Tom Ruetten

11 Why do the Dutch swear with diseases?

Abstract: In this paper, we address the general problem of dealing with a complex and interwoven set of influential factors, potentially both linguistic and extra-linguistic factors, behind linguistic choices. We do so by investigating the rare phenomenon of swearing with diseases in Dutch, rather than with the more common Western taboo concepts of (among others) religion and sexuality. The standing hypothesis is that swearing with diseases is related to the Calvinistic cultural background of the Dutch. Methodologically speaking, we perform a corpus-based analysis of a large database of location-specific tweets from Flanders and The Netherlands in which we observe “bad language”. Since the hypothesized influential factor of Calvinism shows an outspoken geographical pattern, we intuitively expect clear overlap of the area in which we observe a preference for disease-based swearing and the area in which Calvinism is the predominant religion. We do not find a straightforward overlap in the geography of Calvinism and disease-based swearing. Although the locations where disease-based swearing is conspicuously frequent are all within the Calvinistic area of The Netherlands, disease-based swearing is more likely in the highly-urbanized region around Amsterdam. Therefore, we can only conclude that urbanity, socio-economic factors, religious affiliation and nationality play intertwined roles in the choice for disease-based swearing versus swearing with words from other lexical domains. Consequently, we end up with a complex and interwoven set of extra-linguistic factors that influence the lexical choice for a taboo word from a specific domain for swearing.

Keywords: Taboo, swearing, Twitter, sociolinguistics, variational linguistics

1 Introduction

In this paper, we address the general problem of dealing with a complex and interwoven set of influential factors, potentially both linguistic and extralinguistic factors, behind linguistic choices. We do so by investigating the rare phenomenon of swearing with diseases in Dutch, rather than with the more common Western taboo concepts of (among others) religion and sexuality. As an example, we wonder why the Dutch consider an annoying alarm clock a ‘teringwekker’ *typhus alarm clock* whereas one could also refer to this as a ‘godverdomde wekker’

<https://doi.org/10.1515/9783110582758-012>

goddamned alarm clock. In addition to this clearly sociolinguistically inspired methodology of working with alternation variables, we adopt a Cognitive Linguistic take on swearing, enriched with insights from taboo research, to address the hypothesis that swearing with diseases is a Northern Dutch phenomenon that is motivated by a Calvinistic cultural background (in contrast to the southern region of the Dutch language area, consisting of Flanders and the South of the Netherlands, which has a historical Roman-Catholic cultural background).

Methodologically speaking, we perform a corpus-based analysis of a large database of location-specific tweets from Flanders and The Netherlands in which we observe “bad language” (Andersson and Trudgill 1990). As such, we can literally chart the geographical variability in the preference for a disease-word over other potential swearwords. In other words, for each location we can measure the ratio of tweets in which the swearwords draw from the domain of disease rather than from other taboo domains such as religion or sexuality. Since the hypothesized influential factor of Calvinism also shows an outspoken geographical pattern, we intuitively expect clear overlap of the area in which we observe a preference for disease-based swearing and the area in which Calvinism is the predominant religion.

However, foreshadowing the results of our empirical investigation, we do not find a straightforward overlap in the geography of Calvinism and disease-based swearing. Although the locations where disease-based swearing is conspicuously frequent are all within the Calvinistic area of The Netherlands, disease-based swearing is more likely in the highly-urbanized region around Amsterdam. Moreover, in the truly outspoken Calvinistic strongholds, disease-based swearing is not booming, despite far reaching efforts to balance the dataset across all locations.

In addition, census data shows a clear levelling of religious affiliation, i.e. no religious preference, across the entire language area, which complicates the contemporary relation between Calvinism and disease-based swearing even more. In total, from the data alone, we can only assume that urbanity, socio-economic factors, religious affiliation and nationality play intertwined roles in the choice for disease-based swearing versus swearing with words from other lexical domains. Consequently, we end up with a complex and interwoven set of extra-linguistic factors that influence the lexical choice for a taboo word from a specific domain for swearing.

In the remainder of this paper, we go into further details of this investigation. Section 2 offers some background on the relationship between taboo and swearing, and also digs deeper into the relationship between Calvinism and swearing with diseases. This section ends with embedding the research of this paper in the framework of Cognitive Sociolinguistics. Section 3 then elaborates on the

methodological apparatus for the empirical part of the corpus-based analysis. Section 4 discusses the results of the corpus analysis, while Section 5 concludes by linking the results to findings in taboo research, Cognitive Linguistics and swearing research.

2 Background

2.1 Taboo and swearing

Taboo concepts are often used for swearing (Cruse 1986: 273). Commonly cited taboo concepts are religion, sexuality, genitals, feces, direct family relationships, etc. (Allan and Burridge 1991, 2006; Nübling and Vogel 2004; Ljung Magnus 2010; Drange, Hasund, and Stenström 2014) But why are these taboo concepts so likely to occur in bad language? The rationale for this is that using taboo language may achieve some speaker-internal cathartic effect (Jay 1992: 241), or it may even have the inter-speaker function of convergence (Dynel 2012: 29), i.e. by swearing a speaker may create a social connection or an atmosphere of solidarity, or it may achieve a speaker-external aggression (Wajnryb 2005: 30). All of this is facilitated by transgressing the social norm that is implicit in taboo, i.e. not to speak about something.

Therefore, it is valuable to look a bit more into what taboo is. An interesting approach is the notion of “tabooness” as developed in Jay (1992: 168): “tabooness is the degree to which a word is inhibited”. Interestingly, Jay (1992: 13) already introduces a sociological component by relating tabooness to the speaker: “Tabooness focuses on the speaker’s inhibitions and his or her decision about what can or cannot be said in a given setting”. Jay then contrasts this with obscenity, which is in the ear of the beholder: “the notion of obscenity focuses on the listeners or recipient of the information”. Consequently, inhibition and obscenity are two concepts that must be negotiated between speaker and hearer (Bartsch 1987; Harder 2012) to reach a maximum of effect in terms of catharsis, inter-speaker convergence and/or speaker-external aggression.

Building on this, we would like to put forward that the negotiation between speaker and hearer is specifically for taboo in the context of swearing of a most interesting nature. Could the effect of a swearword lie in finding a delicate, implicitly socially debated balance in acceptability? More specifically, one could hypothesize that if a swearword is not taboo enough, the cathartic effect of swearing will not be obtained; to obtain inter-speaker convergence, inhibition and obscenity must be carefully balanced; and to obtain speaker-external aggression,

the speaker must be capable to transgress his/her own inhibition far enough to surpass a threshold of obscenity that triggers aggression in the hearer.

Nonetheless, a logical consequence of this clear socio-interactional perspective is that the gradual nature of “tabooness” must coincide with variability dimensions in both the choice for swearwords and also the extra-linguistic realm. Quite obviously, we know that taboo is culturally influenced, and many examples for this can be found in scientific (Aman 1996; Ljung Magnus 2010) and popular articles. However, due to the social nature of taboo, it is not unthinkable that variation exists below a holistic cultural level, i.e. communities of practice. As mentioned above, we touch upon this more sociological aspect by focusing on a geographical dimension.

2.2 Calvinism and swearing with diseases

Of course, several hypotheses about why Dutch use disease words for swearing have been put forward. The hypothesis we are pursuing (van Sterkenburg 2000, 2008) revolves around the notion that there is a taboo in Dutch society on being in poor health, which can be linked to Calvinism. Being unhealthy is supposedly considered a sign of damnation, from a religious conviction that God decides over one’s wellbeing. Indeed, in the North of the language area (cf. Figure 1), the predominant religious confession is Calvinism, and some aspects of the Calvinistic tradition have been said to incentivize a healthier lifestyle than Roman Catholics (Mackenbach 2007), the historically predominant religious confession in the South of the language area. This could be a historical-cultural foundation for the taboo, which in its turn may have inspired swearing. Notice that in the hypothesis of Van Sterkenburg, there is no mention of Flanders.

Figure 1: Calvinism in The Netherlands (1899 on the left, 2002 on the right). Black indicates 100 % Calvinistic population, white indicates 0 % Calvinistic population). Source: cbs.nl

Swearing with diseases is conspicuously a part of contemporary Dutch, but not exclusively. Currently, swearing with diseases outside of Dutch is indeed quite rare, and only some attestations in Yiddish and Polish (Nübling and Vogel 2004: 30–31) are regularly passed on from publication to publication. Moreover, Lötscher (1981) already observed disease-inspired swearing as early as the 14th/15th century in what is now known as the south of Germany, e.g. he found in R. Brandstetter’s “*Blasphemiae accusatae*” the phrase *das dich das nun vallende übel an gienge* ‘that you may get epilepsy’ – basically a polite predecessor of the *krijg X* ‘get X’ template (see this section below for an excursion into the idea of templates). This is clearly well before the Calvinistic era. And also later on, in Shakespeare’s *Romeo and Juliet*, Mercutio famously wishes “a plague on your houses” (plague is considered to be referring to the pox). So, from the onset, we can already point out that it is not Calvinism, and Calvinism alone, that is driving this contemporary Dutch peculiarity of swearing with diseases.

Nonetheless, it is valuable to investigate the relation between Calvinism and swearing with diseases. Indeed, even if Calvinism is not the instigator for the productive disease-based swearing we observe today, it may have been a power to conserve the existing forms of disease-based swearing from earlier times. Although our contemporary Twitter data is not the most suitable dataset to address this historical-diachronic question, it is not unlikely that a trace of that potential conservatism is still present today, in the form of a geographic distribution.

Finally, we have not found literature in which the disease-based forms of swearing in Yiddish or Polish are described as productive, in the sense that new forms of swearing are being produced. This could however happen, perhaps by referring to recent diseases such as ebola or AIDS. Curiously enough, we do find observations in our Twitter corpus of swearing with AIDS and also informally of swearing with ebola. Although this is a weak signal, it indicates that disease-based swearing is potentially productive; whether it is productive due to taboo-related mechanisms or purely for paradigmatic reasons, we cannot conclude.

2.3 Cognitive Sociolinguistics

Given this framework of lexical variability (in choosing a word from one taboo domain rather than a word from another taboo domain), where words function as expletives due to their special taboo status, which is in its turn heavily influenced by a range of cultural, historical and societal factors, it is straightforward to position this paper in the realm of Cognitive Sociolinguistics (Geeraerts, Kristiansen, and Peirsman 2010; Kristiansen and Dirven 2008). Especially when one considers the tangible definition put forward by Martin Putz: “Cognitive Sociolinguistics would [...] take the characteristics of meaning as a non-discrete yet structured category to account for the variation of and within dialects, sociolects, cultures, registers, styles, stereotypes, etc.” (Pütz, Robinson, and Reif 2014: 9). Following this definition, taboo gets “raised” to a part of the meaning of a word, which is, we think, in line with what Cognitive Semantics would allow for.

Apart from looking at taboo from a Cognitive Sociolinguistic perspective, this study also contributes methodologically to the field by presenting a unique and freely available database of geo-specific swearing utterances. Moreover, it also sheds light on the possibility of investigating rare phenomena with corpus material, because it is relatively straightforward to collect unprecedented amounts of words from social media. This is also showcased by the work of Jack Grieve et al. (Forthcoming), who has also compiled a very large corpus of geo-annotated tweets and investigates the rise of neologisms.

In this section, we would also like to point out that we will not go into the details of the purely linguistic side of swearing. In the current study, we have quite broadly defined the variation space to be words in utterances that can be bad language, with the alternation taking place between lexical fields, i.e. we observe that a bad word from one lexical field is used, rather than from some other lexical fields. However, it is not unthinkable that the use of a specific swearing word is not so interchangeable as we present it due to purely linguistic reasons. As an example, alliteration between the swearword and the object that

is being cursed may be a reason to select one swearword, rather than another. Ignoring these potential linguistic constraints, we can try to see swearwords as fillers for slots in templates. And indeed, a number of interesting linguistic structures could be observed. One of them is the *krijg de X* ‘get the X’ template, with X being a disease: Sanders and Tempelaars (1998) position the first attestations of that popular template in the 16th century (Nübling and Vogel 2004: 22), which is indeed the era in which Calvinism is emerging widely (since more or less 1530). Moreover, the template is still productive, with attestations in our Twitter corpus (cf. Section 3 below) of *krijg aids* ‘get AIDS’, and an informally observed rise in the use of *krijg ebola* ‘get ebola’ during the ebola crisis of 2015. Clearly, such linguistic structures or (extra)linguistic constraints need to be investigated in more detail.

3 Data and methodology

In this section, we go into some details regarding the collected data and the methodology that we follow to perform a qualitative and quantitative analysis.

3.1 Twitter corpus

Investigations into swearing have either based themselves on anecdotal evidence (Montagu 1967; Aman 1996), textbook or undocumented examples (Nübling and Vogel 2004), questionnaire data (van Sterkenburg 2000) or psycholinguistic research (Jay 1992; Jay 1999). In recent years, one can also see a more corpus-oriented approach (McEnery 2004; Beers Fägersten 2012; Han and Wang 2014; Drange, Hasund, and Stenström 2014). The current paper subscribes to that corpus linguistic approach by investigating swearing by means of a geo-specific corpus of Twitter messages. Twitter is an application with which individuals can broadcast a short and usually public 140-character message. This message is brought to the attention of other Twitter users that follow the sender. They have the ability to re-broadcast that message to their followers, called “retweeting”, which may have a recursive and quadratic effect on audience size. The messages are, in most cases, accessible to everybody, and not just the followers. From 2014 onwards, Twitter has become more of a branding tool, also used by companies, marketing agencies and professionals. This change in function of the medium has potentially had an impact on both the audience and the kind of language that can be observed. Nonetheless, even in today’s format, personal messages and conversations remain on Twitter, albeit it to a lesser extent due to the rise of

media such as Snapchat or Whatsapp that restrict public access to these conversations.

The messages in our corpus were collected in 2012 and 2013. The messages were selected from users that identify in their user profile a location; in many cases this aligns with where they are based (by crossing the GPS location of the tweet with the reported location). We did not look at the actual GPS location from where the tweet is sent, but considered the location in the user profile to be a better indication of the region with which the Twitter user identifies. To obtain a more or less balanced corpus, with respect to the (reported) locations, we tried to find about 200 Twitter users per location (by recursively collecting the followers of a number of seed users). Then for each Twitter user per location, we downloaded the 200 most recent messages. By doing so, we collected on average about 2.000 words per user and 350.000 words per location. In total, we look at about 75 million words from 6 million tweets by 46.000 Twitter users in 213 locations.

This corpus is available for download¹.

3.2 Search words

From this corpus, of course, we needed to derive a dataset of tweets that contains swearwords. We start from a list of taboo source domains with taboo lexical items that we could easily get from Sanders and Tempelaars (1998), a dictionary of swearwords, ordered by source domain. We extracted from that dictionary lexical items and their spelling variants for which we could search in the corpus. At some places, we added a few English and French swearwords that are known to be borrowed, e.g. *shit* and *putain* ‘prostitute’. These lexical items are summarized in Table 1.

¹ http://perswww.kuleuven.be/tom_ruette/data/twitter, last access September 3, 2016.

Table 1: Source domains and lexical items for swearing

Source domain	Lexical items
GENITALS	<i>Kut</i> vagina, <i>vagina</i> vagina, <i>spleet</i> vagina, <i>reet</i> vagina, <i>mutts</i> vagina, <i>eikel</i> penis, <i>lul</i> penis, <i>piemel</i> penis, <i>penis</i> penis, <i>pik</i> penis, <i>dick</i> penis, <i>kloot</i> balls, <i>klote</i> balls, <i>aars</i> ass, <i>kont</i> ass
FECES	<i>Kak</i> 'shit', <i>shit</i> 'shit', <i>crap</i> 'shit', <i>schijt</i> 'shit', <i>stront</i> 'shit', <i>keutel</i> 'shit', <i>pipi</i> 'piss', <i>merde</i> 'shit', <i>scheisse</i> 'shit', <i>diarree</i> 'diarrhea', <i>poep</i> 'shit', <i>pis</i> 'piss', <i>zeik</i> 'piss'
SEXUALITY	<i>Fuck</i> 'fuck', <i>hoer</i> 'prostitute', <i>slet</i> 'slut', <i>putain</i> 'prostitute', <i>neuk</i> 'copulate', <i>pijp</i> 'blow', <i>lek</i> 'lick', <i>kus</i> 'kiss'
RELIGION	<i>God</i> 'god', <i>dju</i> 'god', <i>dieu</i> 'god', <i>gatver</i> 'god damn', <i>jesus</i> 'jesus', <i>christus</i> 'christ', <i>duivel</i> 'devil', <i>hemel</i> 'heaven', <i>hel</i> 'hell', <i>heilig</i> 'holy'
DISEASES	<i>weitig</i> 'disease', <i>syfilis</i> 'syphilis', <i>stuipen</i> 'stroke', <i>beroerte</i> 'stroke', <i>mazelen</i> 'measels', <i>Lazarus</i> 'lazarus', <i>kramp</i> 'cramp', <i>krets</i> 'infection', <i>koorts</i> 'fever', <i>tering</i> 'tuberculosis', <i>tyfus</i> 'typhus', <i>bult</i> 'tumor', <i>influenza</i> 'influenza', <i>griep</i> 'influenza', <i>kanker</i> 'cancer', <i>pleuris</i> 'pleurisy', <i>jicht</i> 'gout', <i>pest</i> 'plague', <i>pokke</i> 'pox', <i>takke</i> 'attack', <i>cholera</i> 'cholera', <i>kolere</i> 'cholera', <i>lepra</i> 'lepra', <i>aids</i> 'AIDS', <i>ebola</i> 'ebola', <i>hoest</i> 'cough', <i>ziekte</i> 'disease', <i>kiezen</i> 'molar', <i>longen</i> 'lungs', <i>koliek</i> 'colic', <i>polio</i> 'polio'

A simple Python script filtered the 6 million tweets in the corpus down to a set of about 70.000 tweets that contain at least one of the lexical items from Table 1. However, not all of these tweets use the lexical items that we identified as a swearword, e.g. when the lexical item is actually used to refer to the disease, rather than using the word figuratively. Therefore, we had to manually go through all these tweets, yielding 20.000 tweets that we consider to be instances of swearing. The details of this annotation step are described in the following section.

3.3 Annotation

We have annotated each tweet, first, for whether or not it contains a swearing incident, and second, with the conceptual domain from which it draws for the swearing. For each tweet, we know from Twitter what the location is to which the Twitter user affiliates from his Twitter profile. The dataset has a format as in Table 2, with a typical KWIC layout, a column to indicate whether or not the observation is to be considered as swearing, an annotation of the conceptual domain from which the swearing draws, and the location. The raw dataset, as well as the R

code that was used for the quantitative analysis can be downloaded² for further analysis.

The annotation of whether or not a tweet contains a lexical item that is used as a swearword is not trivial, mostly because it is impossible to classically define a linguistic category such as “swearing” with necessary and sufficient conditions. We followed a guideline that can be summarized as follows: under swearwords we understand expressive figurative language that draws from a domain that is typically considered to be taboo; the use of swearwords must have the function of emphasizing positive or negative emotions; furthermore, even if a lexical item from a potentially taboo domain is used in the above described fashion, it is not considered to be a swearword if that lexical item is not considered to be offensive, clearly crosses an inhibition or may be perceived as obscene.

Table 2: Excerpt from the raw dataset³

left	kw	acc?	right	domain	norm_loc
HaHa	!@ FUCKI- PEDIA		gets a compliment, thinks he's GOD, take the wrong path, doesn't know exactly who RT him & blocked me!Who's the faggot?	sexual	Heerlen, Nether- lands, Heerlen, Netherlands
Wet renault afschaffen en da uit nen bek van socialist,het word tijd dat de politiekers gansters hunne kop is uit hunne	aars	t	trekken	genitals	Mortsel, Belgium, Mortsel, Belgium
Delen	!Fuck		! 🎵 Epic Soundtracks http://t.co/a1ZjbTga #Spotify	sexual	Bilzen, Belgium, Bilzen, Belgium

² http://perswww.kuleuven.be/tom_ruette/data/twitter_invectives/, last access September 3, 2016.

³ Column “left” shows the context words on the left of the keyword, column “kw” shows the keyword that was found in the corpus, column “acc?” indicates whether the observation was considered to be an actual observation of swearing, column “right” shows the context words on the right of the keyword, column “domain” shows the taboo domain from which the swearword draws, and column “norm_loc” shows the cleaned location names of the user profiles.

Table 2 (continued)

left	kw	acc?	right	domain	norm_loc
Nog voor de #kerst 2 nieuwe modellen #GedenkkastjeART op gedenkkastje.nl! Icoon Maria	& Jezus		en Engelen van Brigita Silovic	religion	Nijmegen, Nether- lands, Nijmegen, Netherlands
@MiekelnC naar de ondemand stream, niet naar de ,ik wil het luisteren als ik tijd heb' stream ;) of kijk ik met mijn #	aars	t	?	genitals	Amsterdam, Nether- lands, Amsterdam, Netherlands
(Schijn	)heilig		boontje op fiets gemhuis in in #Losser. Alter-ego bgm #Sijbom. En nu weer naar buiten: optocht start! http://t. co/5CNsPt9m	religion	Netherlands, Neth- erlands
*Tousse	*Bullshit		*Tousse* http://t.co/ X8wYf2KbCe	faeces	Namur, Belgium, Namur, Belgium
Spitaels	,"Dieu		„is overleden.Allez,ik zal maar beginnen bleiten zeker?	religion	Kortrijk, Belgium, Kortrijk, Belgium
"Landgenoten	,Fuck	t	this shit, I'm outta here! Latertjes, woord aan je moeder." #aftreden	sexual	Gouda, Netherlands, Gouda, Netherlands

3.4 Quantitative analysis

Given our research focus on the geographical pattern of swearing with diseases, and the potential relation with the geographical spread of Calvinism, we have to fall back to a geospatial analytical instrument. We identified Spatial Autocorrelation as the most appropriate statistical method to detect geographical patterns in data.

Spatial Autocorrelation comes in two flavors. Global Spatial Autocorrelation indicates whether the data contains a geographical pattern. Local Spatial Autocorrelation indicates for each observed location whether it belongs to an area in

which the phenomenon under investigation reaches high values or low values. This method has already found an application in linguistics (and is described in detail by Grieve 2014; Grieve 2012; Grieve 2013; Grieve, Speelman, and Geeraerts 2013; Grieve, Speelman, and Geeraerts 2011). For details on the statistical method, we refer to the aforementioned publications; here, we restrict ourselves to introducing the intuition behind the method.

3.4.1 Global spatial autocorrelation

The first flavour of spatial autocorrelation is Global Spatial Autocorrelation and indicates whether the phenomenon at hand exhibits a geographical pattern. The strength of the geographical pattern is expressed with Moran's I (Moran 1948; Odland 1988), which can be compared to an effect size in a more common inferential statistical analysis. Moran's I lies between -1 and 1 , with 0 indicating a random geographical pattern. A value of Moran's I close to -1 indicates that nearby locations have different values; a value of Moran's I close to 1 indicates that nearby locations have similar values. In the case of our swearing patterns, we are expecting that nearby locations will have similar values, so that we can discern areas, i.e. groups of locations, with a specific swearing pattern. As with any effect size, it is insightful to also return a p value that indicates the statistical significance of that effect size. A Moran's I value close to 1 might be indicative of a strong geographical pattern, but if the p value suggests that this pattern is just a chance finding, one needs to attenuate the conclusions. Vice versa, a statistically significant pattern ($p < 0.05$) may very well be random, i.e. with Moran's $I = 0$.

3.4.2 Local spatial autocorrelation

Whereas Global Spatial Autocorrelation indicates whether the phenomenon under investigation exhibits a geographical pattern in the full territory, Local Spatial Autocorrelation zooms in on the individual locations. For each location, Local Spatial Autocorrelation calculates a value that indicates “the degree to which that particular location is part of a high- or low-value cluster” (Grieve 2011: 13). So, rather than returning the one value of Moran's I for the full territory as in Global Spatial Autocorrelation, Local Spatial Autocorrelation returns one value for each location. The returned value is referred to as G_i^* (Ord and Getis 1995) and it is at heart a z -score “indicating the degree to which a location is surrounded by locations with similar values” (Grieve 2011: 13).

We visualize these z -scores per location with color-coded dots on a map. Each location in the dataset is represented by a dot, and the color ranges from blue over white to red. A white dot indicates that the location is in an area with transitional values, somewhere between the highest and lowest values that are observed. A blue dot indicates that the location is in an area with very low values, and a red dot indicates that the location is in an area with very high values. Note that these dots are thus indicative of the area in which the location resides, and not about the location itself. As an example, a location with a very low observed value that is amidst a group of locations with very high observed values will nonetheless receive a reddish color. Because of this, it is also possible to think of local spatial autocorrelation as a geographical smoothing function that smothers out local aberrations.

3.4.3 Spatial weighting function

Both Global and Local Spatial Autocorrelation depend heavily on a so-called Spatial Weighting Function. For Global Spatial Autocorrelation, the Spatial Weighting Function defines what ‘nearby’ means in the calculation of Moran’s I . For Local Spatial Autocorrelation, the function defines the scope of the area to which a location is compared to come up with a z -score.

The most straightforward approach for this Spatial Weighting Function is to work with a threshold: locations within a range of X kilometers are considered nearby. Of course, this rather crude function does not take into account that some locations may be isolated, so that no other locations fall within this threshold and no reference values can be calculated.

Rather than a threshold based approach, we decided to apply a continuous weighting function. The inverse of the distance is used as the Spatial Weighting Function. In other words, the farther away a reference location is to the location that is currently being tested, the less impact this reference location will have. Although we used a linear function, the idea of continuous weighing is in line with a fundamental finding of dialectology, i.e. that geographical distance is non-linearly and inversely correlated with linguistic similarity (Séguy 1971). This weighting function is also called a reciprocal Spatial Weighting Function, and is explained in more detail in Grieve, Speelman, and Geeraerts (2013).

3.4.4 Values per location

The spatial autocorrelation method assumes that we have been able to quantify the phenomenon at hand, i.e. swearing with words from taboo domains, into a value for each location. From the dataset we gathered, this is not difficult, as it boils down to constructing a commonly used pivot table. In Table 3, an excerpt of the frequency counts of the different types of swearing in a couple of locations is presented.

Table 3: Quantitative representation of swearing with lexical items from taboo domains per location

<i>Locations</i>	<i>Diseases</i>	<i>Religion</i>	<i>Feces</i>	<i>Sexuality</i>	<i>Genitals</i>
<i>Antwerpen</i>	8	72	133	68	88
<i>Gent</i>	15	172	234	156	193
<i>Amsterdam</i>	12	63	213	79	193
<i>Groningen</i>	18	34	68	34	110

From the absolute values in Table 1, we have to calculate relative frequencies, to be able to say something about how people swear across the territory. However, relative frequencies can be calculated in several ways. First, one could calculate the frequency of tweets containing swearwords relative to the total number of tweets per location. As an example, consider Figure 2, in which the number of tweets containing swearing in a certain location is made relative to the total number of tweets in that location.

Note that, as explained above, the maps that are shown in this section have been treated with a Local Spatial Autocorrelation analysis to bring forward geographical patterns. That is why we see the values per location as red-white-blue dots. In addition, we also provide the summary statistic from the Global Spatial Autocorrelation. In Figure 2, we see that Flanders is an area in which we observe more swearing per location than in the Netherlands. This pattern is (globally speaking) significant, with the indication from Moran’s *I* that this is only a modestly strong pattern.

4 Results and discussion

So, we will investigate the geographical pattern of disease-based swearing, as it occurs in contrast to the other domains for swearing that we incorporated, i.e. RELIGION, GENITALS, FECES, and SEXUAL. The pattern that we observe is presented in Figure 3. On the left-hand side, we show the raw proportions that we observed, with darker dots indicating a higher proportion of disease-based swearing tweets. On the right-hand side, we show the map after treating it with spatial autocorrelation. The pattern we observe has a modest effect size, with Moran's I at 0.1374, and a very small p value.

Figure 2: Disease-based swearing in contrast to swearing based on other domains.

We observe that disease-based swearing occurs with elevated probability in the Rim city area (South of the lake). In fact, almost all red dots are located north of the river Maas and Waal, which is historically speaking the border between the Calvinistic North and the Catholic South. Some red dots do occur below these rivers, especially in the neighborhood of 's-Hertogenbosch. 's-Hertogenbosch is the capital of North Brabant, and an important economic center, home to the brewery Heineken (among other international industries), so it may attract people from a wide area. The red dots are mostly absent in the North of the Netherlands, and along the border with Germany.

In Flanders, we observe mostly blue dots, indicating that swearing with diseases is very uncommon. Here and there we do see some white dots, which indicate a region of transition. Rather than attributing too much importance to and trying to interpret these white dots, we can safely say that the coloring of the dots may be here an artifact of the quantitative method and may be potentially due to some data scarceness. As a matter of fact, due to the use of a reciprocal Spatial Weighting Function, there will always be a range of white dots between the blue and the red dots, so one should not over-interpret this.

We observe that there is clearly an overlap between the area where we observe disease-based swearing and the historical area in which Calvinism was the strongest religion (cf. map in Figure 1). The overlap is not complete, though. In the Calvinistic North-East of the territory, and also in the Calvinistic area of Zeeland (in the South) we do not observe many disease-based swearing tweets. Moreover, in the Catholic province of North Brabant, we do observe disease-based swearing in the vicinity of 's-Hertogenbosch. The highest concentration of disease-based swearing is to be found in the socio-economic center of the Netherlands, around Amsterdam in the so-called Rim City.

5 Discussion and conclusion

In this study, we started from the observation that within the contiguous Dutch language area (North of Belgium and The Netherlands), some Dutch speakers draw swearwords from the DISEASE domain, whereas others don't. One of the standing hypotheses is that this can be related to Calvinism – whereby Calvinism puts diseases under a spell of taboo, and thus making them accessible for swearing. Empirical support for this hypothesis was found, but only to the extent that disease-based swearing seems to occur practically exclusively in the traditionally Calvinistic region of The Netherlands, and not in the traditionally roman Catholic South. However, disease-based swearing could not be observed in the entire Calvinistic region, but predominantly in the Rim City area in and around Amsterdam. This latter observation indicates that also sociological factors, e.g. urban/non-urban, influence the way people swear or, perhaps more basal, what they perceive as taboo.

The question one may ask now is related to untangling the phenomena of swearing, tabooess and cultural background. We know that there are ties between swearing and taboo, and it is also trivial that taboo relates to cultural background. But these relationships are not impervious to confounding influences. As a consequence, we are left with wondering: does the urban/non-urban

cline influence the tabooeness of diseases, and thus indirectly the geographical swearing pattern we observe? Or is the urban/non-urban distinction directly operating on the level of the linguistic practice of swearing, without the go-between of taboo? In other words, do people in urban regions swear with diseases because this is how their peers swear and how people commonly swear in their community of practice? Or do people in urban regions experience the taboo on diseases differently than the people in non-urban regions?

The latter possibility seems to be supported by research that was performed by the Dutch “Bond tegen vloeken” (Union again swearing) in 2015⁴. This research showed that *kanker* ‘cancer’ and *tering* ‘tuberculosis’ are both in the top 5 of the most hurting swearwords (in The Netherlands). And some years earlier, a campaign was launched to keep people from swearing with *kanker* ‘cancer’⁵. These are two examples that indicate that diseases still have a high symbolic – certainly taboo? – value in the culture of the Dutch.

So, why do the Dutch swear with diseases? At the end of this paper, we are still in doubt whether it is because they have a taboo on diseases, or if it is due to their Calvinistic background, or maybe it has to do with a unique urban fad that is currently hip in Amsterdam? The swearing pattern that we observe indicates at least that the choice for a disease as a swearword is a complex interaction of sociological influences on both language use and the tabooeness of concepts.

6 References

- Allan, Keith and Burridge, Kate. 1991. *Euphemism and dysphemism: language used as shield and weapon*. Oxford: Oxford University Press.
- Allan, Keith and Burridge, Kate. 2006. *Forbidden words: taboo and the censoring of language*. New York: Cambridge University Press.
- Aman, Reinhold. 1996. *Opus Maledictorum: A Book of Bad Words*. New York: Marlowe & Company.
- Anderson, Lars-Gunnar and Trudgil, Peter. 1990. *Bad language*. Oxford: Basil Blackwell.
- Bartsch, Renate. 1987. *Norms of Language: Theoretical and Practical Aspects*. London & New York: Addison-Wesley Longman Limited.
- Beers-Fägersten, Kristy. 2012. *Who's Swearing Now?: The Social Aspects of Conversational Swearing*. Newcastle: Cambridge Scholars Publishing.
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

⁴ <http://www.bondtegenvloeken.nl/files/Bond%20tegen%20vloeken/onderzoeken/adviesrapport-bond-tegen-vloeken-2015---samenvatting.pdf>, last access September 3, 2016.

⁵ <http://www.stopscheldenmetkanker.nl/>, last access September 3, 2016.

- Drange, Eli-Marie Danbolt, Ingrid Kristine Hasund, and Anna-Brita Stenström. 2014. 'Your Mum! Teenagers' Swearing by Mother in English, Spanish and Norwegian. *International Journal of Corpus Linguistics* 19 (1). 29–59.
- Dynel, Marta. 2012. Swearing Methodologically: The (im) Politeness of Expletives in Anonymous Commentaries on Youtube. *Journal of English Studies*, no. 10. 25–50.
- Geeraerts, Dirk, Gitte Kristiansen, and Yves Peirsman. 2010. *Advances in Cognitive Sociolinguistics*. Berlin, Boston: De Gruyter Mouton. <https://www.degruyter.com/view/product/177073>.
- Grieve, Jack. 2011. A Regional Analysis of Contraction Rate in Written Standard American English. *International Journal of Corpus Linguistics* 16 (4). 514–46.
- Grieve, Jack. 2012. A Statistical Analysis of Regional Variation in Adverb Position in a Corpus of Written Standard American English. *Corpus Linguistics and Linguistic Theory* 8 (1). 39–72.
- Grieve, Jack. 2013. A Statistical Comparison of Regional Phonetic and Lexical Variation in American English. *Literary and Linguistic Computing* 28 (1). 82–107.
- Grieve, Jack. 2014. A Comparison of Statistical Methods for the Aggregation of Regional Linguistic Variation. In Szmrecsanyi, Benedikt & Bernhard Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, 53–88. Berlin & New York: Walter De Gruyter.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A Statistical Method for the Identification and Aggregation of Regional Linguistic Variation. *Language Variation and Change* 23 (02). 193–221.
- Grieve, Jack. 2013. A Multivariate Spatial Analysis of Vowel Formants in American English. *Journal of Linguistic Geography* 1 (01). 31–51.
- Grieve, Jack, Andrea Nini and Diansheng Guo. Forthcoming. Analyzing lexical emergence in American English online. *English Language and Linguistics*.
- Han, Chong, and Kenny Wang. 2014. Subtitling Swearwords in Reality TV Series from English into Chinese: A Corpus-Based Study of The Family. *Translation & Interpreting* 6 (2). 1–17.
- Harder, Peter. 2012. Variation, Structure and Norms. *Review of Cognitive Linguistics* 10 (2). 294–314.
- Jay, Timothy. 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets*. Amsterdam/Philadelphia: John Benjamins Publishing.
- Jay, Timothy. 1999. *Why We Curse: A Neuro-Psycho-Social Theory of Speech*. Amsterdam/Philadelphia: John Benjamins Publishing.
- Kristiansen, Gitte, and René Dirven. 2008. *Cognitive Sociolinguistics, Language Variation, Cultural Models, Social Systems*. Berlin/Boston: De Gruyter Mouton. <https://www.degruyter.com/view/product/40053>.
- Ljung Magnus. 2010. *Swearing: A Cross-Cultural Linguistic Study*. New York: Palgrave Macmillan.
- Lötscher, Andreas. 1981. Zur Sprachgeschichte Des Fluchens Und Beschimpfens Im Schweizerdeutschen. *Zeitschrift Für Dialektologie Und Linguistik* 48. 145–60.
- Mackenbach, Johan. 2007. Jean Calvin, Calvinism, and Population Health: Impressions from Switzerland. *The European Journal of Public Health* 17 (1). 1.
- McEnergy, A. 2004. Swearing in Modern British English: The Case of Fuck in the BNC. *Language and Literature* 13 (3). 235–68.
- Montagu, Ashley. 1967. *The Anatomy of Swearing*. Philadelphia: University of Pennsylvania press.

- Moran, P. A. P. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2). 243–51.
- Nübling, Damaris, and Marianne Vogel. 2004. Fluchen Und Schimpfen Kontrastiv. *Germanistische Mitteilungen* 59. 19–33.
- Odland, John. 1988. *Spatial Autocorrelation*. London: Sage Pub.
- Ord, J. Keith, and Arthur Getis. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27 (4). 286–306.
- Pütz, Martin, Justyna A. Robinson, and Monika Reif, eds. 2014. *Cognitive Sociolinguistics: Social and Cultural Variation in Cognition and Language Use*. Vol. 59. Benjamins Current Topics. Amsterdam: John Benjamins Publishing Company. <http://www.jbe-platform.com/content/books/9789027270276>.
- Sanders, Ewoud, and Rob Tempelaars. 1998. *Krijg de Vinkentering!* Amsterdam/Antwerp: Contact.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35. 335–357.
- van Sterkenburg, Piet. 2000. *Vloeken: Een Cultuurbepaalde Reactie Op Woede, Irritatie En Frustratie*. Den Haag: Sdu Uitgeverij.
- van Sterkenburg, Piet. 2008. Vloeken Is Niet Meer Wat Het Geweest Is. *Taalschrift.org*.
- Wajnryb, Ruth. 2005. *Expletive Deleted: A Good Look at Bad Language*. New York: Simon and Schuster.