Mapping the Structure of Language Repertoires: on the Use of Sociolectometric Methods

Abstract: In this paper, we will discuss the importance of (socio)lectometric research in attempts to empirically map the range and internal structure of language repertoires, on both the level of the individual and the level of the speech community. Spoken data from 10 East Flemish women (city of Ghent) in 5 speech settings are analysed using 3 multivariate statistical techniques: correspondence regression, multidimensional scaling, and cluster analysis. These analyses show that on a community level, the language repertoire in Ghent can be called – following Auer (2005) – “diaglossic”, with a continuum stretching from dialect to standard language with intermediate variations in between. In this continuum, focal points or prototypes can be identified, such as dialect, tussentaal and Standard Dutch. On the level of the individual language user, however, the repertoires are not necessarily diaglossic. Methodologically, we will argue that different multivariate techniques are ideally combined and complemented, as each technique has both advantages and drawbacks, and only structures emerging in multiple analyses can be considered robust. The Flemish case-study will moreover demonstrate that the techniques can only be usefully implemented in repertoire structure studies when stylistically stratified datasets are available and a diverse set of language features is observed.

Keywords: Language repertoire, sociolectometry, Dutch, cluster analysis, correspondence regression, multidimensional scaling

1. Introduction

In his by now famous 2005-article, Auer distinguishes five types of dialect/standard constellations in Europe, depending on (1) the endo- or exoglossic nature of the standard, (2) the functionality of the standard as an exclusively written or also spoken medium, (3) the presence or absence of a continuum of intermediate variants in between standard and dialect, and (4) the vitality of the spoken dialects. Whereas diglossic repertoires assume that language users switch between structurally and functionally separated systems (such as dialect and standard language), diaglossic repertoires imply more subtle style-shifts on a continuum that stretches from dialect to standard and comprises a whole range of intermediate variation. In the last decade, Auer’s theoretical framework has served as a starting point for several European linguists characterizing the language repertoires in their research areas (see for example Rys/Taelde man 2007, Gooskens/Kürschner 2009, Hernández-Cam poy/Villena-Ponsoda 2009), and has increased interest in generalizable patterns and dynamics. The question however arises how the nature and internal structure of language repertoires can be empirically established. In this paper, we will discuss the usefulness of (socio)lectometric research in attempts to empirically map language repertoires on both the level of the individual and the level of the speech community. We first elaborate on Auer’s typology of dialect/standard constellations, after which we introduce the Ghentish dataset that will be central in this paper. In sections 4, 5 and 6, the introduced dataset is analysed using three multivariate statistical techniques – correspondence regression, multidimensional scaling and cluster analysis respectively – which will lead to a discussion in section 7 of repertoire structure in Ghent and appropriate ways to map the structure of language repertoires.

2. Nature and structure of language repertoires

The term “language repertoire” is traditionally used to refer to the “totality of linguistic resources (i.e. including both invariant forms and variables) available to members of particular communities” (Gumperz 1972: 20–21, cf. also Berruto 1989, Auer 2005). Blommaert/Backus (2013) however raise the question what is exactly meant with
“community” in Gumperz’ definition and argue that the concept of stable communities is anachronistic in times of “superdiversity”. In their view, the individual should be at the centre of scientific attention. Blommaert/Backus (2013) therefore redefine language repertoires as unique “indexical biographies, […] biographically assembled patchworks of functionally distributed communicative resources, constantly exhibiting variation and change” (Blommaert/Backus 2013: 23). This definition strongly reminds of Hymes’ concept of “communicative competence” (Hymes 1972), the unique structural, functional and social language knowledge of a language user. The unicity of individual repertoires or competences does however not necessarily imply the bankruptcy of the community repertoire or the “Gesamtsprachsystem” (Schmidt/Herrgen 2011: 31); an important part of language variants, norms and attitudes is after all shared among broader groups of language users (cf. Geeraerts 2010, Schmidt/Herrgen 2011). In this paper, we therefore use the term “individual language repertoire” to refer to the totality of linguistic resources available to a specific speaker, and we assume the existence of general, higher-level structures which can then be considered typical of the community repertoire.1

In 1989, Gaetano Berruto remarked that “[t]here are many works on the functional and social status of languages and language varieties, but [that] relatively few of them deal specifically with the nature and the internal qualitative structure of repertoires” (Berruto 1989: 552). Even though almost 30 years have passed in the meantime, his observation still holds for many speech communities. Theoretically, however, significant headway has been made with Auer’s typology of European dialect/standard constellations. This typology distinguishes between five types of dialect/standard constellations, which are schematically represented in Figure 1:

1. “exoglossic diglossia”, assuming code-switches between dialects as colloquial language and an exogenous or exoglossic standard as “high variety”;
2. “medial diglossia with an endoglossic standard”, i.e. repertoires with two clearly separated varieties, an endogenous standard for writing and dialects for speech;
3. “spoken diglossia”, with dialects as informal spoken language and standard for both writing and formal speech settings;
4. “diaglossia”, with between dialect and standard a whole continuum of intermediate variations, and
5. “dialect loss repertoires”, in which the dialect end of the repertoire has eroded.

Many concepts in Auer’s typology are anything but new – the concept of diglossia can for instance be traced back to Ferguson (1959), whereas the idea of diaglossia is related to Berruto’s dilalia (Berruto 1989) and to earlier work on intermediate variaties in between dialect and standard language (Hufschmidt et al. 1983, Hoppenbrouwers 1990) – but the added value of the framework is that it interrelates the different repertoire types diachronically. Diaglossia, according to Auer (2005) the most frequently

---

1 A community can be defined at multiple levels: an individual for instance belongs to a certain gender group, but also to the community of a certain social class, a regional community, …
Fig. 1: Auer's (2005) typology of European dialect/standard constellations schematically represented
occurring repertoire type in Europe, would for instance evolve from diglossia and subsequently become subject to dialect loss. As such, the model also accounts for the diachronic dimension which marks community repertoires (but see Rutten/Vosters/Vandenbussche 2014: 14 for a critical account).

In describing diaglossic repertoires, Auer (2011: 491) rightfully emphasizes that careful empirical study is necessary to determine how this repertoire type is internally structured: “Terms such as regional dialect or regiolect suggest that the intermediate layer of speech between dialect and standard has the status of a variety. This, however is not necessarily the case, and careful empirical studies are necessary to establish variety boundaries.”

Whether or not varieties and variety boundaries can be distinguished in dialect/standard constellations strongly depends on the theoretical interpretation of the term variety (see Ghyselen/De Vogelaer 2018 for an in-depth discussion), but most linguists recognize covariance as an important criterion, conceptualizing varieties as sets of variants strongly correlating in their socio-situative behaviour (cf. Weinreich/Labov/Herzog 1968, Berruto 2010, Geeraerts 2010). In attempts to map the nature and structure of language repertoires, it is hence essential to gain insight into covariance patterns. This paper will tackle the question how covariance structures in language repertoires (and hence also the nature of language repertoires) can be empirically studied, building on data from the city of Ghent (capital of the province East Flanders in Belgium). We more specifically address the questions (1) whether the spoken language repertoire in Ghent should be described as di- or rather as diaglossic and (2) whether varieties can be distinguished within this repertoire.

3. Flanders as a case-study

This paper analyses the speech of ten Flemish speakers, who were observed within the context of the first author’s PhD project (cf. Ghyselen 2016b). The goal of that project was to map the internal structure of language repertoires in the cities of Antwerp, Ghent and Ypres (situated in respectively the Brabantine, East Flemish and West Flemish dialect area, cf. Fig. 2), and to study the diachronic dynamics marking these repertoires. In this paper, we zoom in on the data from Ghent.

The Flemish language repertoire is generally described as diaglossic (see for example Rys/Taeldeman 2007, Auer 2005, Swanenberg/Van Hout 2013): between dialect and standard language a whole continuum of intermediate variations is found, which have been dubbed tuusentaal (‘in-between language’, Taeldeman 1992), Soapvlaams (‘Soap Flemish’, Geeraerts 1999) and Verkavelingsvlaams (‘allotment Flemish’, Van Istendael 1989). Repertoire structure is in Flanders however also regionally de-

2 The term “Flanders” is used in its political sense to refer to the northern, Dutch speaking part of Belgium.
3 These intermediate variations are to varying degrees (depending on speaker/situation/…) regionally coloured.
termined: while the repertoires in the Limburgian, Brabantine and East Flemish dialect area are said to evolve from diaglossia to a situation of dialect loss (Vandekerckhove 2009, Ghyselen/Van Keymeulen 2014), the West-Flemish repertoire was until recently described as largely diglossic, with speakers switching between dialect and some form of standard (Willemyns/Vandenbussche 2008, De Caluwe 2009). This exceptional position of West Flanders can be linked to dialect functionality: due to a fairly limited migration from outside the area, the local dialects in West Flanders have remained exceptionally vital for a long time, which seems to have curbed the development toward a diaglossic repertoire (cf. Ghyselen/De Vogelaer 2013). Recent research has however demonstrated that intermediate language use is nowadays also omnipresent in West Flanders (Gabel 2010, Ghyselen 2016a). The West Flemish repertoire hence also seems to evolve toward a diaglossic structure.

Whereas most usually, the continuous nature of Flemish diaglossia is emphasized – varieties such as standard language, tussentaal, and dialect would not be discernable (Ammon 1977: 80, Willemyns 1982: 80, De Caluwe 2009, Plevoets 2008) – more recently, there are also studies mentioning “stable” dialect features in Flemish tussentaal (for example Rys/Taeldeman 2007, Taeldeman 2008, De Decker/Vandekerckhove 2012), which would occur in the tussentaal of all Flemings and as such constitute the heart of a homogeneizing tussentaal variety. Empirical research into this phenomenon and repertoire structure more generally is however rare (but see Plevoets 2008 for an exception).

An empirical study of repertoire structure ideally departs from corpora comprising the spoken and/or written language use of a wide range of language users in a wide range of situations. For this reason, a new corpus was compiled between 2012 and 2014 comprising the speech of 30 highly educated women from Ypres, Ghent and Antwerp (cf. Ghyselen 2015, 2016a). The studied women were born between 1981 and 1986 (n = 15) or between 1955 and 1961 (n = 15), and were recorded in five speech settings: (1) a dialect

---

4 Speakers of the younger age group have the letter “a” in their speaker code (for example a1), whereas the older speakers have the letter “b” (for example b2). In the cluster and MDS-plots, the prefix “Ovla” stands
test ("dia" in the plots and tables of this paper), (2) a standard language test ("st"), (3) a conversation with a friend from the same city ("reg"), (4) a conversation with a friend from a different dialect area ("sup") and (5) a sociolinguistic interview with an unknown interviewer from a different dialect area ("int"). During the sociolinguistic interviews, data were gathered about the linguistic background of the informants and their perceptions of their own language use and language in Flanders in general. In the dialect and standard tests, the informants heard stimuli sentences spoken in either standard Dutch or in the local dialect, which they had to translate into respectively the dialect of the oldest people of their town and standard Dutch "as heard during news broadcasts". These tests were used to determine the informant’s proficiency in the most acrolectal vs. basilectal speech styles available in a relevant location. The gathered recordings were transcribed orthographically using the Praat software (Boersma/Weenink 2011); with the software package EXMARaLDA (Schmidt/Wörner 2009) a searchable corpus was built.

To analyse covariance patterns, it is key to study the use of the language variants in correlation with the studied situations and the different speakers. In a diglossic scenario, we expect to find two clusters of language variants, one type with features showing strong correlations with informal settings and another with more formal ones (cf. Berruto 2010). In a diaglossic scenario by contrast, we expect either more than two clusters or a continuum from on the one hand features which are exclusively used in informal settings to on the other hand features which are heard in formal settings mainly, and in between features which are heard in both and which may be used differently by different speakers.

In this paper, we analyse the East Flemish data quantitatively adopting a correlative sociolinguistic approach. More specifically, the distributions of 25 phonological and morphosyntactic features are studied in the five types of speech settings. Table 1 offers an overview of the selected linguistic variables and the variants attested in Ghent. These variables were selected from an inventory of potentially interesting variation phenomena drawn up on the basis of earlier variationist research in Flanders (Barbiers et al. 2005, Barbiers et al. 2008, Goossens/TaeldeMan/Verleyen 1998, Goossens/TaeldeMan/Verleyen 2000, De Wulf/Goossens/TaeldeMan 2005, De Schutter et al. 2005, Goeman et al. 2008, Rys/TaeldeMan 2007, TaeldeMan 2008). In selecting the variables, both frequency and regional spread were important criteria. Firstly, low frequent variables were avoided as these would cause problems for the statistical analyses. Secondly, the features were selected in such way that there was variation in the regional spread of the studied dialectal variants. Whereas t-deletion in

---

5 Gender was not controlled for in these conversations, as it was already difficult finding suitable informants without making demands on the gender of the speech partners. 6 of the 20 conversations with friends (of the same or of a different dialect area) were mixed sex conversations; the majority were same sex conversations.

6 Of each conversation with a friend 30 minutes were transcribed; the interviews and dialect and standard tests were transcribed entirely.
the word *niet* for instance occurs in almost all Flemish dialect, the realisation of Standard Dutch [ɛ.i] as [ɑː] is characteristic of the dialects in the Ghent region only. Regional spread is known to have an important impact on the diachronic and stylistic dynamics of a dialect feature (Schirmunski 1928/1929, Taeldeman 2009), hence it was important to factor this criterion in when selecting variables. It is beyond the scope of this paper to discuss all studied variables in detail; we refer to Ghyselen (2016b) for an in-depth description.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Attested variants (&quot;code&quot;) + type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonetic/phonological variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Representation Standard Dutch [ɛ.i]</strong> (not before r or in auslautposition) (n = 2595)</td>
<td>&gt; [ɑː] (&quot;maan&quot;)*</td>
<td>[mɑː.n] (‘mine’) &gt; long monophthong or diphthong (&quot;mijn&quot;)**</td>
</tr>
<tr>
<td><strong>Representation Standard Dutch [œ.y]</strong> (&gt; wgm. ù) (n = 670)</td>
<td>&gt; [ɑː] (&quot;haës&quot;)*</td>
<td>[hɑː.s] (‘house’) &gt; [y] (&quot;hus&quot;)***</td>
</tr>
<tr>
<td><strong>Representation Standard Dutch [ɔ.u]</strong> before [t] of [d] (n = 212)</td>
<td>&gt; [ɑː] or [œ.u] (&quot;kaëd&quot;)*</td>
<td>[ka:t] (‘cold’)</td>
</tr>
<tr>
<td><strong>Representation Standard Dutch [ɣ]</strong> (n = 5232)</td>
<td>&gt; laryngalization (&quot;h&quot;)****</td>
<td>[hjoed] [h]edaan (‘well done’)</td>
</tr>
<tr>
<td><strong>Preservation of non-suffixal final schwa</strong> (n = 249)</td>
<td>&gt; with schwa (&quot;bedde&quot;)*</td>
<td>bedde (‘bed’)</td>
</tr>
<tr>
<td><strong>Representation Standard Dutch [o:]</strong> (&gt; wgm. û in open syllables) (n = 241)</td>
<td>&gt; palatalized form (&quot;zeun&quot;)*</td>
<td>[zøn.a] (‘son’)</td>
</tr>
<tr>
<td><strong>Representation of Standard Dutch initial h in a selection of words</strong> (n = 1350)</td>
<td>&gt; h-procope (&quot;hdel&quot;)*</td>
<td>øevel (‘how much’)</td>
</tr>
</tbody>
</table>

7 These codes will be used in the graphs of this paper.
8 This categorisation is based on data from the “Syntactic Atlas of the Dutch Dialects” (Barbiers et al. 2005, Barbiers et al. 2008), the “Phonological Atlas of the Dutch Dialects” (Goossens et al. 1998, Goossens et al. 2000, De Wulf et al. 2005) and the “Morphological Atlas of the Dutch Dialects” (De Schutter et al. 2005, Goeman et al. 2008). For a number of variables, specialized dialectological descriptions were consulted (de Vogelaer 2008 on subject marking; Cornips/De Vogelaer 2009 on gender in Dutch; de Vogelaer/Vandenberghe 2006 on indefinite pronouns and adverbs). This a-priori classification of language variants does not necessarily correspond to laymen’s classifications of the variants; it merely offers an insight into how the variants are characterized in traditional dialect and standard language descriptions.
9 For this variable and also for the realization of Standard Dutch [œ.y] and [ɔ.u], no distinction was made between respectively the long monophthongs [e.i], [œ.i], [ɔ] on the one hand and the diphthongs [ɛ.i], [œ.y] and [ɔ.u] on the other. Without acoustic analyses, those distinctions proved too difficult to make objectively.
### Variable | Attested variants (“code”) + type | Example
---|---|---
**t-deletion in niet (‘not’) or in dat (‘that’) + consonant (n = 3,820)** |  &gt; **t-apocope** (‘niedaC’)*  
 &gt; realization final consonant (‘niedatC’)**  | je moet da nie doen. (‘You do not have to do that’)  
je moet dat niet doen.

**t-deletion in dat (‘that’) + vowel (n = 935)** |  &gt; **t-apocope** (‘daV’)**  
 &gt; realization final consonant (‘datV’)**  | da ook (‘that too’)  
dat ook

**Intervocalic d (n = 229)** |  &gt; glide (‘VwjV’)**  
 &gt; [d] (‘VdV’)**  | dooie vent (‘dead man’)  
dode vent

### Morphosyntactic variables

**Male singular indefinite article (n = 545)** |  &gt; ne (‘ne’)*  
 &gt; e (‘e’)**  
 &gt; een (‘een’)**  | ne vent (‘a guy’)  
e vent  
een

**Verb form present simple 1st singular thematic verbs (in sentences without inversion) (n = 567)** |  &gt; root + e (‘ikmake’)*  
 &gt; infinitive (‘ikmaken’)**  
 &gt; root (‘ikmaak’)**  | ik make (‘I play’)  
i k maken  
i k maak

**Verb form present simple 1st singular athematic verbs (n = 244)** |  &gt; infinitive (‘ikstaan’)**  
 &gt; root (‘iksta’)**  | ik staan (‘I stand’)  
 i k sta

**Personal pronoun he – weak form in postverbal position or after conjunctions (n = 286)** |  &gt; subject doubling (‘hij2dub’)*  
 &gt; em (‘hij2em’)**  
 &gt; ie (‘hijzie’)**  
 &gt; hij (‘hijz’)**  | Komtjij ook? (with jij being a contraction of ie and ij; ‘is he coming too?’)  
Komt em ook?  
Komtie ook?  
Komt hij ook?

**Subject doubling 3 singular masculine/feminine, 1 plural, 3 plural in sentences with inversion and dependent clauses, with a full subject (n = 437)** |  &gt; subject doubling (‘sub31’)*  
 &gt; no subject doubling (‘gnssub31’)**  | A me wider komen … (Lit: ‘if we we come’)  
Als wij komen … (Lit: ‘if we come’)

**Auxiliary in present perfect with zijn (‘to be’), tegenkomen (‘meet’) and vallen (‘fall’) as main verbs (n = 158)** |  &gt; hebben (‘hebben’)*  
 &gt; zijn (‘zijn’)**  | Ik heb ziek geweest. (Lit: ‘I have ill been’)  
Ik ben ziek geweest. (Lit: ‘I am ill been’)

**Subject doubling 2 singular/plural and 1 singular in sentences with inversion and dependent clauses, with a full subject (n = 516)** |  &gt; subject doubling (‘sub12’)*  
 &gt; no subject doubling (‘gnssub12’)**  | Morgen kom ek ik ook. (Lit: ‘tomorrow come I I too’).  
Morgen kom ik ook. (Lit: ‘tomorrow come I too’).

**Preposition in subclauses with to-infinitives (n = 184)** |  &gt; preposition voor (‘voor’)*  
 &gt; preposition om (‘om’)**  | Dat kost veel voor te wassen. (Lit: ‘that costs much for wash’)  
Dat kost veel om te wassen. (Lit: ‘that costs much to wash’)

---

10 No distinction is made between the variants da[t] and da[d], since that distinction is often difficult to make without acoustic analyses.

11 In these cases, subject doubling with a weak pronoun is obliged in the local dialect (cf. De Vogelaer 2008: 326).
<table>
<thead>
<tr>
<th>Variable</th>
<th>Attested variants (&quot;code&quot;) + type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expletive dat</strong> (‘that’) after the conjunctions wie, wat, waar, hoe, wanneer en of (n = 353)</td>
<td>&gt; with expletive dat (“exdat”)*</td>
<td><em>Ik weet niet wie dat er komt.</em> (Lit: ‘I know not who that is coming’)</td>
</tr>
<tr>
<td></td>
<td>&gt; without expletive dat (“gnexdat”)**</td>
<td><em>Ik weet niet wie er komt.</em> (Lit: ‘I know not who is coming’)</td>
</tr>
<tr>
<td><strong>Personal pronoun</strong> second singular, weak form in preverbal position (n = 366)</td>
<td>&gt; ge (“get”)*</td>
<td>*ge speelt (‘you are playing’)</td>
</tr>
<tr>
<td></td>
<td>&gt; je (“jel”)**</td>
<td><em>je speelt</em></td>
</tr>
<tr>
<td></td>
<td>&gt; je ... gie/ge ... gie (“je1dub”)*</td>
<td><em>je speelt gie</em></td>
</tr>
<tr>
<td><strong>Personal pronoun</strong> second singular, weak form in postverbal position (n = 483)</td>
<td>&gt; (d)e (“(d)e2”)*</td>
<td>*speel de? (‘are you playing?’)</td>
</tr>
<tr>
<td></td>
<td>&gt; de gij (“je2dub”)*</td>
<td><em>speel de gij?</em></td>
</tr>
<tr>
<td></td>
<td>&gt; je (“je2”)**</td>
<td><em>speel je?</em></td>
</tr>
<tr>
<td></td>
<td>&gt; ge (“ge2”***</td>
<td><em>speelt ge?</em></td>
</tr>
<tr>
<td><strong>Diminutives</strong> with nouns of which the root does not end in [t] (n = 149)</td>
<td>&gt; ke-diminutive (“kedim”)*</td>
<td>*bloemke/bloemeke (‘little flower’)</td>
</tr>
<tr>
<td></td>
<td>&gt; je-diminutive (“jedim”)**</td>
<td><em>bloempje</em></td>
</tr>
<tr>
<td><strong>Negation</strong> in sentences with nooit (‘never’), niemand (‘no one’), nergens (‘nowhere’) (n = 104)</td>
<td>&gt; double negation (“dneg”)*</td>
<td>*Ik ga dat nooit nie doen. (Lit: ‘I go that never not do’)</td>
</tr>
<tr>
<td></td>
<td>&gt; single negation (“eneg”)**</td>
<td>*Ik ga dat nooit doen. (Lit: ‘I go that never do’)</td>
</tr>
<tr>
<td><strong>Possessive pronoun</strong> 1 plural – inflection before female singular nouns, male singular nouns referring to a family relationship, or before plural nouns (n = 20)</td>
<td>&gt; no inflection (“onsz”)*</td>
<td>*ons moeder (‘our mother’)</td>
</tr>
<tr>
<td></td>
<td>&gt; with inflection (“onzez”)**</td>
<td><em>onze moeder</em></td>
</tr>
</tbody>
</table>

Table 1: Overview of studied variables and variants (*= typical of traditional dialect in Ghent, **= typical of Standard Dutch, ***= variant does not occur in traditional dialect nor in the standard, ****= variant occurs in both the traditional dialect and the standard)

To study how the attested variants correlate to each other and to the independent variables age, speech setting and speaker, we will look at the results from three analysis techniques: correspondence regression (section 4), cluster analysis (section 5) and non-metric multidimensional scaling (section 6). These techniques are combined, given that the first and latter technique reduce a part of the original variance and it is hence interesting to compare the results of different multivariate techniques. Results emerging in multiple analysis types can be considered more reliable than results occurring in only one analysis.

12 The allomorphy within the *je*-suffix was not taken into account as this complicates the calculation of distance measures: some of the *je*-suffixes in the Ieper dialect coincide for instance with Standard Dutch *je*-suffixes (*bloemetje, ‘little flower’), whereas others have another allomorph (*boeksje versus boekje, ‘little book’).
4. Correspondence regression

4.1 About the technique

Correspondence regression (also known as profile based correspondence analysis, cf. de Sutter/Delaere/Plevoets 2012) is a descriptive data analysis technique which studies correspondences or associations between rows and columns of a data matrix and tries to visualize these comprehensibly. In the case of the here discussed data set, correspondence regression departs from a cross table in which the attested variants are listed in the rows, and the combinations of speaker and situation or age and situation in the columns (cf. Table 2). The columns in each row of Table 2 can be considered dimensions of an \( n \)-dimensional space, in which every language variant occupies a specific position (cf. Fig. 3) and as such, a data cloud of points emerges (Plevoets 2008: 65). The row values in each column can in their turn be considered dimensions in a \( k \)-dimensional space, in which every combination of speaker and situation occupies a specific position.

As a first step in the correspondence regression, two distance matrices are calculated, one for the distances between the rows of a table on the one hand (for example between “maan” and “ne”) and one for the distances between the columns of the table on the other hand (for example between “a1-dia” and “a1-reg”). Those distances – calculated using the chi-square measure, designed for frequency data – yield information on the associations between the rows, and the associations between the columns. In the case of Table 2, we for instance calculate all distances between the \( k \) variants in the \( n \)-dimensional space and all distances between the \( n \) columns in the \( k \)-dimensional space.

A second step in correspondence regression is to map the distances calculated in the \( k \)- or \( n \)-dimensional spaces as accurately as possible in a lowerdimensional plot (usually two- or threedimensional). To this end, the original \( k \)- or \( n \)-dimensional matrices are reduced to lowerdimensional spaces. This is done using “singular value decomposition”, a technique which reduces the dimensionality of the data while trying to retain as much relevant information as possible (cf. Strang 2009). It is up to the analyst to decide how many dimensions should be retained in the lowdimensional representation of the

<table>
<thead>
<tr>
<th></th>
<th>a1-dia</th>
<th>a1-reg</th>
<th>a1-sup</th>
<th>a1-int</th>
<th>a1-st</th>
<th>a2-dia</th>
<th>a2-reg</th>
<th>( \ldots )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>“maan”</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>“mijn”</td>
<td>0</td>
<td>4</td>
<td>50</td>
<td>( 58 )</td>
<td>6</td>
<td>0</td>
<td>8</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>“ne”</td>
<td>3</td>
<td>27</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>“e”</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>“een”</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( k )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

Table 2: Example of a cross table which can serve as input for correspondence regression (with absolute frequencies in the cells)
distance matrices. The challenge is to find an equilibrium between surveyability on the one hand (five-dimensional spaces are for instance not easy to interpret) and representativeness on the other hand (reducing the original variance too much can lead to wrong or oversimplified conclusions). In a scree plot or singular value diagram, which visualizes the variance explained by each dimension in decreasing order, the analyst can look for an “elbow”, a point at which a further increase in the data dimensionality implies only minor increases in the explained variance. Once decided how many dimensions one wants to retain, the distances in the two reduced matrices are visualized together in biplots, in which the Euclidean distance between two datapoints is an indication of the \( \chi^2 \)-association between those two points. The axes in the biplots do not have predetermined interpretations; important are especially the distances between the data points and the ways in which they cluster (cf. Geeraerts 2010 and Plevoets this issue).

Correspondence regression differs from “classic” correspondence analysis in that it is “profile-based” (hence also the name profile-based correspondence analysis in De Sutter et al. 2012). Contrary to classical correspondence analysis, correspondence regression does not treat language variants as autonomous data points, but rather as sublevels of an overarching variable. In the case of our data, the masculine indefinite articles \( ne \), \( e \) and \( een \) are for instance treated as sublevels of the variable \( een \), and not as three autonomous variables. Mathematically, correspondence regression differs from the classical variant in the way in which distance matrices are calculated. In a non-profile-based perspective, \( \chi^2 \)-distances are calculated by first comparing the absolute frequency of a variant in a specific context to the row total, i.e. the absolute frequency of that variant in the corpus. Subsequently, the different row proportions are compared. In a profile-based perspective by contrast, the absolute frequency of a variant in a specific context is first weighed against the absolute frequency of the variable in that context. This approach has

Fig. 3: Example of a multidimensional representation of a cross table (based on Table 2)
as big advantage that the obtained distances are “not influenced by any disproportions among the overarching variables, which may arise in the corpus compilation” (De Sutter et al. 2012: 331). If a certain linguistic variable for instance for some reason occurs less in one speech setting than in others, it has important consequences on the row proportions in a non-profile-based perspective. In Table 2 for instance, only considering the row proportions for the article ne would lead to the conclusion that speaker a1 uses that article as much in the regional informal conversations as in the supraregional informal conversations. That conclusion is however not accurate, given that many equivalent article variants are used in the supraregional informal conversations (e and een), while that is not the case in the regional informal conversation. In a profile-based perspective, the absolute frequencies of a variant are therefore first weighed against the variable frequency – in the case of the article ne for speaker a1 in the regional informal conversations, the absolute frequency (27) is for instance divided by the variable frequency (27+1+1) – to only subsequently focus on row proportions. For more mathematical details on this profile-based approach, see Speelman/Grondelaers/Geeraerts (2003), Plevoets (2008), and De Sutter/Delaere/Plevoets (2012).

Another aspect in which the used correspondence regression differs from more “traditional” versions of correspondence analysis is that a hypothesis testing dimension is added; the technique is hence not purely descriptive. More specifically, confidence ellipses are drawn on the basis of “bootstrapping” (Plevoets 2015). If the ellipses of two categories (for example two age groups) do not overlap, this can be considered an indication that the distance between those two is significant; overlap points to a lack of statistical significance (cf. Plevoets 2013).

4.2 Results

Figure 4 shows the scree plot of the performed regression analysis, indicating an elbow at dimension 2, which means that a large part of relevant information can be captured in two dimensions (54.24% of the original variance). Considering the confidence intervals, debate is possible on the usefulness of also analysing a third dimension, as there is only a slight overlap between the confidence intervals of the second and third dimen-

![Fig. 4: Scree plot correspondence regression Ghent (% of explained variance on y-axis)](image-url)
Mapping the Structure of Language Repertoires: on the Use of Sociolectometric Methods

Indicating that this third dimension might also have some explanatory power. A visualisation of the third dimension is enclosed in appendix, but as it proved difficult to identify or explain any patterns in this dimension, the focus in what follows will be on dimensions 1 and 2. These two dimensions are visualized in Figure 5, a biplot of all language variants attested in Ghent (grey), with the main effects for situation (black). Age and speaker were also added as variables in the correspondence regression, but these effects are not plotted in Figure 5 as they will be discussed further on in this paper.

In interpreting the biplot in Figure 5, it is important to focus on the distances between the different data points. Variants plotted close to each other, behave alike (considering the ten speakers in five situations); large distances indicate weak correlations. In the top right corner of Figure 5, we for instance see a relatively small distance and hence strong correlation between the realisation of final-t in the function words niet and dat on the one hand ("nietdatC") and the preverbal pronoun form je ("je1") on the other hand. A large distance can by contrast be found between "nietdatC" and "bedde", the preservation of non-suffixal final schwa, which indicates that those two variants behave
very differently. A speaker realizing the non-suffigal final schwa is in other words very unlikely to also realize her final t’s in niet and dat.

As was already highlighted above, the axes in a biplot do not have predetermined interpretations. Meaning can however be allocated to the dimensions by looking for patterns in the data points (Geeraerts 2010: 244). Figure 5 for instance shows a clear horizontal continuum stretching from dialect variants on the left side (see “bedde” or “zeun”, the palatalisation of wgm. Ġ in open syllables) to – on the right side – standard language variants (for example “nietdatC”) and variants which are not endogenous in the Ghent dialect (according to existing dialect descriptions, cf. Barbiers et al. 2005, Barbiers et al. 2008, Goossens et al. 1998, Goossens et al. 2000, De Wulf et al. 2005, De Schutter et al. 2005, Goeman et al. 2008, Rys/Taeldeman 2007, Taeldeman 2008), such as the West-Flemish realisation of wgm. Ġ as [y] (“hus”). The x-axis hence seems related to “localness” (left: local, right: non-local/also in standard). The y-axis is more difficult to interpret. In the top right corner, features can be found which are usually associated with formal standard language, such as the realisation of final t’s (“nietdatC”), the personal pronoun je in second person singular (“je”), and the realisation of initial h (“gnhdel”). In the bottom of the plot, a number of features can be found which are not endogenous in the Ghent dialect (again: according to existing dialect descriptions), such as the West-Flemish realisation of wgm. Ġ as [y] (“hus”), the Brabantically flavoured realisation of wgm. au as [u] (“groot”) and the West-Flemish n-suffix in the first person singular of thematic verbs in the present (“ikmaken”). In the centre of the biplot a cluster of both dialect variants (such as ke-diminutes “kedim” and the personal pronoun ge) and standard language features can be found (such as the realisation of intervocalic d “VdV” or the realisation of wgm. au as [oː] “groot”), a cluster which seems to correspond to what is called “tussentaal” in the literature.13

Studying the associations between language variants on the one hand and situation on the other, a strong association (small distance) can be seen between the central cluster in the biplot and the regional and supraregional informal conversations. These two speech settings overlap in the biplot, indicating that the language use in these settings barely differs (at least when considering the studied features). The regional background of the speech partner hence seems to play only a minimal role for the studied speakers in Ghent. Moreover, the informal speech situations show only weak associations with the dialect features in the left of the biplot, hinting at a weak position of the traditional dialect in the language repertoire of highly educated women in Ghent. The dialect test (cf. Fig. 5) shows somewhat stronger, but still not overwhelming associations with the dialect features in the left of the biplot, which indicates that even when speakers are instructed to speak the traditional dialect, they do not all realize the features typically associated with that dialect. This observation can be considered a sign of weak dialect competence. When we compare the language use in the interview setting with the speech in the regional and supraregional conversations, a stronger association with the

13 Because of the strong associations (and hence the small distances), the central labels are difficult to read. For this reason, Figure 6 (see appendix) zooms in on the centrum of the biplot (coordinates [-1; 0,5]), with more readable labels as result.
standard language features in the top right of Figure 5 can be observed for the interview setting. The rather formal interview setting – in which the speech partner was unknown and spoke standard language – seems to have triggered “more standard” language use. The strongest association with the standard language features is however unsurprisingly found in the standard language test (cf. Fig. 5), indicating that the informants did not fully exploit their standard language competence in the interview setting. The gap between the interview setting and the standard language test is probably related to the fact that Flemish speakers are known to feel uncomfortable about the VRT-Dutch norm (cf. Geeraerts 2001 on the “sunday suit mentality”) and as such in the spontaneous interview, in which no target variety was imposed, depart from a somewhat looser norm. Also at play is probably that speakers had to focus on form and content during the interview, whereas in the standard language test, all attention could be devoted to the language used.

Overall, when considering the language use of all speakers together, the repertoire in Ghent qualifies for the label diaglossic, with a continuum stretching from dialect variants on the left to standard variants on the right. The question however arises whether this continuum also implies diaglossic individual repertoires among the studied women. The overall diaglossic image might after all also result from a combination of mainly diglossic or even monoglossic personal repertoires, each comprising slightly different language codes (cf. Ghyselen 2016a). In this respect, it is interesting to study the situation effects for each speaker individually. Figures 7 and 8 (see appendix) display the interaction effects between speaker and situation for the younger and older speakers respectively. In these plots, confidence ellipses are not visualized for reasons of surveyability; Table 3 however indicates for each speaker which situation effects displayed overlapping confidence ellipses. The second column for instance lists the situation effects with which the effect for the dialect test overlapped (per speaker).

In Figures 7 and 8 (see appendix), clear interspeaker variation can be observed when it comes to the dialect pole (or the absence of one) of the personal continua. Whereas almost all speakers display strong associations with the standard language features in the top right corner of the biplot in the standard language test, there is considerable idiolectal variation in the dialect test: speakers b2, b3 and b4 show strong associations with the dialect variants in the left side of the biplot, speakers a2, b1 and b5 moderate associations and speakers a1, a3, a4 and a5 weak associations. Speaker a4 moreover displays relatively strong associations with the non-standard, dialect exogenous features in the bottom of the graph. In an attempt towards the Ghent dialect, this speaker realizes West Flemish dialect variants – such as [y] for wgm. û (“hus”) and the en-suffix in the present of athematic verbs in the first person singular (“ikmaken”) – which indicates a lack of proficiency in the Ghent dialect, similar to the other speakers displaying weak associations with the Ghent dialect variants in the dialect test. Interpreting the data in

---

14 For practical reasons, coloured figures are added in the appendix to this paper.
the apparent-time, the stronger associations with the dialect test among older speakers point towards declining dialect knowledge and hence also dialect loss in Ghent.

When comparing the dialect tests on the one hand with the regional and supraregional conversations on the other hand, idiolectal variation is clearly more limited in the latter speech settings. The overall situation effects already indicated that the regional origin of a speech partner seems to have only a minimal impact on the language use in Ghent. The interaction plots demonstrate that this is the case for all studied speakers: the situation effects for the two informal conversations are for each speaker in fairly close proximity (cf. also the many overlapping confidence ellipses summarized in Table 3). Most speakers display strong associations with the central cluster in the biplot, which consists of a combination of dialect features and standard language features. Speaker b5 constitutes an exception however; her language use in the informal conversations associates more strongly with the dialect variants left in the biplot. When comparing the language use in those informal conversations with the language spoken in the interview setting, it is striking that the distances are not overwhelming (see Ghyselen 2016a on West-Flanders for a contrasting result). For speakers a3, a4, b1 and b4, the confidence ellipses for the interview effects overlap with those of other spontaneous speech settings; speakers a1 and b5 however switch more clearly (larger and significant distance) to more standard language usage in the interview.

The available data do not allow straightforward interpretations as to whether the individual repertoires are diaglossic or not. When neglecting the fairly artificial language use in the dialect and standard language test, some speakers seem to have largely monoglossic language repertoires (a3 and a4), with as main code a type of language deviating from both VRT-Dutch and the Ghent dialect. Other speakers (a1) however switch between this intermediate type of language and a more standard type of language; b5 was observed to realize more dialectal features when speaking to friends. It has to be kept
in mind however that the analyses build on data from only a limited number of speech settings. It is fairly probable that there are other settings – not studied here – in which the speakers from Ghent do realize the language observed in the dialect and standard test spontaneously, and that the personal repertoires are hence “broader” than the discussed data suggest. In this respect, the interview metadata offer an interesting extra perspective. Whereas the older speakers b2, b3 and b4 indicate that they do sometimes speak more dialectal – in situations not recorded in the study (for example talking to older family members) – several of the younger speakers report, in line with the studied production data, not mastering or speaking the local dialect. We can conclude that the overall repertoire structure in Ghent is diaglossic, but that that is not uncontroversially the case for individual highly educated speakers.

5. Cluster analysis

5.1 About the technique

In her study on repertoire structure in Wittlich, Ленц (2003) applies cluster analysis to explore the covariance patterns in her data. Hierarchical cluster analysis is closely related to correspondence regression, as it is a multivariate technique which aims at identifying clusters in multivariate data in such a way that “the members of one group are very similar to each other and at the same time very dissimilar to members of other groups” (Гриц 2009: 337). The output of a hierarchical cluster analysis is a “dendrogram” (cf. Fig. 9), in which “objects” – for example speakers or language variants – are clustered on the basis of linguistic distances.

Correspondence and cluster analysis are both descriptive multivariate techniques – multiple dependent variables (in our case all language variants) are studied simultaneously – which group variables on the basis of mutual correspondences. Mathematically, however, the techniques differ significantly. In correspondence regression, features are projected on a low-dimensional space after dimension reduction, whereas in hierarchical cluster analysis features are grouped step by step on the basis of distances. To illustrate the methodology, we depart from Table 2 again. The first step in cluster analysis is comparable to correspondence regression: a distance matrix is calculated for the distances between the $k$ rows of our dataset. In contrast to the correspondence regression approach, only one distance matrix is calculated in cluster analysis (a matrix with distances between the rows, not for distances between columns), and other distance measures than chi-square distances can be opted for (cf. Cox/Cox 2008: 317–319, Гриц 2013: 336–349).

In a second step of the hierarchical cluster process, pairs of objects are selected which show the strongest association, i.e. features with the smallest distance in the distance matrix. These objects are then combined in a cluster. In the example of the dendrogram in Figure 9 objects I and D – at a distance of approximately 0.1 – were combined in a first phase. The next step in the cluster process consists of combining this small cluster
with other objects resulting in a larger cluster. To compute the distance between clusters, different methods have been developed. In this study, we use the “Ward method”, also known as the “minimum variance method”, which generates clusters in such way that the variance within each cluster is as small as possible, and the distance between clusters as large as possible (Gries 2009: 317, Janssens et al. 2008). On the basis of this linkage method clusters are conjoined step by step until in the top of the dendrogram one large cluster of all objects is obtained. It is then up to the researcher to interpret the dendrogram.

Cluster analysis has frequently been criticized for the instability of the results it obtains: because cluster analysis departs from minimal elements, small changes in the input data can lead to very diverging results (Nerbonne et al. 2008: 648). For this reason, new variants of cluster analysis have been developed which allow assessing the stability of the cluster results. In this paper, we will use “bootstrap clustering” to circumvent the stability problem (Nerbonne et al. 2008, Suzuki/Shimodaira 2006). This type of cluster analysis first generates a large number of subsamples of the data (in our study 5,000) by combining random observations from the original dataset. Subsequently, a dendrogram is created for each of these subsamples, which results in a large number (in our case 5,000) of dendrograms. These dendrograms are then compared; clusters which occur in many versions (and hence have a high “bootstrap probability value”) are more reliable than clusters which occur in a few dendrograms only. To calculate these bootstrap probabilities, we use the R-package pvclust (Suzuki/Shimodaira 2013), which generates dendrograms which mention both the bootstrap probability (BP) and the approximately unbiased probability (AU). The BP-value is obtained by means of traditional bootstrapping, in which no demands are imposed on the size of the subsamples, whereas the AU-value is calculated after “multiscale bootstrapping”, a procedure during which the size of the subsamples is consciously altered. Given that the AU-values are more reliable according to Suzuki/Shimodaira (2006), we will rely on these values when evaluating the cluster solutions. AU-values of 95% or larger – which correspond to a p-value of 0.05 or smaller – point towards a strong reliability of the distinguished clusters. These AU-values will be important to determine the total number of relevant
clusters, which is one of the main challenges in interpreting a dendrogram (cf. Everitt 1972).

5.2 Results

On the basis of a frequency table with the relative frequencies per variant for each situation, we built a distance matrix with Euclidean distances, which then served as input for the clustering algorithm. Figure 10 (see appendix) shows the resultant dendrogram. For a division of this dendrogram into clusters, a cut-off point of 2 was used, as this cut-off point yields sufficiently large clusters (marked in different colours for ease of interpretation) which are convincingly supported by the data; the AU-values of most clusters equals or exceeds 95. Only the last cluster on the right (marked in purple) is not convincingly supported by the data (AU-value of 10); we should hence bear in mind that variants such as “datV”, “g” and “VdV” might display situational behaviour differing from the other variants in the cluster.

When studying the variants in the different clusters, the yellow cluster seems to be a dialectal one: all variants in this cluster are either known to occur in the Ghent dialect (“gruut”, “voor”, “zeun”, “bedde”, …) or can be considered hyperdialectals, i.e. approximations of the dialect deviating from both source and target varieties (cf. Lenz 2004: 291). As was already discussed in section 4.2., variants such as the West-Flemish realisation of wgm. û as [y] (“hus”), the Brabantically flavoured realisation of wgm. au as [u] (“groet”) and the West-Flemish n-suffix in the first person singular of thematic verbs in the present (“ikmaken”) can be considered the result of a limited dialect competence.

The grey and red clusters on the other hand can be labelled standard Dutch: except for the postverbal pronoun ge (“ge2”), all variants in these clusters can be considered standard. It is interesting that also in this analysis a distinction is made between (1) features which are usually associated with formal standard language, such as the realisation of final t’s (“niedatC”), the personal pronoun je in second person singular (“je1”), and the realisation of initial h (“gnhdel”) and (2) other standard language features (red cluster), such as as the standard conjugation of verbs in the first person singular (“ikmaak”), which are more closely related to the green cluster. The latter cluster contains only dialectal features, such as the preverbal pronoun ge (“ge1”) and the deletion of final -t in the function words niet and that (“niedatC”), features which are traditionally described as

15 For each variant we divided the token frequency by the number of times the variant could occur in the situation.
16 One informant literally remarked in the middle of the dialect test that her knowledge of the dialect of Ghent was so poor that she even started mingling it with West Flemish variants.
17 Interestingly, the pronoun ge in postverbal position can be considered a tussentaal feature pur sang, as it is not typical of any dialect in Flanders, nor of standard Dutch (cf. De Decker 2014: 107). Van Hoof (2013: 540) considers this postverbal ge as an instance of “colourless usage”, “a less marked, less dialectal and hence more “neutral” sounding realisation of the second person in inversion than the dialectal counterparts” [own translation, ASG, DS & KP].
occurring frequently in *tussentaal* also (cf. Rys/Taeldeman 2007, Taeldeman 2008). The purple cluster, to conclude, contains a mixture of (1) standard Dutch features ("jedim" and "gnsb12"), (2) features which occur both in standard Dutch and the Ghent dialect ("datV", "g", "VdV" and "onze2") and (3) features which do not occur in standard Dutch, nor in the Ghent dialect ("h", "VwjV" and "daV").

A next logical question is in what sense the clustered features behave alike. Contrary to correspondence regression, cluster analysis only visualizes the distances between the rows of the original data matrix; it does not visualize the association between the clusters on the one hand and the studied situations on the other hand, and as such does not explain why certain variants are clustered together. For this purpose, the researcher is obliged to go back to the original frequency table and look for patterns. In this table, it for instance becomes clear that the features in the grey cluster are generally low frequent (relative frequencies of < 27%) except in the standard language test (relative frequencies exceeding 70%). A low overall frequency is also typical of the variants in the yellow cluster (which explains why the yellow and grey clusters are fairly close in the dendrogram), but the yellow variants are found most frequently (if they are realized at all) in the dialect test. The features in the red cluster by contrast are all generally high frequent, a trend which also goes for the variants in the green cluster, except that the latter features are almost always suppressed in the standard language test and to a larger degree also in the interview setting.

The typology of language variants built on the basis of the cluster analysis indicates – in line with the correspondence regression results – that a diaglossic model is most suited to describe the community repertoire of highly educated women in Ghent. After all, the cluster analysis reveals multiple clusters differing in stylistic behaviour, among which also a cluster of dialect features (marked in green) correlating with other dialect features in some speech settings and with standard Dutch features in other settings. The demarcation of this cluster in the dendrogram should however not be automatically interpreted as evidence of a separate variety (cf. discussion on “stabilization” of *tussentaal* mentioned in section 3); cluster analysis always yields clusters, whether or not there are gradual transitions from one cluster to another or not (see Ghyseelen/De Vogelaer 2018 for an in-depth discussion).

To compare the variant clustering in the hierarchical cluster analysis and the correspondence regression in more detail, Figure 11 (see appendix) shows the correspondence biplot discussed in section 4 with colours marking the cluster analysis classification. When ignoring the yellow cluster, the cluster analysis results align fairly well with the patterns emerging in the correspondence biplot. The standard language variants combined in the red cluster in Figure 10 for instance display strong correlations in the correspondence biplot too (see strong overlap in midpoint of graph). The grey standard Dutch features – which were described as standard variants occurring in the standard language test only – occur together in the top right of the correspondence biplot, displaying strong associations with the standard language test. The purple features form a continuous transition from the red cluster towards the grey one and are only weakly correlated, which aligns with the cluster analysis observation above that there
Fig. 12: Dendrogram of the speech setting effects per speaker on the basis of the relative frequencies per variant for each informant in the five studied situations (rectangles indicate clusters highly supported by the data, AU-value ≥ 95).
is not much evidence to consider these features as a separate cluster. Near the central red cluster, the features in green show a strong association with the informal speech settings and a weaker one with the interview setting. This green cluster fades smoothly into the red cluster (cf. also transition from red to purple and grey features), indicating that the clusters in Figure 10 should not necessarily be interpreted as strictly separated varieties. The features combined in the yellow cluster in the dendrogram warrant some more attention as they occur scattered in the correspondence biplot, with one group of features (“bedde”, “kaëd”, “zeun”, “maan”, etc.) clearly separated from other features (“e”, “hijzem”, “mën”, “zuun”). These groups can to some degree also be distinguished in the dendrogram (cf. “kaëd” en “bedde” in one subcluster of the yellow branch and “hus” and “dneg” in another subcluster), but the distinction in the dendrogram is not as clear as in the biplot. The clearer distinction between different types of yellow features in the biplot should probably be attributed to the fact that age was also included as parameter in the correspondence regression; when studying the individual speaker effects and the frequency tables, it becomes clear that the yellow features in the left of the biplot are mainly associated with the older speakers, whereas the other yellow features are used by both.

<table>
<thead>
<tr>
<th></th>
<th>Dia</th>
<th>Reg</th>
<th>sup</th>
<th>Int</th>
<th>St</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>a2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>a5</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b1</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b2</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b3</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b4</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Schematic representation of the interactions between speaker and situation based on the cluster dendrogram in Figure 12. Conversations marked in the same way are included in the same cluster in Figure 12. An asterisk indicates that the speech setting is included in a cluster that is not convincingly supported by the data (AU p-value < 95).

As was the case for the correspondence regression, the overall diaglossic image should not be held as an indication that each speaker shifts from dialect to standard on a continuum with intermediate variations. To study the variation among speakers, a cluster analysis was performed on the relative frequencies per variant for each combination of speaker and situation.\(^{18}\) Figure 12 shows a dendrogram of the speech setting effects per

---

\(^{18}\) As the pv-clust package has difficulties handling missing cases, only linguistic variables were included which occurred in the speech of all speakers in all speech settings.
speaker, with rectangles marking clusters sufficiently supported by the data. For ease of interpretation, Table 4 indicates per speaker in which cluster the situation effects are included.

The dendrogram in Figure 12 first of all indicates little interspeaker variation in the standard language test. Corroborating the correspondence regression results, the cluster analysis combines all individual standard language tests – except the one by speaker a4 – in one cluster. On the basis of the original frequency table, we can assume that the exceptional position of a4 should be attributed to the use of non-standard features such as expletive dat (“that”), h-deletion and t-deletion in the standard language test (cf. biplot in Fig. 7 displaying a stronger association of speaker a4 with non-standard features in the standard language test). Also in line with the correspondence regression are the dialect test effects: these effects are scattered over several clusters, especially separating younger speakers from older ones. Here again, an age effect in dialect competence emerges.

Looking at the spontaneous informal speech settings, Table 4 shows that for the majority of speakers, the regional and supraregional informal setting can be found in the same cluster, hinting at a similar type of language use in both settings. The interpretation of the individual interview setting effects is more complicated, as many of these effects are found in clusters which are not convincingly supported by the data. In general, however, Table 4 indicates that several speakers do differentiate between the informal conversations with friends and the interview setting, as these types of speech settings are found in different clusters. All conversations of speaker b4 however can be found in the same cluster. As was the case with the correspondence regression, the cluster analysis indicates fairly monoglossic repertoires for some speakers, and di- or diaglossic personal repertoires for others.

6. Non-metric multidimensional scaling

6.1 About the technique

A third technique to be discussed here is multidimensional scaling (also MDS), which is closely related to correspondence regression.19 The techniques have in common that they attempt to find structures in distance matrices by creating lowdimensional representations in which Euclidean distances between data points maximally represent the original distances from the multidimensional space (Baayen 2008: 136). The outputs of multidimensional scaling and correspondence regression are hence largely the same: multidimensional scaling also generates plots in which the distances between data points are indications of the associations between those data points. Multidimensional scaling differs from correspondence regression in the type of data which is expected

19 Correspondence analysis is even often considered a type of multidimensional scaling (see for example Baayen 2008, Cox/Cox 2008).
as input: whereas correspondence regression focuses on count data, multidimensional scaling is much more free concerning data input. Secondly, the type of multidimensional scaling we will present here – non-metric multidimensional scaling – uses a different algorithm in the dimensionality reduction. The singular value decomposition which is used in correspondence regression ensures that the distances obtained after dimension reduction approximate linear transformations of the original distances in the distance matrix. In non-metric multidimensional scaling by contrast, *monotonically increasing relations* are strived at, meaning that only the order of rank of the original distances should be maintained (Cox/Cox 2008: 323, Kruskal 1964). The degree to which the original distances correspond with the distances after dimension reduction is measured by means of the stress-value $S$ in multidimensional scaling.\(^{20}\) A value smaller than 0.15 or 15% would point towards an acceptable dimension reduction (Borg/Groenen 2005: 47), although a lot depends on the original multidimensionality of the data. An $S$-value of 0.15 for an originally 50-dimensional dataset reduced to 2 dimensions for instance has a different meaning than an $S$-value of 0.15 for a 5-dimensional dataset reduced to 4 dimensions. Borg/Groenen (2005: 47–48) therefore suggest to use as guideline a scree plot in which the stress-value is visualized in function of the number of dimensions (cf. correspondence regression). Similar to the correspondence regression procedure, one can then look for the point at which a further increase in dimensionality does not significantly reduce the stress value.

In this study, we complement correspondence regression with non-metric multidimensional scaling using the isomultidimensional scaling-function in the R-package MASS. As already mentioned in section 3, the prime reason for this combination of methods is that in dimension reduction analyses a part of the original information is always lost and it is hence interesting to compare the results of different dimension reduction techniques. Results emerging in both analyses can be considered more reliable than results occurring in only one analysis.

### 6.2 Results

Figure 13 shows the MDS-plot with the situation effects emerging in the language use of the ten studied East Flemish women. For this analysis, a frequency table was first built with for each situation relative frequencies\(^{21}\) per variant, after which a distance matrix was calculated on the basis of euclidean distance measure. When comparing the MDS-plot with the biplot in Figure 5, largely similar structures can be observed. In the MDS-analysis, the regional and supraregional speech situations and the interview setting also strongly correlate (cf. biplot correspondence regression), while there are

\(^{20}\) For mathematical details on this stress value, we refer to Cox/Cox (2008: 323) and Borg/Groenen (2005: 41–42).

\(^{21}\) The relative frequencies were calculated by dividing the number of times a variant occurred in a specific situation by the number of times it could occur (variable frequency).
clear differences between those situations on the one hand and the dialect test and the standard language test on the other hand. We can conclude that the situation effects in the multidimensional scaling analysis are very similar to those in the correspondence regression and that the pattern can hence be considered robust.

Secondly, it is also important to verify to which degree the variant patterns observed in sections 4 and 5 are also supported by a MDS-analysis, given that these patterns are fundamental in describing repertoire structure. Figure 14 (see appendix) shows a MDS-plot generated on the basis of relative frequencies per situation for each variant attested in Ghent. Euclidean distances served as input for the multidimensional scaling-analysis; colour codes mark the variant clusters distinguished in section 5 on the basis of a cluster analysis.

When comparing the multidimensional scaling-plot with the correspondence bi-plot, the patterns at first sight seem very divergent. Recall from section 4 that in the correspondence bi-plot the x-axis seemed related to standardness, whereas on the y-axis
exogeneous dialect features were found at the bottom, standard features on the top and variants typically associated with *tussentaal* at the centre. In Figure 14 (see appendix), however, other dimensions structure the MDS-plot. In this plot, variants are ordered in a circle-like shape, with equivalent variants of binary variables – such as the variants expletive *dat* ("exdat") and the absence of expletive *dat* ("gnexdat") – mirrored along axes running through the midpoint of the graph. The clusters distinguished in the cluster analyses can be clearly recognized:

- the high frequent standard language features (marked in red) are plotted in the left of the graph,
- the low frequent dialect variants in the right (marked in yellow) – without distinction between variants that are or are not endogenous in the Ghent dialect,
- the low frequent standard language features (marked in grey) in the top right corner,
- and the high frequent dialect features (typically associated with *tussentaal*, marked in green) in the bottom left.

The features which were combined in the purple cluster, a cluster not convincingly supported by the data in the cluster analysis, can in the multidimensional scaling-plot be observed dispersed around the midpoint and in the top of the graph. The big distances between the purple features in the MDS-plot mirror the low AU p-value for this cluster in the cluster analysis. The x-axis in the multidimensional scaling-plot is related to general frequency: in the left high frequent variants are plotted, in the right low frequent variants. The y-axis on the other hand is connected with standardness, stretching from non-standard variants in the bottom to standard in the top.

Compared to the biplot resulting from correspondence regression, the opposition between equivalent variants and the general corpus frequency comes more clearly to the fore in the multidimensional scaling-analysis. This has to be attributed to the implicit row normalisation applied in correspondence regression to calculate $X^2$-distances. As was already described in section 4.1, correspondence regression departs from row proportions when calculating the distances between rows, whereas in multidimensional scaling the values in the crosstable are not automatically rescaled to row- or column proportions. This explains why the inversely proportional relation in the frequencies of variants is more prominent in the multidimensional scaling plot than in the correspondence regression result.\(^{22}\)

Leaving the dimensionality of the plots aside and focussing on the different correlations between variants, the multidimensional scaling plot shows roughly the same stylistic patterns as the correspondence biplot. As in the cluster analysis, however, the multidimensional scaling plot does not distinguish between different types of "yellow" variants, given that no age groups were distinguished in the frequency table used as input.

\(^{22}\) When the multidimensional scaling analysis is repeated using a frequency table in which row proportions are calculated, the patterns in the multidimensional scaling plot are identical to the ones in the biplot resulting from correspondence analysis.
To study the individual language repertoires, to conclude, a frequency table was built with the relative frequencies of each variant for every combination of speaker and situation. On the basis of this frequency table a distance matrix with Euclidean distances was generated, which served as input for the multidimensional scaling analyses. Figures 15 and 16 (see appendix) show the results of this analysis, with the situation effects for respectively the 20–35 year old and 50–65 year old speakers.

Interpreting the multidimensional scaling plot with the interaction effects for speaker and situation is more complicated than was the case for the correspondence biplot, as we have no confidence ellipses to facilitate the interpretation. In general however, Figures 15 and 16 confirm the observation that idiolectal variation is limited in the standard language test, and substantial in the dialect test, with the older speakers (plotted in Fig. 16) diverging more clearly from the language use in the spontaneous speech settings than the younger ones (plotted in Fig. 15). Also in line with the cluster and correspondence analyses is the variation observable in “repertoire width” when focussing on the spontaneous speech settings: for some speakers, the language use in the regional and supraregional informal conversations and the interview barely differs (cf. speaker b4, a3), whereas others differentiate more clearly between informal conversations on the one hand and the interview on the other (cf. b1, b5). Here again, the structures observed in the correspondence and cluster analysis are confirmed by non-metric multidimensional scaling.

7. Discussion & conclusion

At the beginning of this paper, we set out to address the question how language repertoire structure can be empirically studied. We focused on East-Flemish highly educated women in five speech settings and formulated two specific research questions: (1) should the spoken language repertoire in Ghent be described as di- or rather as diaglossic and (2) can varieties be distinguished within this repertoire?

Concerning question 1, our data indicate that everything depends on the level at which the repertoire is studied. When focussing on the community repertoire or the “Gesamtsprachsystem” (Schmidt/Herrgen 2011: 31) – in our case the “community” of highly educated Ghentish women – a diaglossic structure emerges: the different analyses all point to the existence of a continuum stretching from dialect variants mainly used in informal situations by older speakers to standard variants mainly used when the Ghentish speakers are explicitly asked to speak standard. In between, features are found which show associations with multiple situations and which are used differently by different speakers. The association between certain dialect features (such as “gruut”, “kaëd” and “zeun”) and the older speakers indicates that the community repertoire is marked by dialect loss (cf. Auer’s Type D repertoire).

When we however analyse the stylistic behaviour of the individual respondents in the five studied speech settings, the individual repertoires are not necessarily diaglossic. When neglecting the fairly artificial language use in the dialect and standard language
test, some speakers for instance barely adapted their language in the different situations (at least regarding the language features studied), resulting in a monoglossic structure, whereas others did shift significantly between the different situations, showing a diaglossic or diglossic repertoire. Of course we have to bear in mind that we only studied a limited number of speakers and a selection of language features, but the results underline the observation made in Ghyselen (2016a: 40) that in characterizing language repertoires, it is essential to make a distinction between the level of the individual speaker and the one of the community: while “individual speakers may have diglossic repertoires with a clear structural distance between dialect and some form of intended standard, a combination of all those individual repertoires may yield a diaglossic overall picture” (Ghyselen 2016a: 40).

Pertaining to question 2, we already highlighted in section 2 that the answer strongly depends on the theoretical interpretation of the variety concept, but that we in any case expect a variety to be marked by covariance, meaning that we cannot distinguish varieties in the studied repertoires if we do not find bundles of language features strongly correlating in their socio-situative behaviour. For the Ghentish data, covariance patterns were studied using three multivariate statistical techniques: correspondence regression, cluster analysis and non-metric multidimensional scaling. These techniques revealed several clusters of features in the community repertoire, such as

- a dialect cluster, with features such as the realisation of standard Dutch [ɔː] as [yː] (“gruut”) and the preservation of non-suffixal final schwa (“bedde”), mainly used by older speakers;
- Two closely correlated clusters with on the one hand frequently used dialect features (such as the preverbal personal pronoun ge and expletive that) and on the other hand frequently used standard features (such as single negation, and the absence of subject doubling in 3 singular masculine/feminine, 1 plural and 3 plural),
- a cluster with standard Dutch features which are only used in very formal settings, such as the realisation of final -t in niet (“not”) and that (“dat”).

A tough question is of course how strong the covariance has to be before a group of features can be labelled a variety. In the correspondence regression, the observed clusters showed relatively smooth transitions from one to the other; only the cluster of low frequent dialect variants was clearly separated from other variant clusters. Building on Geeraerts (2010) and Ghyselen/De Vogelaer (2018), we could argue that the overall diaglossic repertoire in Ghent is marked by prototype structures, with prototypical instances of dialect, tussentaal and standard language, but also instances of language use which cannot easily be categorized. Interestingly, the cluster structure detected in Ghent differs clearly from the structures observed by Ghyselen/De Vogelaer (2018) in Ypres, a city at a distance of 64 km (celestial latitude) from Ghent.23 In the Ypres study, a cluster emerged between dialect and tussentaal, a type of speech which

---

23 Ghyselen/De Vogelaer (2018) use the same research design as the here-presented study, hence the results of the two studies lend themselves well for comparison.
was labelled “cleaned-up dialect” by the speakers themselves. The differing structures in Ghent and Ypres highlight the importance of empirically studying repertoire structure.

When studying covariance patterns, it is interesting to also consider the metalinguistic status of these patterns. To what degree are the patterns detected mathematically in language production data also perceived as linguistic types or varieties by language users? It is beyond the scope of this paper to thoroughly discuss all the perceptual data collected in the project, but it is interesting to note that all speakers studied in this paper, when asked to describe the existing language variation in Ghent, report a diaglossic language repertoire: besides the Ghentish dialect (called Gents ‘Ghentish’, plat Gents ‘broad Ghentish’ or echt Gents ‘real Ghentish’) and standard language (labelled Algemeen Nederlands ‘General Dutch’, echt Nederlands ‘real Dutch’ or Algemeen Beschaafd Nederlands ‘General Civilized Dutch’), all speakers distinguish one or more types of intermediate language use.

Speakers a1, a2, a3, a4, a5, b1 and b2 divide the repertoire into three categories, characterizing the intermediate language use as a poging tot standaardtaal (a1, ‘attempt towards standard language’), standaardtaal met een ljuie uitspraak (a2, ‘standard language with a sloppy pronunciation’), iets tussen de twee (a3, ‘something in between’), tussentaal (a4, a5, ‘in-between-language’), onder mekaar taalgebruik (b1, ‘among-one-another language use’) or deftig Vlaams met een Gents accent (b2, ‘decent Flemish with a Ghentish accent’). This intermediate language use is reported to be the daily colloquial language and would be closer to standard language than to dialect. On the basis of such descriptions, it can be assumed that the described language use coincides with the attested intermediate cluster.

Speakers b3, b4 and b5 distinguish two “types” of language use in between dialect and standard language: gekuist dialect (b3, b4, ‘cleaned-up dialect’) or tussentaal die naar dialect toe neigt (b4, ‘in-between-language closer to dialect’) on the one hand and gekuist Nederlands (b3, ‘cleaned-up Dutch’), slordig Nederlands (b3, ‘sloppy Dutch’), slordig ABN (b5, ‘sloppy General Civilized Dutch’) or tussentaal die naar Algemeen Nederlands toe neigt (b4) on the other hand. The “cleaned-up dialect” is according to b4 her former home language; b5 reports speaking it with her children, and b3 mentions it as a levelled type of dialect that would increasingly be spoken in Flanders. None of the speaker reports having spoken cleaned-up dialect in the recordings. The “sloppy Dutch” conversely is the language the informants report to speak in the interview and generally also in daily life. This type of language seems to correspond to the detected intermediary cluster.

Considering the perceptual data, it is intriguing that the metalinguistic reflections of the informants more often than not fit the patterns detected in the production data. The number of participants in our study is of course too small to allow far-reaching conclusions pertaining to the relation between production and perception patterns when studying the structure of language repertoires, but the data do suggest an interesting connection between the two. In this context, Ghyselen/De Vogelaer (2018: 16) argue that the ontological status of production patterns should be factored in when studying variety structure, as including perceptual criteria allows distinguishing categories which are not
only statistically, but also cognitively real. The question is of course how production and perception patterns are exactly related to each other. In a cognitive sociolinguistic framework, the process of social interaction is usually referred to to explain the emergence of linguistic systematicity (cf. Geeraerts 2010: 238, Schmidt/Herrgen 2011). In interaction, members of a speech community or a community of practice seek effective communication and often social cohesion, which results in mutual adaptions of the language behavior towards each other and as such centripetally creates regularities within the community. These regularities are social in nature, but also cognitive (for example Geeraerts 2010: 238), as language users have mental representations of the existing regularities. On the basis of lower-level schemata, for example pertaining to the language use of a single speaker, language users would generalize higher-level schemata, concerning the language use of larger social groups or of specific situations (Kemmer/Barlow 2000). These cognitive schemata are constantly “reconfigured” as language users participate in more and new usage events. Further research is in our view however needed – combining insights of cognitive linguistics and sociolinguistics – to establish the ways in which cognitive patterns and structures in usage data mutually influence each other.

Methodologically, the study shows that multivariate statistical techniques are vital to advance our understanding of the “nature and the internal qualitative structure of repertoires” (Berruto 1989: 552), as only such methods allow studying how language variants correlate in their socio-situative behaviour. The outcome of such multivariate analyses however strongly depends on the socio-situative dimensions along which the data vary and on the language features studied. An important prerequisite for empirical studies on repertoire structure is hence that stylistically stratified datasets are available and that a diverse set of language features is observed. Given the discussed tension between the individual and the community (cf. above), the datasets ideally also allow studying the stylistic behaviour of individuals. The latter criterion is however challenging: in the previous decades large-scale speech and written corpora have been built, but few of them record language users in multiple speech settings. An exception to this rule is the German language area, where three complementary projects were initiated between 2006 and 2008 to systematically map dialect-standard variation:

The “Deutsch heute” project (IDS Mannheim), following König (1989) aims to survey near-standard variation in the speech of informants of both sexes with the highest level of secondary education from all of the German-speaking countries (160 locations; cf. Brinckmann et al. 2008); the “Sprachvariation in Norddeutschland (SiN)” project (Universities of Bielefeld, Frankfurt/Oder, Hamburg, Kiel, Münster, and Potsdam) sets out to examine the entire variative spectrum of female informants of various ages and differing dialect competence from 36 small North German towns (cf. Schröder/Elmentaler 2008); the goal of the “regionalsprache.de (REDE)” project (Marburg) is to capture the entire variative spectrum of male informants of different ages and social backgrounds from 150 locations across the German Federal Republic. (Schmidt 2011: 158)
In 2016, the Austrian DiÖ-project (“Deutsch in Österreich”) was launched, a large-scale project of the Universities of Vienna, Graz and Salzburg and the Austrian Academy of Science, which focuses on language variation, language contact and language perceptions in Austria, and aims at building a stylistically stratified dataset. In the Dutch language area, however, no comparable data collections are available. The Corpus of Spoken Dutch (CGN, “Corpus Gesproken Nederlands”) does contain different speech settings, but the speakers (1) were instructed to speak Standard Dutch and (2) were mostly registered in one type of speech setting only. The material collected for this research paper does contain speech of individuals in multiple situations, but is undeniably limited in regional, situational and social scope. We can only hope that further steps will be undertaken to build a large-scale corpus of intersituational spoken and/or written Dutch focusing on the same language users in multiple settings.

In this paper, we have discussed three multivariate statistical techniques: profile-based correspondence regression, bootstrap clustering and non-metric multidimensional scaling. Notwithstanding the mathematical differences between profile-based correspondence regression, bootstrap clustering and non-metric multidimensional scaling, the analyses presented in sections 4 to 6 led to largely similar conclusions. Each technique however has clear advantages and disadvantages when applied in repertoire structure studies.

The used correspondence regression, to begin with, has as first and foremost advantage that it visualizes the structure of multiple dependent variables in their association with independent parameters such as situation, age and speaker, which makes it easier to interpret the identified structures. Another advantage of the used correspondence regression for our type of research is that it is geared towards count data and that it provides an automated profile-based approach building on the relative frequencies within variables. In the performed cluster analysis and multidimensional scaling, we had to add the profile-based dimension ourselves by first generating a table with relative frequencies, and subsequently feeding this table as input to the cluster/multidimensional scaling-analysis. A third advantage of the performed correspondence regression is that the confidence ellipses generated by means of bootstrapping in the corregp-package are very useful in interpreting emerging structures. A disadvantage however is that the dimension reduction, a core feature of correspondence regression (and also of multidimensional scaling-analysis), always implies a loss of information, and that it is difficult to assess how relevant this lost information might be from a sociolinguistic viewpoint.

The advantages and drawbacks outlined for the correspondence regression immediately highlight the drawbacks and advantages of the other two techniques used, cluster analysis and non-metric multidimensional scaling. A practical disadvantage of the cluster analysis and multidimensional scaling as applied in our study is that the techniques are not profile-based *an sich*; the researcher has to model the data in a matrix with relative frequencies him- or herself. Secondly, the techniques display distances between data points, but offer little information on the sense in which certain data points correlate. The researcher has to go back to the original frequency table to explain the observed patterns, but when datasets get larger and more complex, this interpretation exercise...
can get very complicated. The big advantage of the performed cluster analysis is however that it does not use dimension reduction algorithms, and that we can hence expect less “loss” of potentially relevant information. Secondly, the approximately unbiased p-values in the pv-clust package are useful in that they provide essential information on the degree to which clusters are supported by the data. Concerning the performed non-metric multidimensional scaling, we already indicated that it is appealing in its use of another dimension reduction technique than singular value decomposition.

The above clearly indicates that each of the three analysed techniques has advantages and drawbacks. The techniques are therefore – in our view – ideally combined and complemented, as only structures emerging in multiple analyses should be considered robust. In this study, we applied cluster analysis on the “raw” frequency data, as we wanted to offer an introduction to the technique, but we could also use the results of the dimension reduction techniques (for example correspondence coordinates of variants in multidimensional space) as input for a cluster analyses (cf. Grieve 2016, Vandeven 2016, Ghyselen/De Vogelaer 2018). As such, the cluster analysis is automatically profile-based; the researcher does not have to model the data in a matrix with relative frequencies him- or herself. Moreover, while only a limited number of dimensions can be plotted surveyably in a correspondence biplot, a larger number of dimensions can serve as input for the cluster analysis. By combining the results of the cluster analysis (i.e. the dendrogram) with that of the correspondence regression (i.e. a biplot), a thorough insight can be achieved in the data structure.

In sum, we hope to have shown that the application of multivariate statistics in linguistic research opens up many possibilities for the empirical exploration of structures in language repertoires. As is the case for all statistical research, the usefulness of the techniques strongly depends on the research design: only when applied insightfully on carefully designed datasets (both in terms of corpus structure and variable selection) do the techniques allow insight in repertoire design. In this paper, 25 linguistic variables were analysed in the speech of 10 Flemish speakers – selected in function of their education level, regional background and sex – in 5 speech settings. We hope that the research can be an impetus towards the development and exploration of larger-scale (open-source) stylistically stratified corpora.
References


Mapping the Structure of Language Repertoires: on the Use of Sociolectometric Methods


Mapping the Structure of Language Repertoires: on the Use of Sociolectometric Methods


Appendix

Fig. 6: Correspondence plot Ghent with main effects for situation (detail [-1:0.5])
Fig. 7: Correspondence plot Ghent with interaction effects between speaker (25–35 years old) and situation (dia = dialect test, reg = conversation with friend of same region, sup = conversation with friend of different region, int = interview, st = standard language test), incl. detail [-0.25:0.25]. Every colour represents one speaker; confidence ellipses are not plotted for reasons of surveyability.
Fig. 8: Correspondence plot Ghent with interaction effects between speaker age 50–65 years old) and situation (dia = dialect test, reg = conversation with friend of same region, sup = conversation with friend of different region, int = interview, st = standard language test), including detail [-0.25:0.35]. Every colour represents one speaker; confidence ellipses are not plotted for reasons of surveyability.
Fig. 10: Dendrogram of variants on the basis of the relative frequencies per variant for the five studied situations

Distance: euclidean
Cluster method: ward.D2
Fig. 11: Correspondence plot Ghent with main effects for situation and colours indicating cluster analysis classification
Fig. 14: MDS-plot with correspondences between variants on the basis of the relative frequencies per situation (reduction to two dimensions, stress: 7.70, colours visualizing the variant groups distinguished in section 5 on the basis of a cluster analysis).
Fig. 15: MDS-plot with situation effects of 25–35 year old speakers on the basis of the relative frequencies per variant (reduction to two dimension, stress: 9.9). Each colour represents one speaker.
Fig. 16: MDS-plot with situation effects of 60–65 year old speakers on the basis of the relative frequencies per variant (reduction to two dimension, stress: 9.9). Each colour represents one speaker.
Fig. 17: Correspondence biplot for dimension 1 vs. 3
Fig. 18: Correspondence biplot for dimension 2 vs. 3