"Tell me who you talk to, and I'll tell you how you talk": Comparing the language use of two interaction based clusters of people in a single Usenet newsgroup

Tom Ruette, Dirk Speelman, Dirk Geeraerts

Quantitative Lexicology and Variational Linguistics, University of Leuven

1 Introduction

This paper discusses a quantitative analysis of the variation of a number of linguistic features in a corpus of Computer Mediated Communication. We start from the the hypothesis that social interaction is a predictor of linguistic variation, independent of demographic background. The hypothesis is tested by means of a study on a corpus of on-line discussions in Usenet. We selected the highly active "be.politics" newsgroup during the year 2009. This newsgroup allows Dutch speaking Belgians to debate current political affairs. Within this newsgroup, we identified two clusters of users, by means of their interaction patterns. To ensure that these clusters are not confounded or motivated by the demographic background of its members, we performed a computer-assisted semi-automatic qualitative analysis of personal information that the members self-reveal. It appeared that that the interaction-based clusters were not confounded nor motivated by the demographic background of its members. Then, the usage of a number of linguistic features is counted in the two clusters and compared by means of Analyses of Variance. We'll show that some linguistic features show significant variation between these clusters. Consequently, we claim that social interaction is a possible predictor for linguistic variation in the case of a Usenet newsgroup on Belgian politics. Further research will show whether this finding can be extrapolated to other registers and topics.

The remainder of this paper consists of 5 parts. In section 2, we introduce the theoretical background that supports our research question. Section 3 gives a small overview of the history and characteristics of Usenet. That same section also tackles the issue of finding clusters of people based on interaction patterns. The following part, section 4, describes the qualitative analysis of the demographic background of (some of) the cluster members. Section 5 gives an overview of the Analyses of Variance and the outcomes, and finally a conclusion is formulated in Section 6.

2 Theoretical background

The development of variationist linguistic research has shown us that the synchronous state of language still allows for considerable variation in the grammar. This variation appears to be linked to a limited amount of explaining dimensions. The first dimension concerns "inter-speaker variation" and has been studied in quantitative socio-linguistic research. Traditionally, this research tries to explain the varying use of linguistic phenomena between (groups of) language users by referring to social characteristics of these users. A number of these characteristics have been studied very extensively. Socio-linguists have looked into social classes (Labov, 1966), communities (Milroy, 1987), gender (Labov, 1990), age (Eckert, 1997), ethnography (Eckert, 1988). The work of Eckert in Belten High seems especially relevant for the current study, as it looks into the linguistic variation between two groups of people. Her study is however not directly interested in the frequency of interaction between group members, but it rather takes the ethnographic characteristics of the people — fashion, world view, authority — as group identifying features. Her study starts from a common social distinction in high schools: jocks versus burnouts. While jocks strive to participate in the institutional, extracurricular structure of the high school community, and receive privileges for this, burnouts do not submit to the school authority. Burnouts tend to be bound to the local job market, and feel that the school and its activities do not prepare them for that perspective. Jocks are institutionally oriented, while burnouts are more locally and personally oriented. This pair of "Communities of Practice" (Wenger, 2000) is the basis for the social groups of which she studies the language use. Of course, she had access to more sociological information, e.g. fashion and hobbies, as she performed socio-linguistic interviews with the children (Eckert, 2000, Chapter 2). On the linguistic level, she focuses on one syntactic variable (negative concord) and six vocalic variables. She annotated socio-linguistic interviews with the pupils and analysed her measures with logistic regressions. She finds out that the major determiners of the use of socio-linguistic variables are jock or burnout affiliation, gender, and engagement in the practices that constitute those categories (Eckert, 2000, Chapter 5). In short, starting from an ethnographic account of the Belten High social situation, instead of traditional ties or socio-economic class, Eckert uses the concept of a "Community of Practice" (CoP) (Wenger, 2000), rather than a typical Milroyesque Social Network. The members of the social groups that she studies are not necessarily friends or family. Rather, they share opinions on life, authority, fashion and school. For Eckert, the language variation between these CoPs is a way of emphasizing their ethnographic differences. The language variation has in that sense a social meaning.

The second dimension has been named "intra-speaker variation" and concerns the variation in language use of one person in different situations. Although Labov already mentions intra-speaker variation since his earliest studies, the theoretical framework that we review here starts from the observation that the same newsreader alters his speech when presenting news reports on different radio stations with a specific audience (Bell, 1984). Other studies noted similar observations, e.g. that Welsh speakers broaden their Welsh accent when they are confronted with a interviewer with a negative opinion on Welsh (Bourhis and Giles, 1977). Bell (1984) claims in his "audience design" that "speakers take most account of hearers in designing their talk". In short, "speakers design their style for their audience". The findings of Bell are in the line of the Speech Accommodation Theory formulated by Giles and Powesland (1975). Bell mentions also that this style shift goes further than phonological changes, and may

also influence the choice of e.g. pronoun choice. Therefore, the "audience design" framework is also valid for a corpus-based study as the present one. The findings of more recent research in this domain appear to be quite similar to the findings of Eckert (2000), mentioned above. Coupland (2007) (also in Coupland (1985) points out that a specific use of language plays an important role for the construction of "identity" in the society. The function and meaning of language variation is social. Register studies in the line of Biber (1988), however, have shown that language variation can also be merely functional. Biber's MultiDimensional Analysis studies discovered a number of dimension that explained the variation of linguistic features in terms of "informational" versus "involved", or "narrative" versus "non-narrative". An obvious discussion that follows from the distinction between two dimensions concerns the question of which dimension is more important than the other one. Bell (1984) himself already addresses the discussion in the second claim of his ten principles for style analysis: "style derives its meaning from the association of linguistic features with particular social groups". Preston (1991) gives quantitative evidence for this claim by pointing out that the variational range of inter-speaker variation is always larger than the range of intra-speaker variation. However, his findings are contradicted by Finegan and Biber (1994). An interesting discussion between Preston, Finegan and Biber can be found in Eckert and Rickford (2001).

Findings of research along the first dimension has set off the social turn in linguistic research (Kristiansen et al., 2008; Harder, Forthcoming) towards Cognitive Sociolinguistics. One of the central issues in Cognitive Sociolinguistics concerns the assumed homogeneity of the speech community, which has been proved wrong by socio-linguistic research. Socio-linguists have shown that there is structural variation in the language and, as such, social differentiation of the speech community is essential in every kind of linguistic research. In addition to that, Kristiansen (2008) adds that research along the second dimension i.e. Bell's "audience design" has shown that the granularity of a speech community should not stop at the level of social differentiation, but that style changes are to be taken seriously, as well. Finally, register studies in the line of Biber (1988) have shown that language also differs according to functional dimensions. All this leads to questioning the level of granularity that a (socio-)linguistic study should adopt. Therefore, this study compiled a corpus of texts from a single register (Usenet/CMC) and topic (politics), written by people with comparable demographic properties (well educated, old males) during one year (2009). Below, we will show that this tightly defined setting still leaves room for considerable variation along the dimension of "frequency of interaction".

Comparing linguistic behavior to social networks that are based purely on the frequencies of interaction is a research interest of scholars that study Computer-Mediated Communication (CMC). The interest in CMC language flourished in the 90s (Herring, 1994; Baym, 1996; Paolillo, 1996; Werry, 1996; Cherny, 1999; Paolillo, 1999), and was maintained by mainly Paolillo (Paolillo, 2001; Paolillo et al., 2005) and Herring (Herring and Paolillo, 2006; Herring et al., 2007, 2009) throughout the first decade of the 21st century. An exemplar case study is described in Paolillo (2001). He performed a Social Network Analysis on Internet Relay Chat (IRC) material from an Indian chat channel, correlating the network structure with linguistic features. Paolillo embeds his research question in the approach that Milroy (1987) proposed for (territorial) Social Networks. Basically, he operationalized the "Network Strength Scale" concept by counting the frequency of interaction. The study starts from the assumption that the findings of Milroy — stronger ties (c.q. higher frequency of interaction) are linked

with more vernacular linguistic norms — can be found in IRC material, as well. On the linguistic level, he looks at low-level orthographic substitutions, such as changing "s" into "z" like in the African American rap and hip-hop subculture. Another linguistic features is the code-switching into Hindi, which is not atypical for an Indian chat channel. Finally, the use of obscenity is counted, as well. As these are all forms of non-legitimized language use, they are claimed to be linked to the use of vernacular. From the findings of Milroy, these linguistic features should be more frequent among members of the strong-tie networks within the chat channel. For the network analysis, Paolillo (2001, p. 192-193) performs a positional analysis, based on Factor Analysis. The result of the Factor Analysis yields categories of participants, classified by having similar patterns of interaction. It showed that 16 participant groups could be ordered on "centrality", which Paolillo linked to tie strength. Paolillo (2001, p. 195-196) checks whether the interaction based groups are motivated by sex or region, and concludes that they are not. He does not have more demographic information on the participants that might confound the groups. To test the initial hypothesis that a more central, strongly tied group would show more vernacular characteristics, the retrieved structure was correlated to the linguistic variables using a logistic regression. However, the hypothesis proved to be wrong. As an explanation, Paolillo points out that a reinterpretation of the social and vernacular status of the linguistic variables might be needed. In the overwhelming flow of IRC messages that follow rapi8y on each other, it is hard to stand out and receive attention. Therefore, the linguistic variables are not only vernacular markers, but also attention-getting strategies. This partially obscures the hypothesized link between tie strength and the linguistic features. As can be seen from Paolillo (2001), this type of research draws heavily on theoretical insights from socio-linguists. Socio-linguists have indeed also studied the relation between linguistic variation and groups of people. In 1963, Labov used fishermen communities in his famous Martha's Vineyard study (Labov, 1963). Later on, he studied the language of Thunderbirds and Jets (Labov, 1972). In that period, during the 70s, Gumperz studied communication in communities from a strong linguistic viewpoint (Gumperz and Hymes, 1970). In the eighties, Milroy (1987) published the Belfast studies. Around that time, also, Eckert (1988) started her long-term study in Belten High.

3 Usenet and Social Network Analysis

The Usenet community is one of the oldest around in the computer world (Hauben and Hauben, 1997, Chapter 2-4). Born in 1979 as an academic experiment between Duke University, University of North Carolina at Chapel Hill and the Physiology Department of the Duke Medical School, Usenet was a low-cost alternative to ARPANET. ARPANET pioneered the networking technology that serves as the foundation of today's global Internet, but was only available through political connections and a lot of money. The main idea of the project was (and still is) to provide a rapid access newsletter, to which everybody can submit articles and replies. The popularity of Usenet grew almost exponentially during the beginning years. From this growth, and its consequent chaos, a structure with posts, threads and newsgroups emerged. Newsgroups are collections of threads around a central topic or theme and commonly hierarchically structured. For an important part of the newsgroups hierarchy, the highest level of the taxonomy is a country prefix. Everybody is free to start a new newsgroup, but the Usenet community has to approve a "charter" — a set of guidelines and mission

statements for the new newsgroups — before it is publicly available. Throughout, threads within a newsgroup can be posted without moderation and almost no newsgroups restrict membership. A thread is a collection of posts, organized in a "in reply to" structure.

Since Hauben and Hauben (1997) published their book, the on-line community situation has changed thoroughly. Not entirely the same as Usenet, a growing number of "social network sites" (SNS) have appeared. Boyd and Ellison (2007) introduce a definition and overview of this kind of websites up to 2007. Some important players in the SNS field are e.g. Facebook, MySpace and Windows Live Spaces, but also Couchsurfing, LinkedIn and Xing. In the SNS definition of Boyd and Ellison (2007), they emphasize the fact that on-line social networks are a reflection of real life social networks, and that "meeting new people" is not the main goal. This is almost the opposite of an on-line discussion platform as Usenet, where strangers discuss a wide variety of topics. Nevertheless, for the current purpose of this study, the emphasis is on people that interact with each other, and in that sense Usenet is not much different from the SNSs. Moreover, although all these SNSs are increasingly popular, Usenet still has the attention of the major players in the computer world. The recent (2009) efforts of Google Inc. to make the Usenet archive publicly available — and to make posting to the Usenet more easy through a web interface — are indicative of that. Moreover, Usenet is, apart from the many ad hoc fora on specific topics, the only community where people use longer stretches of text to make their opinions clear in public. Because of the large amount of textual data, Usenet is therefore the most appropriate medium for our study. One of the more important newsgroups in Belgian Usenet is "be.politics", which is still very active, with more than 4000 posts (going up to more than 8000) per month. The topic of the discussions is always related to current political, usually local Belgian affairs. "be.politics" is a tight and stable newsgroup with long-time members. Occasionally, In-Real-Life meetings take place, and some of the members revealed there actual name, giving up the anonymity of the Internet. The frequency of posting per hour shows us that the members regularly spend several hours per day on the newsgroup. Most members post between 17:00 and 23:00 hour. Therefore, "be.politics" is probably not their only way of communicating with others, but it is most certainly an important medium for them.

The nature of a Usenet discussion thread makes it very straightforward to retrieve the interaction patterns of the users. One could simply define interaction by checking which Usenet posters participate in one thread. However, the Usenet meta information of every post, stored in the header lines, makes a more fine-grained analysis possible. Every post (except the Original Post (OP), which is the starting point for the discussion) contains a header line that lists the previous posts that are being referenced. This reference header line keeps a historical track of the discussion: if the fifth posts references the fourth one, while the fourth one referenced other posts, the reference line of the fifth post will contain the fourth post, but also all references of the fourth post. By using this information, we also measure the frequency of interaction within a thread between specific users. This approach results in a so-called "edge list", where every poster is a "vertix" (plural: vertices) and each interaction is an "edge". We analyze this edge list for clusters by means of a Social Network Analysis ("igraph" package for R `http://www.r-project.org/` and a cluster finder, described by Clauset et al. (2004)). For simplicity reasons, we define the edge list as undirected. For the same reasons, we set a lower boundary to the amount of times (> 100) that two posters need to interact before their edge is taken into account. The result of this analysis can be

found in Figure 1. The basic principles of this analyses root in algebra. First, the adjacency matrix A of the edge list is calculated, which gives an idea of the closeness of every pair of vertices. Then, a matrix P containing the probability that there is an edge between each pair of vertices is subtracted from this adjacency matrix A, resulting in the modularity matrix B. This modularity matrix B gives a numerical idea of how well a division in clusters would be. From this modularity matrix B, the eigenvector for the largest positive eigenvalue is calculated. Based on the sign of their corresponding element in that eigenvector, vertices are split into two clusters. If all vertices have the same sign, there is no underlying cluster structure.

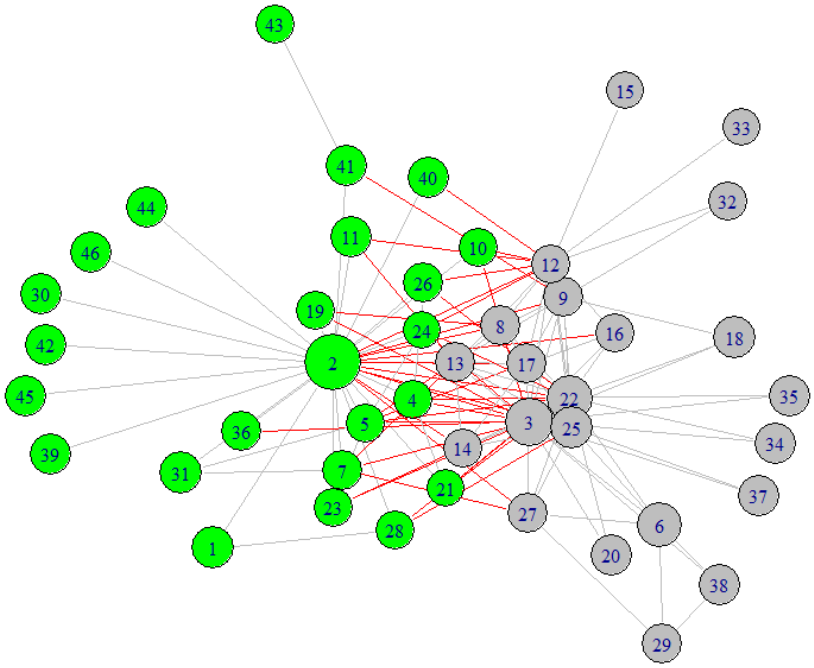

Figure 1: *Clusters in the Usenet be.politics group, 2 Newmann clusters*

This graph reveals at least two clusters within "be.politics" around "2" (cluster 1) and "3" (cluster 2), with respectively 24 and 22 members. This does not strictly mean that it is impossible to find even more clusters, but it definitely means that there is at least a twofold structure based on the interaction patterns. For the sake of simplicity in the statistical analysis below, we will use this two-cluster solution. The corpus then consists of Usenet posts, selected on two criteria. First, they need to be written by a member of these clusters, and second, at least one of the addressees of the post should be a member. Of course, the posts come from the 2009 "be.politics" newsgroup.

4 Demographic background

It is possible that the two interaction-based clusters from the Social Network Analysis resemble an off-line demographic speech community structure. Especially on a

culturally important topic as politics, we might find that people with the same socio-economic background interact more often. However, as the qualitative analysis below will show, this hypothesis does not seem valid for the "be.politics" newsgroup, where people from a relatively constant demographic background seem to interact. We observe the clear tendency that the average "be.politics" user is an old well educated male from the socio-economic center of Belgium (the geographic center and the region to the west of that center).

The qualitative analysis actually began with a small scale inquiry in the newsgroup. This provided us with first hand demographic information for some of the members. Other members link to their personal website that contains a curriculum vitae. This information can be found in Table 1. For a number of the members, we can find some of their demographic background in their posts on Usenet. For this, we created a small Python script that filtered out the posts of a certain member and searched in the texts of these posts for a number of demographic clues. These demographic clues are listed in Appendix D. Concerning the privacy of the members, we used mapped their Usenet "nicknames" to anonymous numbers. Therefore, we feel confident that this study is ethically acceptable. An overview of the inferred demographic information can be found in Table 2

alias	year of birth	occupation	education	region	sex	cluster
28	1940s	consultant	university	central	M	1
36	1970s	shop owner	college	west	M	1
3	1930s	retired doctor	university	west	M	2
13	1940s	electrician	secondary school	NA	M	2
9	1950s	electrician	college	west	M	2

Table 1: *Direct demographic information*

Although we were not able to get a full demographic picture of the clusters, it appears from this sample that there are no demographic preferences for cluster membership. Both clusters have mostly older, well educated males. Although the sample is limited, it seems to show that there is no bias towards demographic background in the construction of the clusters. This is in itself a remarkable observation, that goes against the normal assumptions of socio-linguistic research. However, an acceptable explanation for this phenomenon might come from the nature of the Usenet register. As mentioned in Section 3, Usenet is a very democratic medium, open to everybody who has an Internet connection. As such, people from many different demographic backgrounds are able, and will, take part in the discussion. Although apparently mostly older and well educated males take up an interest in politics, their demographic characteristics are to a certain extent hidden, in so far as the participants choose. This is a first, passive explanation for the fact that clusters do not seem to correlate with demographic background. An active explanation can be found as well. As it seems that the "be.politics" members do not mind to reveal a part of their demographic background, the indirect nature of Computer Mediated Communication might encourage the members to cross the demographic borders.

5 Quantitative analysis

As mentioned above, the variationist part of this study is corpus-based and therefore, we will not look into phonological features. Rather, a collection of functional features,

alias	age	occupation	education	region	sex	cluster
2	old	NA	NA	NA	M	1
23	old	technical	high	west	M	1
10	old	NA	high	NA	M	1
11	old	NA	high	NA	M	1
19	old	NA	NA	NA	NA	1
5	young	administratio	high	west	M	1
42	old	computers	high	NA	M	1
26	old	NA	high	NA	NA	1
24	old	teacher	high	NA	M	1
31	old	NA	NA	NA	M	1
45	old	NA	NA	NA	M	1
total (with	old: 9		high: 8	center: 1	M: 11	
direct	young: 2		modest: 1	west: 3	F: 0	
information)	NA: 1		NA: 5	NA: 9	NA: 0	
8	young	administration	high	NA	M	2
22	old	NA	NA	center	F	2
12	old	programmer	high	NA	M	2
13	old	NA	NA	NA	M	2
14	young	NA	high	NA	M	2
20	NA	NA	NA	NA	M	2
15	young	NA	NA	NA	M	2
16	NA	NA	NA	NA	M	2
17	NA	computers	high	NA	M	2
29	old	NA	NA	center	M	2
32	old	NA	high	NA	M	2
37	old	NA	modest	NA	M	2
33	old	consultant	high	NA	M	2
total (with	old: 10		high: 8	center: 2	M: 12	
direct	young: 3		modest: 2	west: 2	F: 1	
information)	NA: 4		NA: 6	NA: 11	NA: 0	

Table 2: *Inferred demographic information*

as used in the register studies of e.g. Biber (1988), some morpho-syntactic alternation variables and a number of lexical markers will be used. The list of functional features is given in Appendix A. For the morpho-syntactic alternation variables, we fall back on De Sutter et al. (2005) and Tummers et al. (2005), who explored regional and stylistic variation in the word order in Dutch verb clusters and the inflection of adjectives in Dutch neuter noun phrases. It is also possible to find lexical alternation variables by means of a synonymy repository, such as EuroWordNet (Vossen, 1998). It has been shown (Geeraerts et al., 1999) that the choice for one of the variables might correlate with sociological (regional) and stylistic variation. Lexical markers — instead of synonymous alternation, sheer frequency patterns are observed — will typically reveal topical differences, but Speelman et al. (2006) has shown that certain lexical markers — so called Stable Lexical Markers — neutralize the topical bias and can be used for the discovery of other kinds of variation. In the current (preliminary) version of this paper, we did not perform a full Stable Lexical Marker analysis, although this is planned for the final version. Instead, we took a number of highly frequent, politics-related keywords from a frequency list for "be.politics". Significant differences in appearance of these words will most probably signal politics-internal topic differences. The current study will look into these three different kinds of variables and compare them to the interaction based cluster structure of the Usenet newsgroup. Previous research has shown that each of these variables appears to link up with a certain type of extra-linguistic variation. Biber's functional features distinguish registers and text-types; alternation variables have been strong indicators of traditional sociological characteristics; and lexical features are related to topical differences. However, cross-overs to link up these features with their atypical extra-linguistic variation have been attempted

by aforementioned Speelman et al. (2006) and Argamon et al. (2007).

The corpus then consists of Usenet posts, selected on two criteria. First, they need to be written by a member of these clusters, and second, at least one of the addressees of the post should be a member. Of course, the posts come from the 2009 "be.politics" newsgroup. We annotated the posts of these clusters automatically by means of a number of Python scripts. This resulted in an observation table of the format of example Table 3. As the posts are relatively short (on average, only 44 words, ranging between 1 and 1964 words), we combined posts into larger sub-corpora. The combination is based on two criteria. First, we combined all posts from each sender (this results in 46 rows in the table). Then, we split these observations again according to "interaction pattern" (see below), which expanded the table to 126 rows (an example of this Table for analysis is also in Table 4). As the observations contain relatively different amount of words (mean: 2824 words, median: 991.5 words), we make the frequency counts of the linguistic features relative to text length (this is not uncommon, cf. Biber (1988)). As mentioned before, some of the features are alternation variables, which are not represented as frequencies relative to the amount of words, but as odds, e.g. the chance for flexed adjective in neuter noun phrases viz. all adjectivized neuter nouns phrases.

	sender	from	to	interaction	var_1	var_2	var…	var_i
$post_1$	sender1	1	1	1-1	f	f	…	f
$post_2$	sender1	1	2	1-2	f	f	…	f
$post_3$	sender2	2	1,2	2-1,2	f	f	…	f
…								

Table 3: *Design of the observation table*

sender	from	communication	var_1	var_2	var…	var_i
$sender_1$	1	within1	f	f	…	f
$sender_1$	1	1-mixed	f	f	…	f
$sender_2$	2	2-mixed	f	f	…	f
…						

Table 4: *Design of the table for analysis*

The previous studies mentioned above make use of multivariate clustering techniques on the side of the extra-linguistic features (De Sutter et al., 2005; Tummers et al., 2005), or on the side of the linguistic features (Biber, 1988). Although we will of course also make use of these clustering models later on, we first want to observe whether there is actual variation in the use of the linguistic features throughout the different clusters by means of simple univariate Analyses of Variance. Therefore, we use the Table of Analysis (Table 4) to answer two questions. Due to non-normal distributions of some of the linguistic features, we transformed the data with a square root. The first question starts from the cluster membership of an author: "Is the average use of a certain linguistic feature different between the two clusters?" The second question takes the interaction pattern into account: "Do people change their language use when they address other people?"

To answer the first question, we compare all posts from cluster 1 to those of cluster 2 for each linguistic feature. Table 5 shows that a number of features are significantly different between the two groups.

For the second question, we compare the interaction patterns for people from a single cluster. We perform Analyses of Variance to see whether communication within the

feature	from2 est.	p-value	variable
belgie	-0.599	0.000	chance of "belgie" viz. "belziek"
belziek	0.599	0.000	chance of "belziek" viz. "belgie"
bepol	0.005	0.013	#bepol / textlength
nederlands	0.010	0.000	#nederlands / textlength
passive	0.0141	0.024	#passives / textlength
perfect	0.011	0.094	#perfect / textlength
prepositions	0.02	0.032	#prepositions / textlength
textlength	21.509	0.000	textlength
vlaming	0.008	0.027	#vlaming / textlength
waal	0.005	0.054	#waal / textlength
waals	0.004	0.036	#waals / textlength

Table 5: *Differences between messages from cluster 1 versus cluster 2*

cluster is different from communication across the cluster. The results are presented in Table 6 and 7.

feature	1-mixed est.	1-mixed p-value	variable
allcaps	-0.024	0.065	#allcaps / textlength
flik	-0.321	0.015	#flik / #politie
politie	0.119	0.018	#politie / #flik
indirect.objects	-0.018	0.045	#indirect object / textlength
passive	-0.015	0.094	#passives / textlength
past	-0.033	0.017	#past / textlength
prepositions	-0.051	0.005	#prepositions / textlength
questions	-0.037	0.002	#questions / textlength
secondPersonPronouns	-0.037	0.018	#second person pronouns / textlength
vb	-0.172	0.092	#vb / political parties
vld	-0.128	0.011	#vld / political parties

Table 6: *Comparing communication within cluster 1 to communication from cluster 1 to a mixed audience*

feature	2-mixed est.	2-mixed p-value	variable
allochtoon	0.288	0.034	#allochtoon / #vreemdeling
vreemdeling	-0.29	0.058	#vreemdeling / #allochtoon
democraat	-0.762	0.003	#democraat / #tsjeef
tsjeef	0.632	0.018	#tsjeef / #democraat
firstPersonPronouns	-0.027	0.053	#first person pronouns / textlength
vlaming	-0.01	0.048	#vlaming / textlength
waals	-0.006	0.063	#waals / textlength

Table 7: *Comparing communication within cluster 2 to communication from cluster 2 to a mixed audience*

In the first step of this analysis, we wanted mainly to establish on a basic level whether there is actual variation between these clusters. The Analyses of Variance have shown that there are some differences in linguistic behavior between the two clusters on lexical and grammatical level, which might indicate stylistic and topical variation. A multivariate analysis might show us which linguistic features have a similar variational pattern, and this might allow us to find a more thorough explanation for the linguistic differences.[1]

Analyzing the separate Analyses of Variance, however, already gives an idea of the linguistic properties of the clusters. First, we compare the language of the two clusters

[1]This endeavor will be part of the final version of this paper.

to each other. A first group of features is clearly lexical (belgie, belziek, bepol, nederlands, vlaming, waal, waals). The features nederlands, vlaming, waal, waals might indicate a topical difference between the groups: cluster 2 seems to address the Belgian federal structure more often than cluster one. The fact that they rather use a nickname "Belziek" than "België" for the country shows that they have a negative view on the federal structure. The feature "bepol" shows that the shibboleth "bepol" (a lexicalized, adjectivized abbreviation for "be.politics") is a group identifying word for cluster 1 members. The second group of features has a grammatical nature (passive, perfect, prepositions). Passive and perfect constructions are syntactically more complex structures than active voice and simple past because of the presence of an auxiliary. This seems to indicate that cluster 2 members write in a somewhat more complex style than members of cluster 1. The fact that they also write longer texts seems to corroborate this. Finally, the augmented use of prepositions might indicate a more "informational" style, if we bluntly translate the findings of Biber (1988) to Dutch.

Secon8y, the differences between communication within the cluster is compared to communication across the clusters. We split the analysis into two parts. First, we check whether members of cluster 1 alter their language use depending on the audience they address. Then, we do the same thing for cluster 2. For cluster 1, Table 6 shows that there are again some differences between speaking for your own cluster and speaking for a "mixed" audience. On the lexical level, we notice that the cluster 1 members use the standard word for police more than the nickname "flik". In combination with the decrease in shouting (all caps), this might indicate a more "civilized" tone. Also, some political parties are not so often mentioned in their posts when speaking to a mixed audience. On the grammatical level, the diminished use of questions and second person pronouns indicates a less interactional style. This ties in with the reduced amount of indirect objects as indirect objects (in Dutch) occur usually with verbs of transfer or communication. This possible reveals a less personal and more distinct way of communication. For cluster 2, Table 7 shows that the linguistic chances that cluster 2 members make when addressing a mixed audience are much more subtle. Mostly lexical changes take place. On a topical level, cluster 2 tends to speak a bit less about the federal structure of the country, when addressing a mixed audience, by using keywords such as "vlaming" and "waals" less. On a stylistic level, they however tend to use the nickname of democrats ("tsjeef") more, and the neutral term ("democraat") less than in their cluster internal discussions. Also, the more technical term for foreigner ("allochtoon") is used more often — and the common term ("vreemdeling") less often — than in posts within the cluster.

6 Conclusions

To conclude this paper, two important questions come to mind: "why are there two clusters within one discussion group of about 50 members, and are the reasons for this dichotomy not confounding the findings above?" and "why actually can we observe significant differences in the use of certain linguistic features between two clusters?" A first and obvious answer for the first question has to do with the number of participants in a discussion. A discussion with the full group of 50 members is practically impossible. As can be seen in Figure 2, the amount of people (other than the original poster) that take part in the discussion is usually just one, and only rarely more than 10.

Therefore, it is understandable that small groups of people will start to form. This

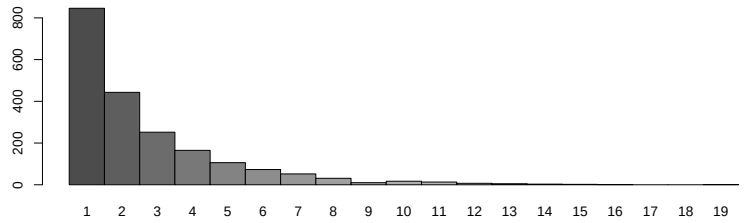

Figure 2: *Amount of threads in which "x" users partake (other than the original poster).*

leads to a second reason for group forming: shared interests or stance. It seems reasonable that members will engage in a discussion if the topic of the discussion interests them or if they think that the original post is worth commenting on. In addition to that, we would like to put emphasis on the fact that — within the Usenet register and this particular newsgroup — the demographic background does not seem to limit the social interaction. It appears that highly educated people discuss political matters with people of modest education, regardless of their age or (regional) origin. In that sense, "be.politics" and possibly the Usenet register is a unique social environment in which unusual socio-linguistic findings are possible.

The third reason for cluster building might be personal preferences with regard to other members. The "be.politics" newsgroup is already fairly old, and it might be that personal feelings — just like in off-line communities — have overcome the relative anonymity of Usenet. In that case, Eckert-style "social meaning" can be attached to the linguistic differences. However, this is not a possible confounding feature for our analysis, but rather an explaining one. As such, this might be an answer — even if it is only a partial answer — to the second question.

It is possible to find even more reasons for the construction of cluster within "be.politics", and — by doing so — to point out possible extra-linguistic confounding features for the analysis. We have e.g. not mentioned "emotions" (is the discussion frien8y or a fight, humorous or serious, etc.) or "time of the day" (later on in the day, people might get tired). However, trying to take care of all these possible confounding features will most certainly lead to data sparseness, making a statistical analyses inappropriate.

Therefore, we claim to have found at least an indication of the fact that variation in language use within this single ancillary register is related to social interaction. Our analyses have shown that groups of frequently interacting people tend to apply certain linguistic characteristics differently. We have found that one cluster uses a syntactically more complex language than the other. Moreover, when the clusters interact with each other, they change their language use as well. One cluster seemed to adopt a more civilized tone when addressing the other cluster.

These first preliminary results encourage further research. A significant amount of work is to be done. The most important job is to try and remedy as many possible confounding extra-linguistic features. In that light, it should be possible to balance the (sub)topical differences between the groups and measure stance. Finally, we should

broaden the perspective and deliberately take extra-linguistic parameters into account as well. For this, we could stay within the Usenet register and take e.g. the "be.sports" newsgroup as a topical reference value. Moreover, adding an extra register to the analysis is also possible. To do this, we might add political and sports articles from a newspaper corpus to the analysis. An analysis in which both the linguistic side as well as the extra-linguistic side have a multivariate structure forms the larger goal of the project in which the current study has been performed.

References

S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

N. Baym. Agreement and disagreement in a computer-mediated group. *Research on Language in Social Interaction*, 29:315–346, 1996.

A. Bell. Language style as audience design. *Language in Society*, 13(2):145–204, 1984.

D. Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

R. Bourhis and H. Giles. The language of intergroup distinctiveness. In H. Giles, editor, *Language, Ethnicity and Intergroup relations*, pages 119–135. London: Academic Press, 1977.

D. Boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1): http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html, 2007.

L. Cherny. *Conversation and Community: Chat in a virtual world*. Stanford, California: CSLI Publication, 1999.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review*, 70(6):066111, 2004.

N. Coupland. *Style: language variation and identity*. Cambridge University Press, 2007.

N. Coupland. "hark, hark the lark": Social motivations for phonological style-shifting. *Language and Communication*, 5(3):49–70, 1985.

G. De Sutter, D. Speelman, and D. Geeraerts. Regionale en stilistische effecten op de woordvolgorde in werkwoordelijke eindgroepen. *Nederlandse Taalkunde*, 10: 97–128, 2005.

P. Eckert. *Linguistic variation as social practice*. Oxford: Blackwell., 2000.

P. Eckert. Sound change and adolescent social structure. *Language in Society*, 17: 183–207, 1988.

P. Eckert. Age as a sociolinguistic variable. In F. Coulmas, editor, *The Handbook of Sociolinguistics*, pages 151–167. Malden, Massachusetts: Blackwell, 1997.

P. Eckert and J. R. Rickford, editors. *Style and sociolinguistic Variation*. Cambridge University Press, 2001.

E. Finegan and D. Biber. Register and social dialect variation: an integrated approach. In D. Biber and E. Finegan, editors, *Sociolinguistic Perspectives on Register*. Oxford University Press, 1994.

D. Geeraerts, S. Grondelaers, and D. Speelman. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut, 1999.

H. Giles and P. Powesland. *Speech style and social evaluation*. Number 7 in European monographs in social psychology. Academic Press, 1975.

J. Gumperz and D. Hymes. Sociolinguistics and communication in small groups, working paper 33. In J. Pride, editor, *Readings in Sociolinguistics*. London: Penguin, 1970.

P. Harder. *The social Turn in cognitive linguistics*. Berlin/New York, Mouton de Gruyter, Forthcoming.

M. Hauben and R. Hauben. *Netizens: On the history and impact of usenet and the internet*. IEEE Computer Society Press: Los Alamitos, California, 1997.

S. Herring. Politeness in computer culture: Why women thank and men flame. In M. Bucholtz, A. Liang, L. Sutton, and C. Hines, editors, *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*, pages 278–294. Berkeley, California: Berkeley Women and Language Group, 1994.

S. Herring and J. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10/4:439–459, 2006.

S. Herring, J. Paolillo, I. R. Vielba, I. Kouper, E. Wright, S. Stoerger, L. Scheidt, and B. Clark. Language networks on livejournal. In *Proceedings of the Fortieth Hawaii International Conference on System Sciences*, 2007.

S. Herring, D. Kutz, J. Paolillo, and A. Zelenkauskaite. Fast talking, fast shooting: Text chat in an online first-person game. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42)*, 2009.

G. Kristiansen. Style-shifting and shifting styles: A socio-cognitive approach to lectal variation. In *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Mouton De Gruyter, Berlin, 2008.

G. Kristiansen, J. R. T. Langacke, R. Dirven, and D. Geeraerts, editors. *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Cognitive Linguistics Research. Mouton De Gruyter, Berlin, 2008.

W. Labov. The social motivation of a sound change. *Word*, 19:273–309, 1963.

W. Labov. *The social stratification of English in New York City*. Center for Applied Linguistics, 1966.

W. Labov. *Language in the Inner City*. Philadelphia: Pennsylvania University Press: Oxford: Blackwell, 1972.

W. Labov. The intersection of sex and social class in the course of linguistic change. In *Language Variation and Change*, volume 2. Cambridge University Press, 1990.

L. Milroy. *Language and social networks*. London; Baltimore: Basil Blackwell; University Park Press, second edition, 1987.

J. Paolillo. Language variation on internet relay chat: A social network approach. *Journal of Sociolinguistics*, 5(2):180–213, 2001.

J. Paolillo. Language choice on soc.culture.punjab. *Electronic Journal of Communication*, 6(3), 1996. http://www.cios.org.

J. Paolillo. The virtual speech community: Social network and language variation on irc. *Journal of Computer-Mediated Communication*, 4(4), 1999. http://www.ascusc.org/jcmc.

J. Paolillo, S. Herring, I. Kouper, L. Scheidt, M. Tyworth, P. Welsch, and E. Wright. Conversations in the blogosphere: An analysis from the bottom up? In *Proceedings of the 38th Hawaii International Conference on System Sciences*. Los Alamitos: IEEE Publications, 2005.

D. Preston. Sorting out the variables in sociolinguistic theory. *American Speech*, 66: 33–56, 1991.

D. Speelman, S. Grondelaers, and D. Geeraerts. A profile-based calculation of region and register variation. In A. Wilson, D. Archer, and P. Rayson, editors, *Corpus linguistics around the world*, volume 56 of *Language & Computers*, pages 181–194. Rodopi, Amsterdam - New York, 2006.

J. Tummers, D. Speelman, and D. Geeraerts. Inflectional variation in belgian and netherlandic dutch: A usage-based account of the adjectival inflection. In N. Delbecque, J. van der Auwera, and D. Geeraerts, editors, *Perspectives on Variation. Sociolinguistic, Historical, Comparative*, pages 93–110. Berlin: Mouton De Gruyter, 2005.

P. Vossen, editor. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht, 1998.

E. Wenger. *Communities of practice*. New York: Cambridge University Press, 2000.

C. Werry. Linguistic and interactional features of internet relay chat. In *Computer-Mediated Communication: Linguistic, Social and Cross-cultural Perspectives*, pages 47–64. Amsterdam: Benjamins, 1996.

A List of functional features

textLength, adjectives, attributiveAdjectives, predicativeAdjectives, adverbs, conjuncts, interjections, nouns, prepositions, verbs, direct.objects, indirect.objects, subclauses, futur, infinitives, past, aux, koppelwerkwoord, passive, perfect, firstPersonPronouns, secondPersonPronouns, thirdPersonPronouns, standard.spp, formal.spp, informal.spp , allcaps, questions, totalWor8ength, types, tokens samenstellingen

(relative to textlength, except for textlength itself)

B List of alternation variables

(rood, groen), (onverbogen.neutraAdjFlex, verbogen.neutraAdjFlex), (vb, cd.v, nva, ldd, spa, vld), (socialist, sos), (tsjeef, democraat), (allochtoon, vreemdeling), (zei, zegde), (wilde, wou), (politie, flik), (belziek, belgie)

C List of lexical features

bepol, poco, vlaams, waals, nederlands, vlaming, waal

(relative to textlength)

D Demographic clues

- 2: “Dat mijn vrouw zo gebrekkig Frans spreekt is ook hinderlijk” (that my wive speaks so little French is also annoying), “Wat dan met mijn vrouw en kinderen” (what about my wife and kids), “dat vervloekte Antwerps van mijn Vlaamse kinderen” (that damned Antwerp dialect that my Flemish kids speak)
- 23: “mijn kleinkinderen” (my grandchildren), “verbleef ik voor mijn werkgever” (I resided for my employer), “mijn vrouw, ook een gentse” (my wife, from Gent also), “ik werk al meer dan 35 jaar bij dezelfde baas” (I work for the same employer for more than 35 years), “Aangezien ik nu bijna met pensioen mag - (1/8/09; geboren 1/7/09) dus op mijn 65ste” (Since I can almost retire on my 65th birthday), “Ik werk bijvoorbeeld ook voor (niet bij) Sonelec alwaar ik in de jaren ’70 o.a. de ”Usine de Lampes” (in Sahouria / Mohammedia / Wilaya de Mascara) heb helpen opstarten. ” (I do not work for Sonelec where I helped starting up the Usine de Lampes): old highly educated male from the west
- 5: “ik woon in een gentse gemeente” (I live in a suburb of Gent), “wat jullie generatie er al doorgejaagd heeft” (what your generation has spent already), “ik en mijn vrouw” (me and my wife), “ik zal dat eens moeten aankaarten bij mijn federale collega. (I should address my federal co-worker about this)”: highly educated young male from the west
- 8: “mijn vrouw is een buitenlandse” (my wife is from abroad), “Mijn vrouw heeft het Nederlands niet als moedertaal” (my wife does not have Dutch as her mother tongue), “ik werk in een federaal bedrijf” (I work in a federal company), “En kids, sjah, eerst aan enkele compatibiliteitsproblemen werken” (and kids, well, first sort out some compatibility problems), “Zo kan het zijn dat ik het mooiste project van collega’s, die zich bezig houden met de implementatie, toch nog een onvoldoende geef” (It is possible that I flunk the most beautiful project of co-workers, who work on the implementation): younger highly educated male
- 10: “Niet alleen werd nu de sakosj van mijn vrouw leeg gemaakt” (Not only was the purse of my wife emptied), “[als ik] mijn kinderen en kleinkinderen kan zien opgroeien is dat voor mij voldoende” (if I can see my children and grandchildren grow up, I’m happy), “Zonder mij een dikke nek te moeten aanmeten ben ik beter thuis in de micro/nono electronica, magnetische velden” (Without boasting, I am better in micro/nono electronics, magnetic fields): old, highly educated male
- 11: “nochtans studeerde ik tot mijn 22ste” (I studied until I was 22), “En ik heb linguistiek gestudeerd, dus dat zijn er behoorlijk wat.” (And I studied linguistics, so that are quite some), “Op de school van mijn kinderen was er tien jaar geleden een idioot die de boel in de fik heeft gestoken” (In the school of my children, some ten years ago, there was a guy who torched everything down): highly educated old male

- 12: “toen ik daar meer dan 20 jaar geleden studeerde” (I studied there 20 years ago), “vooral in hitechmekka SV en SFO, en wanneer ik daar studeerde” (especially in high technology mechanics SV and SFO, and when I studied there): old highly educated male
- 13: “die mij en mijn vrouw toebeet” (who snapped at my wife), “mijn kinderen lopen hier minstens een keer per week binnen” (my children pass by at least once a week), “nadat mijn beide dochters afgestudeerd waren” (after the graduation of both my daughters), “opdrachten waar ik geen zin in heb schuif ik door naar een jongere collega” (I pass assignments that I do not like on to my younger co-worker): old male
- 14: “ ik heb ook in Groningen gewerkt/gestudeerd voor mijn phd” (I worked/studied in Groningen for my PhD): highly educated male
- 20: “zodadelijk gaat mijn vrouw naar de film” (in an instant, my wife is going to the movies): male
- 15: “onze 4 (kleine) kindjes zijn dan bij iemand anders”, “: mijn vrouw speelt geregeld ”Impromptus et Moments Musicaux” op de piano.”: male, young
- 16: “wat heeft mijn vrouw hier mee te maken” (what has my wife to do with this?): male
- 17: “mijn vrouw en ik” (my wife and I), “Ik werk zelf met Xen in het datacenter” (Personally, I work with Xen in the datacenter): highly educated male
- 29: “zolang ze mijn vrouw gerust laten” (as long as they leave my wife alone), “ik woon al dertig jaar in het statiekwartier van Antwerpen” (I live in the sailors quarter of Antwerp for 30 years now): old male
- 42: “ik werk al vele jaren in IT” (I work in IT for many years now): old highly educated male
- 19: “Als mijn kinderen het over mama hebben, spreken ze beide lettergrepen op dezelfde -redelijk snelle- manier uit” (if my kids speak about “mama”, they pronounce both syllables in the same — rather quick — way), “ ik ben een Vlaming en ik woon in Vlaanderen” (I am Flemish and I live in Flanders), “voor mijn vrouw haar verjaardag had ik anti-rimpelcreme gekocht” (I bought anti-aging creme for the birthday of my wife): old male
- 26: “Ik werk in de prive, en ben dus geen ambtenaar” (I work in the private sector and I am therefore not a ?), “Ik heb dus 7 jaar kunnen werken terwijl ik thuis een peulschil betaalde voor kost en inwoon” (I was able to work for 7 years while I lived at home and payed almost nothing for living and eating), “ik heb een leuk huisje, lieve vriendin en goedbetaalde job” (I have a nice house, a sweet girlfriend and a well payed job): highly educated and young
- 24: “Vooral Ghingis slaagt erin me elke dag weer te verbazen - wat mijn vrouw allang niemeer kan”, “In het begin der zeuventiger jaren van vorige eeuw - jaa, de tijd vliegt - was ik een schoolcollega van Helmut’s vader Luc (die godsdienst doceerde!)” (In the beginning of the 70s, I was a co-worker with Helmut’s dad Luc (who taugth religion): male, old, highly educated
- 31: “Zowel ik al mijn vrouw komen uit een gezin die het niet breed hadden” (Both I and my wife come from families that did not have much money), “Dus als ik een derde kind zou willen met mijn vrouw” (so if I would like to have a third child with my wife): old male
- 32: “moment was dat mijn vrouw een bubbelbad nam.” (at that moment, my wife was taking a bath), “De geomorfologie van de grote geulen verandert niet dramatisch, dat vak heb ik gestudeerd” (the geomorphology of the larger ? does not change dramatically, I studied that course), “We zijn inmiddels 17 jaar getrouwd” (we are married for 17 years now): highly educated old male
- 1: “Ik werk in een bedrijf waar ik af-en-toe wel eens prive-gegevens van personen zie passeren.” (I work in a company where now and than private information of people pass by): highly educated old male

- 37: “Niet alleen mijzelf.. maar ook mijn vrouw en kinderen.” (not only I, but also my wife and children), “ik ben al 25 jaar gelukkig gescheiden” (I am happily separated for 25 years now), “Of ze een gemeenschappelijke toekomst hebben.. dat laat ik aan de afgestudeerden hier..” (whether they have a common future, that I leave to the people who have studied): modestly educated old male
- 45: “want mijn kinderen gaan ook naar school” (because my children go to school as well), “dat mijn vrouw een paar keer loopbaanonderbrekeing heeft genomen” (that my wife took a suspension of her job): old male
- 33: “toen ik daar nog neurofysiologie studeerde” (when I studied neurophysiology), “In die functie heb ik 20 jaar lang expertises gedaan voor de gerechtsarts” (In that function, I consulted for 20 years the justice doctor), “ toen ik mijn vrouw - god hebbe haar ziel - leerde autorijden, was het een uitgemaakte zaak” (when I taught my wife — God have her soul — how to drive, it was a clear case): highly educated old male
- 22: “wanneer ik zwanger was van mijn oudste kind” (when I was pregnant with my first child), “Ik heb 4 kinderen” (I have four children), “ik woon in Antwerpen” (I live in Antwerp): female, old, living in the center