing no text preprocessing whatsoever. Relevant for the current investigation is that character n-grams are also popular in language and dialect recognition (e.g. Lui and Baldwin 2012; Zampieri et al. 2017). However, whereas character n-grams have been shown to be useful in text classification tasks, their exact relation to the text classes is not always obvious.

	Alt_C1	Alt_C2	Alt_C3
ik→ich	0.36	-0.22	0.75
mij→mich	0.95	-0.10	0.22
jou→dich	0.96	-0.11	0.16
s-→sj-	0.15	-0.27	0.81
aa→oa/ao	0.97	-0.06	-0.06
ij→ie	0.98	-0.01	-0.06
ui→uu	-0.06	0.76	-0.12
ui→oe	0.92	-0.11	-0.04
niet→nie	-0.10	0.94	-0.10
niet→neet	0.96	-0.10	0.04
niet→nit	0.03	-0.00	-0.40
niet→neit	-0.10	-0.01	0.41
dat→da	-0.01	0.87	-0.10
dat→dea/dae	0.92	-0.01	0.11
dat→det	0.94	0.04	-0.12

Table 5: Component loadings for the first three components of a principal components analysis (with varimax) of the alternation measurements

In our investigation, we expected them to at least partially reflect the dialect variation, as many of the effects we know are based on pronunciation, which should be visible in letter trigrams. Also, our smoothing procedure which abstracts from the individual user to the geographical environment of the individual user should ensure that any analysis we apply shows regional variation. Still, before we embark on such an analysis, we want to show a few cautionary examples.

Figure 9 shows maps for five individual trigrams, intuitively chosen by the authors on the basis of world knowledge and experience with Twitter data. Pane 1 shows the (smoothed) distribution of the trigram *vvv*. Rather than pronunciation or Twitter spelling it represents the abbreviation for a) the Tourist Office (likely to be found in the whole of Limburg) and b) the football club VVV-Venlo (likely to be found in and around Venlo). The map shows that indeed the most likely region is highlighted. This means that we have to be aware that there are more "local words" than just dialectal ones.

Pane 2 shows the trigram *ich*, which could naively be expected to follow the dialectal use of *ich*, *mich* and *dich* as described in Section 4.1. A look at the map immediately shows that this is only very partly the case. The trigram does occur more than average in the corresponding areas, but it is mostly concentrated in Maastricht as it is part of the name of that city. Here we can conclude that dialectal use of n-grams can be overshadowed by other, idiosyncratic "local uses".

In the other three panes, we show trigrams that are more linked to sociolects than to dialects. Pane 3 shows *the*, probably the most prominent markers for the use of English. It is not surprising that this is very strong in Maastricht, with its more international stature and university. The other marked locations, however, are less easy to explain. Pane 4 shows *omg* ("oh my god!") as one of the strongest markers of young Twitter users, supported by pane 5 with *gwn*, the standard abbreviation in their sociolect for *gewoon* ("normal", "just"). Here we see that even sociolects may have regional concentrations, for the interpretation of which we would have to investigate the demographics of the various towns and/or neighbourhoods. This figure again shows that our smoothing procedure has the capacity to deliver fine-grained geographical patterns and that this fine-grainedness is important for better interpretation of the areal distribution. Furthermore, it becomes clear that there is not one single young Twitter user sociolect, but that its use may vary per community. After these examples,

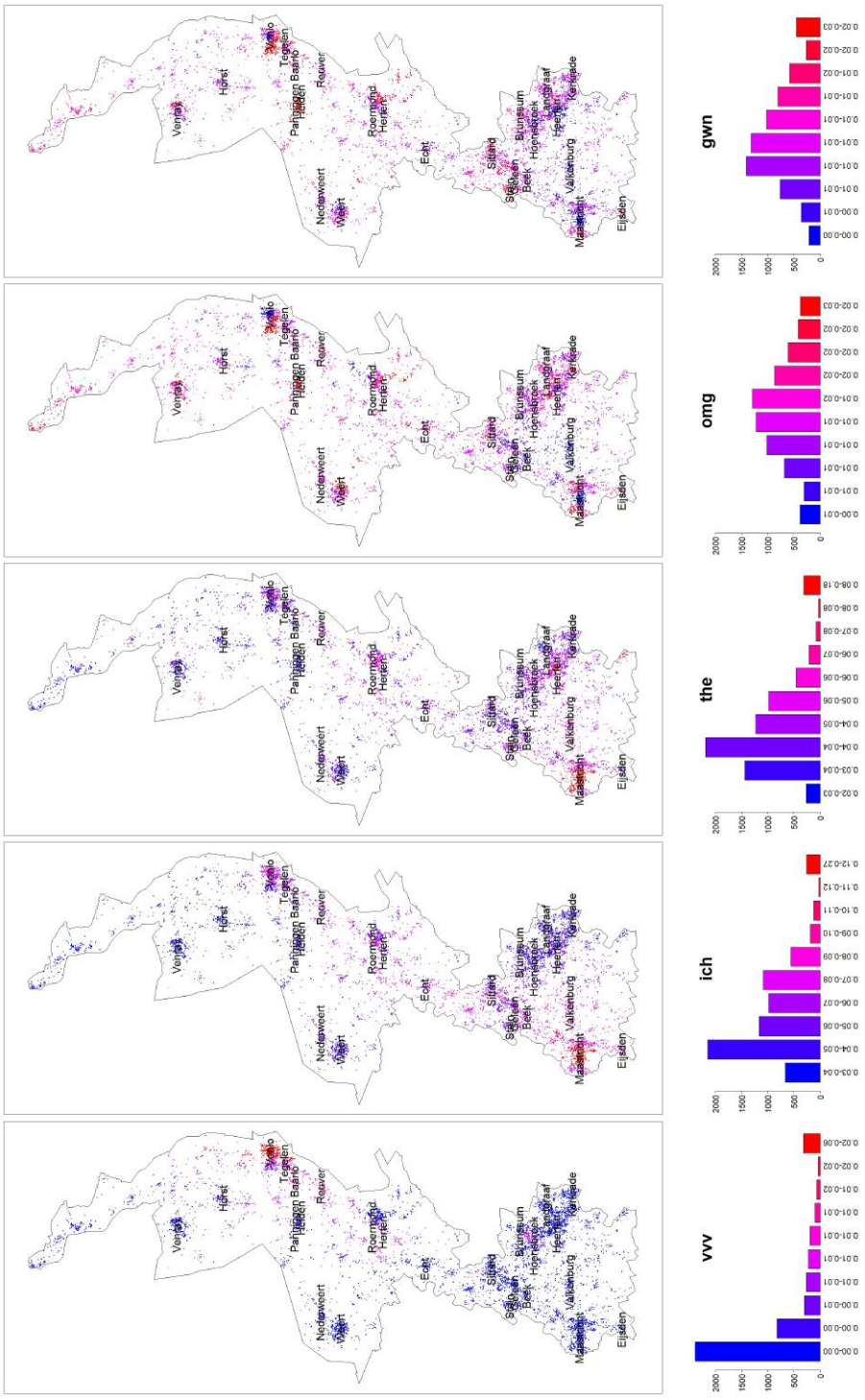

Figure 9: Trigram strength maps for *vvv*, *ich*, *the*, *omg* and *gwn*

it is no longer a surprise that such sociolects also come to the fore in the principal component analysis of the trigram measurements in the next section.

5.2 Principal Component Analysis for Trigrams

Figure 10 shows the maps for the first five principal components (after varimax rotation) for the trigram measurements (together explaining 24.6% of the variance).

The first component (Tri_C1; pane 1) turns out to be not dialectal but rather sociolectal. Trigrams with high loadings on this component include various English trigrams, pointing to many English loan words, which is indicative for younger informants, higher education levels, and perhaps larger companies. The component is concentrated in the largest five towns in Limburg: Maastricht, Venlo, Heerlen, Roermond, Sittard-Geleen. Trigrams from within *Maastricht*, such as *ast* also contribute strongly to this component.

The second component (Tri_C2; pane 2) does appear to be dialectal, highlighting the same strip from Venlo to Maastricht as Alt_C1 (Figure 8, pane 1), although with a different distribution of strengths. In the trigrams with high loadings, we observe a significant number of trigrams with the dialectal spellings *ae* and *ao*. Trigrams from within *Venlo* are more prominent here than in Tri_C1.

With the third component (Tri_C3; pane 3) we return to sociolect, in this case that shown by young Twitter users, as demonstrated by abbreviations such as the already mentioned *omg* and *gwn* and *drm* (short for *daarom*, "because of that"), and by intensifiers such as *fuck* and *kk* (short for *kanker*, "cancer"). The highlighted locations are compatible with those for *omg* and *gwn* as shown above, and are similarly awaiting future research for a more detailed analysis.

The fourth component (Tri_C4; pane 4) shows a region tripartition in a) the north plus Weert and surroundings, b) the region from Venlo to Roermond, and c) the rest of southern Limburg. The loadings give only a few sensible hints, such as trigrams from *sjat* ("darling") for the south and for the north trigrams from *Nijmegen* (name of the town just off the map to the north) and *kei* (an intensifier popular in the province of North Brabant, just west of the blue area). Most of the other strong trigrams cannot be readily interpreted.

In the fifth component (Tri_C5; pane 5), especially many trigrams with *oa* have influence, and some with *ea*. As such there seems to be a dialectal influence. However, the map does not match any known areal distribution.

5.3 Relation between trigrams and alternations

As three of the principal components of the trigram measurements appear to have a dialect background, we have a partial answer to our second research question in that at least some dialect information is being discovered. For the rest of the answer, we need to determine to which degree the discovered information overlaps with what is known about Limburg dialects, or in our case, to which degree it overlaps with the information found with the alternation measurements. We investigated this on the basis of the values for all selected Twitter users in the rotated principal component spaces. We took two approaches. First, we calculated the correlation between the two sets of measurements (Table 6). Next, we used multiple regression (using the function `lm` in `R` (R Development Core Team 2008)) to predict the values for the alternation components from those of the trigram components and vice versa (Tables 7 and 8).

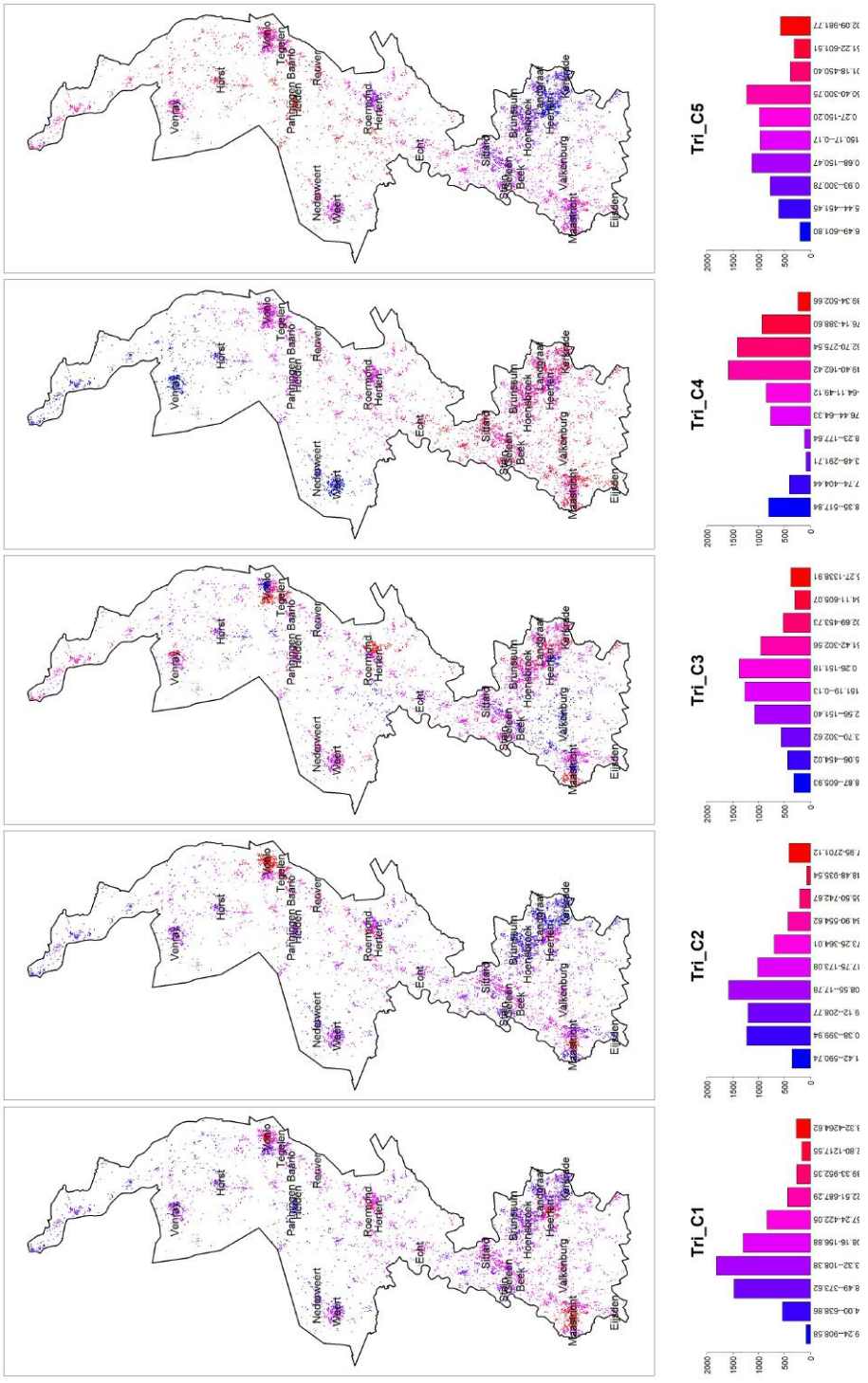

Figure 10: Strength maps for the first five components of a principal components analysis (with varimax) of the trigram measurements

components	Alt_C1	Alt_C2	Alt_C3
Tri_C1	.09	-.25	-.11
Tri_C2	.76	-.02	-.16
Tri_C3	-.16	.04	-.13
Tri_C4	.11	-.70	.31
Tri_C5	.61	.41	-.02

Table 6: Correlations between the five trigram components and the three dialect components.

components	Alt_C1	Alt_C2	Alt_C3
t-value Intercept	0	0	0
t-value Tri_C1	-76***	0	-11***
t-value Tri_C2	224***	-6***	-10***
t-value Tri_C3	54***	-0.5	-16***
t-value Tri_C4	103***	-72***	29***
t-value Tri_C5	132***	25***	3**
R^2	0.937	0.535	0.168

Table 7: Multiple regression, predicting the alternation components with the trigram components.

components	Tri_C1	Tri_C2	Tri_C3	Tri_C4	Tri_C5
t-value Intercept	0	0	0	0	0
t-value Alt_C1	8***	103***	-14***	14***	76***
t-value Alt_C2	-22***	-2*	4***	-92***	51***
t-value Alt_C3	-10***	-21***	-11***	40***	-3**
R^2	0.082	0.607	0.044	0.589	0.538

Table 8: Multiple regression, predicting the trigram components with the alternation components.

The results confirm what we already posited when describing the trigram components. Tri_C1 and Tri_C3 are rather sociolectal and have only tentative connections to the alternation components. Tri_C2, Tri_C4 and Tri_C5, however, are strongly related to dialect use. The most striking individual relation is that between the two strongest dialectal components, Alt_C1 and Tri_C2, as visible in the correlations as well as in both regressions. Still, they are hardly identical as shown by Figure 8, pane 1 and Figure 10, pane 2.

Even more striking is the fact that Alt_C1 can be predicted almost perfectly (an R^2 of 0.937) by the first five trigram components. In other words, the most important component of the dialectal information is indeed being found by way of trigrams. However, the information is distributed in a rather different way, so that there seems to be no direct way to find the dialect information automatically purely on the basis of trigram measurements. The question how feasible this would be with the help of a dialectologist awaits further research.

6. Conclusion

For the Dutch province of Limburg, which is known to have a strong use of dialect, we investigated whether Twitter can be used as a source of dialectal information, thus supplementing and potentially enhancing existing research methods. For this, we identified 7,201 Twitter users in (Dutch) Limburg and measured their use of word forms representing various alternations between Standard Dutch and the dialects in Limburg (knowledge-rich approach), after which we compared the regional patterns of our dialect features with regions and isoglosses known from dialectological sources. In addition, we investigated whether the dialect areas might also be discovered using a knowledge-poor approach, namely measurement of letter (and space) trigrams. Finally, we compared the results from the trigram measurements to those of the alternation ones.

Although we still want to investigate whether other statistical approaches and smoothing techniques may help us to define dialect areas in a more precise way (cf. Rahimi et al. 2017; Eisenstein et al. 2014; Grieve 2013), we found (Section 4) that, all in all, the areas discovered by applying our current techniques to the alternation measurements match well with our expectations based on dialectological knowledge. As such, the use of Twitter as a data source for dialect studies appears viable. We also found (Appendix A) that our newly introduced smoothing technique is better than kriging at preserving variation at a more fine-grained geographical level. However, the use of Twitter and the mapping techniques is not trivial, as we observe various differences between the Twitter-based and the traditionally recognized areas, which need further study. In particular, we observed three confounding effects.

First of all, many alternation maps show a higher variant form use in specific locations outside the expected area, e.g. the use of *da* instead of *dat* in Maastricht. In these cases, it may well be that those variant forms are used in a different sense in those locations. As a result, we have to be careful in selecting word forms for dialect studies, and sometimes may even need to disambiguate between uses of the form.

The second confounding effect is that our Twitter measurements add a quantitative component to the mostly qualitative data from traditional studies. In principle this is positive, but in practice we have not yet learned how to handle the influence of this component properly. As an example, the first principal component of our alternation measurements does not so much select an area in which specific variant forms are being used, but rather the area in which the dialect use is strongest. Further research is needed to separate choice of forms from pure frequency of use.

Finally, we observe that the borderline between variants on our maps is more gradual and scattered than what traditional dialect maps deliver. There are various explanations for this. We already suggested merging populations in the case of Venlo and Tegelen, and population mobility in the case of the Panningen isogloss. Past immigration from outside Limburg is another process that may have triggered lower dialect figures in the eastern mine area. However, in less populated regions, the smoothing procedure used for the maps might also contribute to gradual rather than sharp

divisions. Further research is needed to determine the exact explanatory sources of scattered and fuzzy boundaries, to understand how dialect boundaries become visible in modern communicative exchanges in the social media, and to develop a more refined methodology for the use of Twitter for identifying dialect region.

As to the viability of a knowledge-poor approach, we were surprised at the degree to which dialectal information can be rediscovered in letter trigram counts (Section 5). The abovementioned first principal component of the alternation measurements, which we interpreted as general level of dialect use, can be predicted almost perfectly with a multiple regression analysis on the first five principal components of the trigram measurements. Still, the fact that the same information is somehow present in knowledge-poor measurements does not mean that it can be extracted automatically. Two of the five principal components that we analysed mirror sociolects rather than dialects, even though they are linked to locations instead of to informants. And for the other three, although mostly dialectal in nature, it is not clear how they relate to known dialect regions. We have to conclude that the knowledge-poor approach might be useful in discovering unknown dialectal effects, but as yet only in the hands of an expert dialectologist.

All in all, we can conclude that, at least for areas with strong dialect use, Twitter can be a valuable source for dialect data. Furthermore, the data is not just bigger data than we are used to in this field, but it is also richer data. All we need now is to develop the knowledge and the methodology to use this data properly and precisely, for which we foresee extensive further interdisciplinary research, in close cooperation between dialectologists and data scientists.

References

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio (2013), *Applied spatial data analysis with R, Second edition*, Springer, NY. http://www.asdar-book.org/.

Gräler, Benedikt, Edzer Pebesma, and Gerard Heuvelink (2016), Spatio-temporal interpolation using gstat, *The R Journal* **8**, pp. 204–218. https://journal.r-project.org/archive/2016-1/na-pebesma-heuvelink.pdf.

Hiemstra, P.H., E.J. Pebesma, C.J.W. Twenhöfel, and G.B.M. Heuvelink (2008), Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network, *Computers & Geosciences*. DOI: http://dx.doi.org/10.1016/j.cageo.2008.10.011.

Isaaks, E. H. and R. M. Srivastava (1989), *An Introduction to Applied Geostatistics*, Oxford: Oxford University Press.

Kaiser, H.F. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, pp. 187.

Keulen, R., T. van de Wijngaard, H. Crompvoets, and F. Walraven (2007), *Riek van klank. Inleiding in de Limburgse dialecten*, Sittard: Veldeke Limburg.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. http://www.R-project.org.

Revelle, William (2018), *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois. R package version 1.8.4. https://CRAN.R-project.org/package=psych.

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

van Halteren, Hans (2015), Metadata induction on a dutch twitter corpus: Initial phases, *Computational Linguistics in the Netherlands Journal* **5**, pp. 37–48.

van Halteren, Hans and Roeland van Hout (2017), *Isoglossen in Twitter? Op zoek naar tweet-geografie*, Leiden: Stichting Nederlandse Dialecten, pp. 53–63.

Wackernagel, H. (2010), *Multivariate Geostatistics. 3rd edn.*, Berlin: Springer-Verlag.

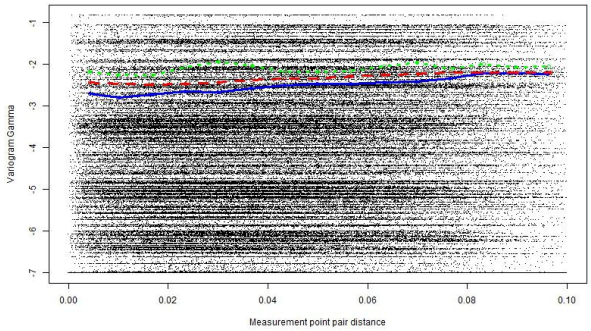

Figure 11: Variogram cloud (with log scale on the y-axis) of measurements for *ik→ich* for the area around Maastricht. The coloured lines represent the variograms for this area (solid blue line), the more sparsely populated area just east of Maastricht (dotted green line) and the whole province of Limburg (dashed red line).

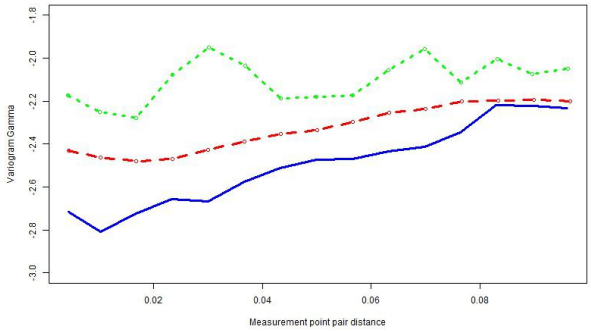

Figure 12: Variograms (with log scale on the y-axis) for the measurements of *ik→ich* for the area around Maastricht (solid blue line), the more sparsely populated area just east of Maastricht (dotted green line) and the whole province of Limburg (dashed red line).

Appendix A.

For our plots, we could not use the raw measurements per user, as there were substantial differences in the measurements per user even within a small region. We therefore needed to smooth the measurements in order to derive measurements for locations. In geographic information systems (GIS), the current standard method is interpolation by *kriging* (Wackernagel 2010), which is based on first fitting a theoretical *variogram* (Isaaks and Srivastava 1989) and then interpolating values for locations from those in the neighbourhood. Grieve (2013) applied this method to dialectal data representing phonetic and lexical variation in the U.S. He used a Gaussian variogram fit and ordinary kriging. However, our data was rather different from Grieve's in three essential respects. First of all, his measurements were rather constant locally, i.e. informants living very close to each other yielded (almost always) very similar measurements. In our case, even in a limited locality, there were vastly

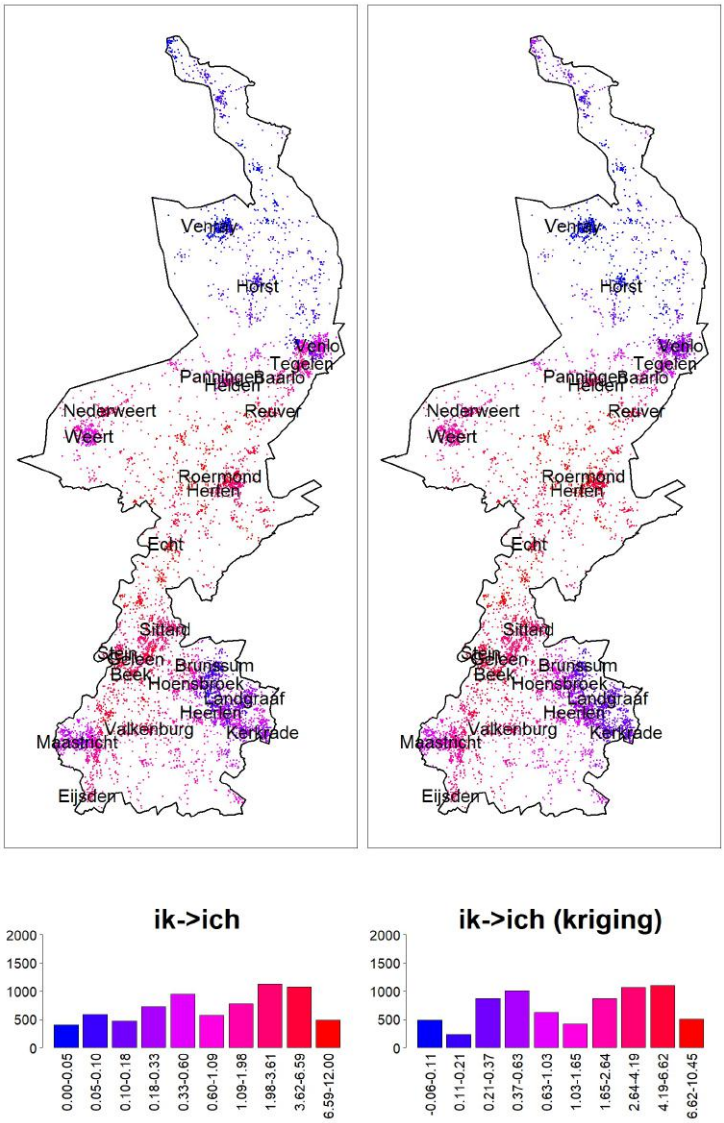

Figure 13: Dialect feature map for *ik→ich*, with two different smoothing techniques. The left pane shows the result with our own technique. The right pane shows the result with *kriging*.

different measurements as the level of dialect use in tweets varies per user, e.g. of two neighbours one might use dialect daily and the other one not at all. The other difference was that Grieve's informants are nicely spread over the U.S. On the East Coast there were some small clusters, but these never even reached ten locations. We on the other hand had very dense clusters in the large cities and very sparse coverage of vast tracts of land elsewhere. We were afraid that using fixed weights for fixed distances would lead to over-smoothing in the cities and insufficient smoothing in sparse areas. Finally, Grieve was forced to interpolate as the phonetic and lexical data sets were built with different informants. We were plotting the actual user locations and did not require interpolation.

Although we expected that kriging would not work as desired, we decided to test whether our negative expectations were correct. We first looked at the variogram. We created variogram clouds and variograms for the alternation *ik→ich* (using `R` (R Development Core Team 2008), the packages `sp` (Bivand et al. 2013) and `gstat` (Gräler et al. 2016), and the function `variogram`), with cutoff set to 0.1 degree, for a) our full dataset, b) a block of 0.1x0.1 degree covering the city of Maastricht (dense coverage), and c) a similar block just to the east of Maastricht (sparse coverage). We show the variogram cloud for the Maastricht block in Figure 11 with the variograms for the three datasets imposed, and just the three variograms in Figure 12. The variogram cloud for this dense region does not show any clear effect of distance, but the corresponding variogram (solid blue line) implies that there is some after all. This is also visible for the full province of Limburg (dashed red line), and overall for the sparse region (dotted green line) although less visible because of the huge swings. The local variability is visible in that all three variograms start with a downward movement, whereas Grieve's variogram (Grieve 2013, Figures 8-14) invariably start with an upward slope. Also worrying for the kriging is that the variogram is placed higher as the region is less densely populated with measurements.

Still hoping that the negative indications need not be in the way of a proper smoothing, we applied kriging to Limburg as a whole for the alternation *ik→ich* (using `R` (R Development Core Team 2008), the package `automap` (Hiemstra et al. 2008) and the function `autoKrige`), again with a block size of 0.1x0.1 degree. In Figure 13 we show the maps for the output of autoKrige and for our own smoothing. The overall picture of Limburg is very similar, but kriging does seem to oversmooth seriously in the big cities, especially in Venlo and Maastricht. We might be able to prevent this by taking a smaller block size, but the shape of the variogram for the sparse region (the dotted green ine in Figure 12) shows that even with 0.1x0.1 we are already getting too few points for a proper estimation. Even though we expect that our distance-based smoothing may also have led to some undesirable artefacts, we prefer it for now to kriging. However, in the (near) future, we want to investigate alternatives, such as a density-sensitive locally varied kriging.