Schmidt redux: How systematic is the linguistic system if variation is rampant?

Dirk Geeraerts
University of Leuven

Abstract

Harder (2003) describes a socially interactionist conception of usage-based linguistics: the structures that emerge from the level of usage have a social status, as part of a shared inventory of linguistic means of expression. However, this social turn in the conception of a usage-based language system has two consequences that are not receiving due attention in Harder's formulation. First, a social view of linguistic structure entails a variationist conception of the linguistic system. Instead of one homogeneous system, 'the' language takes the form of a cluster of lectal systems, each of them fragmentary with regard to what we would traditionally consider to be 'the' language. Second, we have to take into account a further degree of heterogeneity, to the extent that lects have prototype structure: the linguistic phenomena that constitute the lect may be more or less typical for the lect in question, and will thus each have to be studied on their own.

1 The ontological status of the language system in a usage-based model

A usage-based conception of language inevitably raises questions about the ontological status of the linguistic system. Usage phenomena are now broadly seen as an integral and crucial part of linguistic description because there is a dialectal relationship between Structure and Use: individual usage events are realizations of an existing systemic structure, but at the same time, it is only through the individual usage events that changes might be introduced into the structure. (For different aspects and versions of the usage-based research paradigm, see Hopper 1998, Barlow and Kemmer 2000, Bybee 2001, 2006, Geeraerts 2002, Tomasello 2003. For an evaluation of the technical state of the art of the usage-based trends in linguistics, see Tummers, Heylen and Geeraerts 2005.) But how then, in such a dialectic view of the relationship between Structure and Use, does the system exist - if at all? The Use pole of the dialectic relationship is readily identifiable: it exists in the form of actual instances of language use, whether active or passive. But where do we find Structure?

In 'The status of linguistic facts: Rethinking the relation between cognition, social institution and utterance from a functional point of view', Peter Harder (2003) offers a foundational contribution to the debate that answers the question in a socially interactionist vein: 'Like all social facts, such as everyday routines, fashion, and the value of money, the state of a language has no precise location in the community. Social facts are sustained by individual mental states without being reducible to them, existing within boundaries of variation that are continually created and modified as a result of feedback mechanisms in networks of interactive practices' (2003: 69). If we attempt an analytic reformulation of this synthetic statement, which is further developed in Boye and Harder (2007), we may highlight the following aspects.

First, language as structure is a social fact, as an observable regularity in the language use

realized by a specific community. Second, it is at the same time a cognitive fact, because the members of the community have an internal representation of the existing regularities (the system) that allows them to realize the same system in their own use of the language. Third, the same mechanism that allows the existing collective regularities to enter the individual minds is also the one that allows regularities to emerge to begin with, viz. mutual influence in social interaction. People influence each other's behavior, basically by co-operative imitation and adaptation, and in some cases by opposition and a desire for distinctiveness. Paying attention to what others do, however subsconsciously, thus creates a mental representation of the collective tendencies in the behavior of the community; adapting one's own behavior to those tendencies, reaffirms and recreates the tendencies. And fourth, in the same way that the existing regularities emerged from actual interaction, changes may emerge; as such, a degree of variation is an inevitable aspect of any synchronic state of the language.

This view ties in with an emerging line of research in Cognitive Linguistics that takes the view that a language can only be adequately conceived of if one takes into account the socially interactive nature of linguistic communication. Examples of this strand of research include Sinha (2007) on language as an epigenetic system, Zlatev (2005) on situated embodiment, Itkonen (2003) on the social nature of the linguistic system, Verhagen (2005) on the central role of intersubjectivity in language, and Geeraerts and Grondelaers (1995), Palmer (1996) and Kövecses (2005) on the cultural aspects of language. While most of the references cited here focus on theoretical arguments, cross-linguistic differences, and historical variation, variationist research linking up with sociolinguistics is still relatively underrepresented within the social tendencies within Cognitive Linguistics. See however Kristiansen and Dirven (2008) for a collection of variational studies within the framework of Cognitive Linguistics. One of the consequences of the present paper, linking up with the argumentation in Geeraerts (2005), is precisely that such an extension towards the sociolinguistic and dialectological realm is inevitable once the ‘social turn in Cognitive Linguistics’ (Harder, Forthcoming) is taken.

A graphical representation of the model derived from the quote taken from Harder (2003) may be found in Figure 1. For each individual, we first distinguish between outward, externally observable usage, and the mental representation that underlies language use. That mental system is graphically represented as a collection of forms, symbolically couched in a 'thought cloud'. The individual's system does not only correspond to the individual's usage, but it is also influenced by other people's usage: what other people do has an influence on what we know about language behavior; by interiorizing the behavior that we notice in other's, our mental representation of language is attuned to that of the community. However, we never interact with the entire community, but we interact in specific networks. The figure distinguishes simplistically between two dimensions that define social networks, each represented by an elliptical set representation at the bottom of the figure: the boys vesus the girls, and the dotted figures versus the others. One individual, needless to say, may be characterized on both dimensions, and in that sense, the interactive networks overlap. Also, it should be kept in mind that lectal differences in actual speech are not just determined by speaker characterisics, but also by contextual features like communicative situations underlying different registers. Because the interactions in the community are not complete, the individual mental representations, and the individual usage behavior, are not identical for all the members of the community. In the figure, this is indicated by the fact that the system component is different for each of the three persons. Iconically, the mental representation of the middle figure (which is situated in the overlapping area of two networks) is represented as composed of both the system of the left

hand figure and that of the right hand figure.

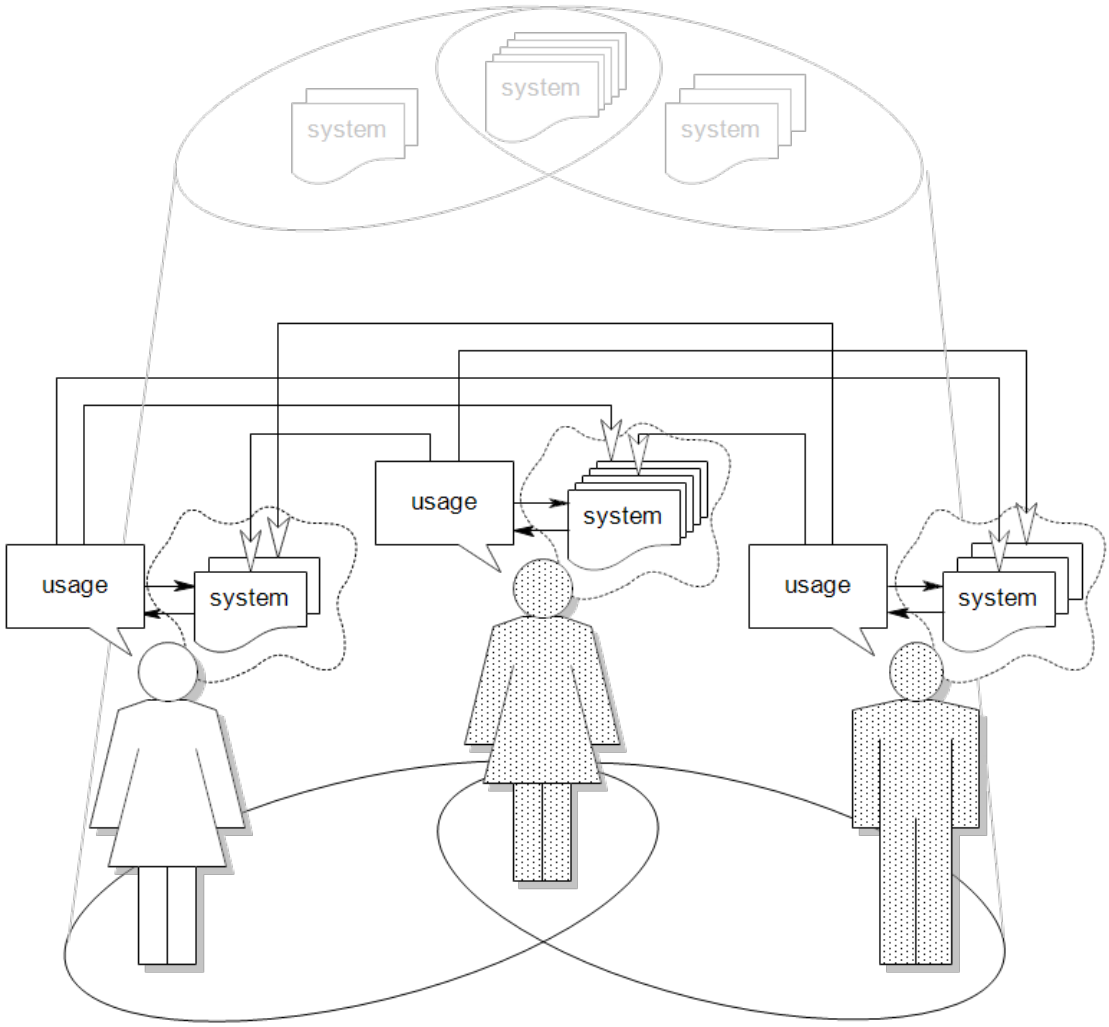

Figure 1
A schematic representation of a usage-based model of language

In the top half of the figure, the observable regularities in the behavior of the characters are asbtracted along the lines of a traditional linguistic description: one that takes the notion of a system for granted (even if the system is a socially structured set of systems), and that does not pay a lot of attention to the dialectic and interactionist aspects. In the usage-based model sketched here, the ultimate reality of the linguistic system resides in the complex dynamic system depicted in the lower half of the figure. The 'system' or 'systems' in the upper half are abstractions only that we should take care not to reify or hypostasize: they have no reality independent of what is going on below. To emphasize the epiphenomenal nature of 'the system' as an abstraction, the upper half of the figure is drawn in grey.

The recognition of social variation in Figure 1 is a crucial addition with regard to the model defined in the quote by Peter Harder. On the one hand, Harder does accept synchronic variation within the lnguistic system, as the quote makes clear, but on the other, the quote refers rather

uniformly to 'the community', without explicitly taking into account the internal social dimensions that shape the interactions within the community. In other words, the variation that is correctly included in the model described by Harder is structured variation, and the structure of the variation is a social one. In that sense, the social model of the language system that is described in Harder (2003) cannot be restricted to a general semiotic or philosophical recognition of the social nature of linguistic facts, but it naturally leads to a sociolinguistic, sociovariationist type of description in which the social dimensions of variation and interaction are an integral part of the description of the linguistic system. *If a usage-based model of the language implies a social conception of linguistic facts, then a social conception of linguistic facts also implies a variationist model of the language.*

2 The unsystematic nature of the linguistic system in a usage-based model

But if we accept the addition of this specification to the model defined by Harder (2003), a further question crops up: how systematic are the subsystems? This question introduces a second dimension of heterogeneity into our conception of the linguistic system. We have already recognized that the linguistic system is not homogeneous within a community, i.e. that there are subcommunities with their own grammar - even though all those grammars cluster together as a system of systems. How homogeneous, then, are the systems of those subcommunities? To arrive at an answer to that question, we have to see that we normally think of a linguistic system as a collection of language forms that behave uniformly under identical lectal circumstances. For instance, we can say that there is a West-Flemish dialect of Dutch because the speakers in the province of West-Flanders share a number of linguistic characteristics, including a monophtongal pronuciation of the standard Dutch diphtong ui and a laryngeal realization of g. In establishing West-Flemish as a (dia)lect, we assume a high degree of co-variation among a number of linguistic forms, more specifically, co-variation in the distribution of those forms over socially distinct groups of speakers - in the case of traditional dialects, a group of speakers that share a geographical origin. We have discarded the idealization of a totally homogeneous linguistic community, but we may still adhere to the view that what varies over the heterogeneous linguistic community are separate linguistic systems: collections of linguistic forms that together constitute an internally coherent unity - systems, in short, in the structuralist sense of mutually co-determining entities. But to what extent is that actually the case?

The general question to be answered, then, is the following: next to the question how heterogeneously groups of language users behave with regard to a given set of linguistic forms, we need to investigate how heterogeneously sets of language forms behave with regard to lectal dimensions. To what extent do language forms behave in bundles, clusters, sets - systems - of distributionally equivalent elements?

It will be readily appreciated that this question is a variant of the neogrammarian controversy. The neogrammarian concept of sound laws, as originally formulated by Brugmann and Delbrück (1886-1916), or Schleicher (1850) implies that changes affect an entire system of forms at the same time. Phonetic systems change as systems, i.e. as wholes in which individual expressions featuring a certain phoneme have no special position (apart from well-established categories of exceptions to the sound laws, which may be lexically particular, like the occurrence of assimilations). Conversely, the alternative view as already formulated by Schuchardt (1885)

argues that the changes in a language may be much less systematic than suggested by the neogrammarians. The later concept of lexical diffusion (Wang 1969, 1977) likewise argues against an all too systemic conception of linguistic change, focusing on the lexical mediation of sound change. Shifting from diachronic changes between languages to synchronic variation between lects, the question we try to address involves a similar distinction between a 'system first' and a 'individual item first' approach.

To see what is conceptually and methodologically involved in such a question, we need to relate to another chapter of nineteenth century linguistics. The methodology and the conceptual analysis we need to answer our question are somewhat similar to the approach that led to the discovery of the 'Rhenish fan' and Schmidt's formulation of the wave theory of linguistic change (Schmidt 1872). Given the various elements of the High German sound shift, do they geographically occur together? Do the pf that distinguishes High German *apfel* from Low German and Dutch *appel*, the f that distinguishes High German *dorf* from Low German and Dutch *dorp*, the ch that distinguishes High German *ich* from Low German and Dutch *ik* always occur as a bundle? As it turns out, the dialect landscape in the neighbourhood of the Rhine shows that the isoglosses for the various sound changes do not coincide. The demarcation between *ik* and *ich* lies much more to the north than that between *dorp* and *dorf*, for instance. Between the 'pure' High German situation and the 'pure' Low German / Dutch situation, a number of dialects exhibit transitional configurations. Metaphorically, each separate sound change spreads like a wave over the dialect landscape, but some waves reach farther than others.

In terms of the systematicity question, the wave theory goes along with the Schuchardtian position that different forms may have their own history. At the same time, the combination of different forms identifies different dialect areas. If we look at the major isoglosses, different intermediate situations can be distinguished between the Dutch *ik – maken – dorp – dat – appel* situation, and the High German *ich – machen – dorf – das – apfel* situation. (The words here are only exemplary for the lexical sets featuring the relevant phonemes.) Schematically, the different dialect areas are demarcated as follows:

ik	maken	dorp	dat	appel
ich	maken	dorp	dat	appel
ich	machen	dorp	dat	appel
ich	machen	dorf	dat	appel
ich	machen	dorf	das	appel
ich	machen	dorf	das	apfel

Each dialect area constitutes a different linguistic system: different forms may have their own distribution when you look at them separately, but clusters of co-occurring forms still define lectal systems. However, a more radical interpretation is possible as well, when it is recognized that the variability is so outspoken that the notion of 'dialect' – the notion of a system of co-occuring formal phenomena – itself loses its substance. This is a point of view expressed by Paris in a lecture of 1888 (Paris 1888), where he states that dialects in reality do not exist: 'Il n'y a réellement pas de dialectes ; il n'y a que des traits linguistiques qui entrent respectivement dans des combinaisons diverses, de telle sorte que le parler d'un endroit contiendra un certain nombre de traits qui lui seront communs, par exemple, avec le parler de chacun des quatre endroits les plus voisins, et un certain nombre de traits qui différeront du parler de chacun d'eux. Chaque trait linguistique occupe d'ailleurs une certaine étendue de

terrain dont on peut reconnaître les limites, mais ces limites ne coïncident que très rarement avec celles d'un autre trait ou de plusieurs autres traits'.

In a similar vein, we now have to ask whether linguistic phenomena that might be susceptible to lectal variation within a language, always occur in clearly distinguishable bundles, and specifically, if the bundles that they occur in correspond to easily identifiable lectal differences. The latter addition is crucial, because by logical necessity, there is always a level at which linguistic phenomena occur in distinctive bundles. But if this were the level of the idiolect, for instance, or even worse, the level of an individual usage event, we would not be inclined to talk of a social system.

Also, the structure of the lectal variables we need to take into account is more complicated than in the original Rhenish fan. In Schmidt's aproach, the basic lectal variables are dialects, distinguished by a single variational dimension, viz. geography. With regard to our contemporary question, we cannot restrict the analysis to a geographic dimension, but we will take into account various dimensions that may lie at the basis of lectal structure: not just speaker characteristics like age and location, but also situational characteristics like style and register. Like Schmidt, we are interested in seeing how lectal variables pattern together with each other and with linguistic variables, in the same sense in which the Rhenish dialects cluster together in dialect areas on the basis of bundles of linguistic variables. However, the lectal space to be explored is a multidimensional one, rather than a monodimensional one. We will see presently how this methodological challenge can be met.

3 Introducing the case study

As an illustration of the issues involved, we will focus on the study of colloquial Belgian Dutch conducted by Koen Plevoets as a PhD project supervised by the author the present chapter (see Plevoets 2008a, 2008b). This project is part of a line of research conducted within the research team Quantitative Lexicology and Variational Linguistics of the university of Leuven that investigates various aspects of a lectally enriched multivariate grammar (De Sutter 2005, De Sutter, Speelman and Geeraerts 2008; Grondelaers, Geeraerts, Speelman and Tummers 2001, Grondelaers, Speelman and Geeraerts 2008; Tummers 2005, Tummers, Speelman and Geeraerts 2004; Van Gijsel 2007, Van Gijsel, Speelman and Geeraerts 2008). The investigation carried out by Plevoets is based on the Belgian Dutch (or Flemish, if one wishes) data from the Spoken Dutch Corpus (Oostdijk 2002). Informally speaking, the investigation tries to establish whether a number of linguistic phenomena that were revealed by previous studies (see Geeraerts 2001 and the publications just mentioned) to be typical of colloquial Belgian Dutch do indeed pattern in a lectally homogeneous way: when we look at the way in which the different linguistic variables occur together, do they occur in bundles that may be readily interpreted in a lectal way, i.e. as corresponding to what we would traditionally consider to language varieties with their own grammar?

In more technical terms, in order to arrive at an objectively based grouping of the target concepts, we use a statistical analysis known as *correspondence analysis*, a type of cluster analysis. Correspondence analysis (which may be considered a counterpart for non-metric data of what principal components analysis achieves for numeric data) is a technique for jointly exploring the relationship between rows and columns in a contingency table. Correspondence analysis can be thought of as trying to plot a cloud of data points (the cloud having height,

width, thickness) on a single plane to give a reasonable summary of the relationships and variation between them (Benzécri 1992, Greenacre 1984). In a correspondence plot, the x-axis and the y-axis do not have an a priori interpretation; rather, the interpretation of the plot involves the pattern that emerges from the grouping of the data points. Points that are positioned close to one another have a positive association between them. The interpretation of the plot, in other words, takes the form of identifying the clusters of points in the plots. These points, of course, will be of two types: points representing linguistic variables on the one hand, and points representing lectal variables on the other.

(It should also be remarked that the correspondence analysis carried out by Plevoets is not a stratightforward correspondence analysis, because it adapted to the 'profile based' methodology developed in Geeraerts, Grondelaers and Speelman 1999, Speelman, Grondelaers and Geeraerts 2003. This is a technical point that need not concern us here, however.)

3. 1 Specifying the lectal variables

The corpus that we use for our study is the Spoken Dutch Corpus (Corpus Gesproken Nederlands, abbreviated as CGN). Its size is 10 million word tokens in sum, two thirds of which stems from The Netherlands, and one third from Flanders. The CGN is a stratified corpus, in that the linguistic material is sampled from different types of speech situations, called 'components'. They are the following fifteen:

- a: Spontaneous conversations ('face-to-face')
- b: Interviews with teachers of Dutch
- c: Spontaneous telephone dialogues (recorded via a switchboard)
- d: Spontaneous telephone dialogues (recorded on MD via a local interface)
- e: Simulated business negotiations
- f: Interviews/discussions/debates (broadcast)
- g: (political) Discussions/debates/meetings (non-broadcast)
- h: Lessons recorded in the classroom
- i: Live (eg sports) commentaries (broadcast)
- j: Newsreports/reportages (broadcast)
- k: News (broadcast)
- l: Commentaries/columns/reviews (broadcast)
- m: Ceremonious speeches/sermons
- n: Lectures/seminars
- o: Read text

These 15 components will prove highly valuable to our analyses, as they enable us to capture the stylistic differences of the variables. One remark to be made beforehand concerns the fact that component e (simulated business negotiations) has material only from The Netherlands and is lacking for Flanders. Also, component o could be removed because it is non-spontaneous, but we have kept it in for comparison.

Furthermore, each utterance is annotated for its speaker's characteristics, such as Region, Age, Sex, Educational level, and Occupational level. Flanders has the following coding scheme, based on the provinces and the traditional dialect areas that they represent:

- brab: Flanders, central region (Antwerpen and Vlaams-Brabant)

- ovl: Flanders, transitional region (Oost-Vlaanderen)
- wvl: Flanders, peripheral region 1 (West-Vlaanderen)
- lim: Flanders, peripheral region 2 (Limburg)

With respect to Age, the CGN only lists the speaker's birthyear. As this level of granularity might be too fine-grained for our analyses, we code instead for the decade in which the speaker was born. Consequently, we code for generations, following the classification of the sociologist Becker (1992). The generations he distinguishes on the basis of sociological criteria are the following:

- pre the pre-war generation, born between 1910 and 1929
- sil the 'silent' generation, born between 1930 and 1939
- pro the protest generation, born between 1940 and 1954
- los the 'lost' generation, born between 1955 and 1970
- pra the pragmatic generation, born after 1970.

The variable Sex makes the obvious distinction between male (M) and female (F) speakers. Educational Level has a ternary structure, distinguishing between high (hie), middle (mid), an dlow (low). Occupational Level, finally, looks as follows:

- occA: occupation in higher management or government
- occB: occupation requiring higher education
- occC: employed on the teaching or research staff in a university or a college
- occD: employed in an administrative office or a service organisation
- occE: occupation not requiring any specified level
- occF: self-employed
- occG: politicians
- occH: employed with the media, entertainment or artistic sector
- occI: student, trainee
- occJ: having no job.

3.2 Specifying the linguistic variables

The investigation carried out by Plevoets focuses on morphological features of colloquial Belgian Dutch. While lexis and syntax have also been studied (see the references mentioned above), morphology has been cited as the most typical characteristic of colloquial Belgian Dutch (Goossens 2000). We consider three groups of variables: diminutive formation, adnominal variables, and pronominal variables.

Dimunitive formation involves the contrast between diminutives on -je, which is the standard form, and diminutives on -ke, which is the colloquial form: s*toeltje, boompje, tafeltje, pakje* versus *stoeleke, bomeke, tafelke, pakske* (small chair, tree, table, package; a full description involves an analysis of allomorphs like -je, -tje, -pje in the standard case and -ke, -eke, -ske in the colloquial case, but we will not go into these details here). In the correspondence analyses that we will present, these variables are represented as dim.j versus dim.k.

The variation in the adnominal variables involves a pattern of inflection that is related to the gender of the nouns, and to the first element of the noun. Dutch has a three-gender system, with masculine, feminine and neuter nouns. The primary variation consists of colloquial forms on schwa appearing next to the standard forms. So we get the following distribution, where the

right hand column indicates the colloquial forms.

indefinite article ('a')	een	ne
negative pronoun ('none')	geen	gene
distal demonstrative ('that')	die	dieje, diene
possessive pronoun 1 sg ('my')	mijn	mijne
possessive pronoun 3 sg masculine ('his')	zijn	zijne
possessive pronoun 3 sg feminine ('her')	haar	hare
possessive pronoun 3 pl ('their')	hun	hunne

In the possessive pronoun of the second person, we get a lexical alternation: the standard forms on j (singular *jouw*, plural *jullie*) are replaced by *uw* in both numbers. In the standard language, *uw* forms would be polite possessives as opposed to the familiar forms on j, but the colloquial register does not make such a distinction between T and V pronouns. As such, the variational value of *uw*-forms is ambiguous.

The basic system is complicated by two factors. To begin with, the colloquial forms on schwa basically occur only with masculine nouns, but the gender system of Dutch is characterized by a simplifying drift towards a two-gender system, distinguishing the nouns that take the definitive aricle *de* (the older masculine and feminine classes) with nouns that take *het* (the neuter class). A number of originally feminine nouns, then, are shifting towards the masculine class, a process that has progressed further in Netherlandic Dutch than in Belgian Dutch.

When measuring the proportion of colloquialisms that we find, we therefore make a distinction between three classes of nouns: the ones that have always been masculine, the ones that are registered in the official spelling dictionary of Dutch as having both genders, and the ones that are considered to be feminine, but that may still be touched by the tendency towards masculinization. The odds of getting colloquial forms in these three classes are different. In the traditionally masculine class, the odds will be higher than in the traditionally feminine class, for instance, because in the latter group, the colloquialisms will basically only occur with the nouns that have shifted towards the masculine gender.

A further complication is the fact that the colloquial forms on schwa get an extra n before vowels, b, d, t, and h. So, we get *ne man* versus *nen hond* ('a man', 'a dog'). Crucially, this variation also occurs with adnominal forms that already have a schwa in their standard form. So, the definite article *de* is the standard form for all masculine and feminine nouns, but with masculine nouns beginning with a vowel, b, d, t, or h, *den* is the colloquial variant. This introduces a set of additional markers of colloquial speech, on top of the n-variants of the list presented above:

definite article ('the')	de	den
distributive pronoun ('each')	elke	elken
universal pronoun ('every')	iedere	iederen
proximal demonstrative ('this')	deze	dezen
possessive pronoun 1 pl ('our')	onze	onzen

Adjectives too are included in this variation. The standard form for adjectives used with masculine or feminine nouns features an inflectional schwa. In colloquial speech, these inflected adjectives get an n-ending under the same conditions as the articles and pronouns; the articles preceding the adjectives are then phonotactically sensitive to the initial element of the adjectives. So, where standard speech has *een magere man, een hongerige man, een magere*

hond, een hongerige hond ('a thin man', 'a hungry man', 'a thin dog', 'a hungry dog'), colloquial Belgian Dutch has *ne magere man, nen hongerige man, ne mageren hond, nen hongerigen hond*.

In the plots that we will present later, these various forms are represented schematically. Thus, de.m stands for the standard use of the definite article with (traditionally) masculine nouns, and de.f with (traditionally) feminine nouns. The labels den.m and den.f likewise indicate the use of the den-variant, before the vowels and consonants that allow for the alternation. In the variables that have a colloquial schwa-variant, the schwa cases and the schwa+n cases are represented together as markers of colloquial speech. So, een.m is the standard form of the indefinite article for masculine nouns, and ne.n.m stands for either the *ne* or the *nen* alternative.

Taken together, we can then distinguish between the variants that indicate standard language, and the variants that indicate colloquial language. Without going into further detail, the following overview may help to interpret the plots. For each of the categories, the first line specifies the standard language forms, while the second line lists the forms that are typical for colloquial Belgian Dutch.

definite article	de.m, de.c, de.f
	den.m, den.c, den.f
indefinitive article	een.m, een.c, een.f
	nen.n.m, ne.n.c, ne.n.f
negative pronoun	geen.m, geen.c, geen.f
	gene.n.m, gene.n.c, gene.n.f
distributive pronoun	elke.m, elke.c, elke.f
	elke.n.m, elke.n.c, elke.n.f
universal pronoun	iedere.m, iedere.c, iedere.f
	iedere.n.m, iedere.n.c, iedere.n.f
proximal demonstrative	deze.m, deze.c, deze.f
	deze.n.m, deze.n.c, deze.n.f
distal demonstrative	die.m, die.c, die.f
	dieje.n.m, dieje.n.c, dieje.n.f, diene.n.m, diene.n.c, diene.n.f
possessive pronoun 1 sg	mijn.m, mijn.c, mijn.f, m.n.m, m.n.c, m.n.f
	mijne.n.m, mijne.n.c, mijne.n.f, m.ne.n.m, m.ne.n.c, m.ne.n.f
possessive pronoun 2 sg/pl	je.m, jouw.m, jullie.m, je.c, jouw.c, jullie.c, je.f, jouw.f, jullie.f
	uwe.n.m, uwe.n.c, uwe.n.f
	(ambiguous) uw.m, uw.c, uw.f
possessive pronoun 3 sg m	zijn.m, zijn.c, zijn.f, z.n.m, z.n.c, z.n.f
	zijne.n.m, zijne.n.c, zijne.n.f, z.ne.n.m, z.ne.n.c, z.ne.n.f
possessive pronoun 3 sg f	haar.m, haar.c, haar.f
	hare.n.m, hare.n.c, hare.n.f
possessive pronoun 1 pl	onze.m, onze.c, onze.f
	onzen.m, onzen.c, onzen.f
possessive pronoun 3 pl	hun.m, hun.c, hun.f
	hunne.n.m, hunne.n.c, hunne.n.f
adjective	adj.e.m, adj.e.c, adj.e.f
	adj.n.m, adj.n.c, adj.n.f

Next to the adnominal phenomena (articles, adjectives, demonstratives and possessives), we include the personal pronouns into the investigation. First, if we have a look at pronouns in subject function, we need to distinguish between two positions, the basic non-inverted one and the inverted one. The non-inverted position for the first person singular takes the standard form *ik* or reduced *'k*, whereas a colloquial form would be more or less emphatic *ikke*, or a reduplicated form: *ik kom (e)kik mee* ('I come along'). In the inverted position, the standard forms are the same as in the non-inverted position. The colloquial Belgian Dutch form is -*kik* or *ekik: kom ekik mee?* The first person plural has *we* and *wij* as standard forms in non-inverted position, and as colloquial forms *me*, and reduplicating *we* and *wij.* In the case of inversion, *we* and *wij* are standard, *me* is colloquial.

The forms for the second person, singular and plural, are characterized by the same double system as the second person possessives: standard Dutch has a system distinguishing T and V pronouns, whereas colloquial Belgian Dutch has a unitary system. In the non-inverted position, familiar *je, jij, jullie* and polite *u* contrast with colloquial ge, gij. The same holds for inversion, but the forms *de* and *degij* (*komde mee? komdegij mee?* 'do you come along?') add to the inventory of colloquialisms.

The third person singular forms feature masculine *hij* and feminine *zij, ze* in non-inverted position; colloquial counterparts are reduplicating *hij* and *ze.* In inversion, the foms are basically the same, except for the addition of *m* as a colloquial masculine form (*komt m mee?* 'does he come along?') The plural third person forms are zij and ze, both in inverted and non-inverted position; colloquial variants are cases of reduplication: *komen ze zij mee?* 'do they come along?'.

If we then consider the personal pronouns in object functions, no specific phenomena need to be mentioned, except for the typical double system in the second person pronouns: j*e, jou, jullie* are standard forms, but *u* can be both a colloquial form (as the counterpart of the subject forms *gij* and reduced *ge*) and a polite standard form.

Finally, with regard to the reflexive pronouns, the typical feature of colloquial Belgian Dutch is the use of eigen: ik was mijn eigen, hij wast zijn eigen versus the standard form ik was me, hij wast zich 'I wash myself, he washes himself'. The reciprocal pronoun *elkaar* 'each other' has a colloquial Belgian Dutch counterpart in *mekaar*.

Following the same conventions as before, we can now give the following overview of the relevant labels.

subject pronoun 1 sg	ik.not, k.not, ik.inv, k.inv ikke, ik.dbl, k.dbl, k.ik
subject pronoun 2 sg/pl	je.not, jij.not, jullie.not, je.inv, jij.inv, jullie.inv ge.not, gij.not, ge.dbl, gij.dbl,ge.inv, gij.inv, de.inv, de.gij (ambiguous) u.not, u.inv
subject pronoun 3 sg	hij.not, hij.inv, ie, ze.not.vr, zij.not.vr, ze.inv.vr, zij.inv.vr hij.dbl, m.inv, ze.dbl.vr, ze.zij.vr
subject pronoun 1 pl	we.not, wij.not, we.inv, wij.inv, me.not, we.dbl, me.inv
subject pronoun 3 pl	ze.not.mv, zij.not.mv, ze.inv.mv, zij.inv.mv ze.dbl.mv, zij.dbl.mv, ze.zij.mv
object pronouns	me.obj, mij.obj, je.obj, jou, jullie.obj, hem, m.obje, haar.obj,

	ze.obj.vr, hen, hun.obj, ze.obj
	(ambiguous) u.obj
reflexive and reciprocal	me.ref, mij.ref, je.ref, ons.ref, zich.3, elkaar
	mijn.eigen, je.eigen, uw.eigen, ons.eigen, eigen.3, mekaar
	(ambiguous) u.ref

4 Results and discussion

As a first step in the analysis of the data, we may have a look at each of the lectal dimensions separately. In each case, the question will be to what extent the potential lects (the values on the lectal dimensions) are characterized by a specific bundle of linguistic variables. It is beyond the scope of the present paper to devote a separate discussion to each of the lectal dimensions, so let us have a look a one of the crucial dimensions: register. Given that colloquial Belgian Dutch is an informal variety, we expect a correlation between the linguistic variables that we consider typical for colloquial Belgian Dutch and the more informal registers.

In Figure 2, the register components of the CGN are plotted against the linguistic variables. We can roughly identify two dimensions among the CGN components. Overall, the horizontal dimension of the plot represents a cline from standardized language use on the left, to colloquial speech on the right. In fact, if we look at the distribution of the standardized and the colloquial linguistic variables, in the way in which we identified them above, we can clearly see a dominance of standardized forms on the left and colloquial forms on the right. The individually visible items to the right are definitely colloquial ones, just as the individually visible ones to the right are standardized. Further, if we zoom in on the central area of the plot, spacing out the cloud of overlapping variables as in Figure 3, we again find the same continuum. (Note that the labels of the components are omitted in Figure 3.)

Does this distribution of the linguistic variables correspond with a plausible distribution of the lectal variables? In the upper left corner, o (read aloud written texts), b (interviews with teachers) and h (classroom lessons), form a cluster, and in the lower left hand corner, f (discussions, interviews, debates in the media) and g (political discussions, interviews, debates, particularly as conducted in parliament) constitute a second cluster. As b, h, f, and g may be subsumed under the label of public speech, these components contrast neatly with the other end of the horizontal dimension, where we find a (spontaneous face to face conversations) and c and d (telephone conversations) as examples of informal private speech. Characteristically, b, h, f, and g pattern specifically with the standardized forms of adress: the polite V-forms in the cluster f-g, and the familiar T-forms in the cluster b-h. This also makes clear that we need to distinguish another structural dimension, orthogonal to the horizontal dimension: at the non-colloquial side of the dimension, a distinction needs to be made between situations triggering polite 2nd person pronouns, and situations triggering familiar 2nd person pronouns.

The plots, in other words, confirms the initial assumption that we have to distinguish a set of colloquialisms in contrast with a set of standaridzed forms, and that the distinction between both sets of variables is indeed at least in part a stylistic one. However, the plots also show that the various linguistic indices (and in particular, the colloquialisms) do not all have the same distribution: they do not coincide in one particular point of the plot, as would be the case if had a Rhenish fan-like situation. If the CGN components constitute homogeneous lects in the sense in which the areas between the isoglosses of the Rhenish fan constitute homogeneous dialects,

then each of the CGN components - or specific clusters of such components - would coincide in the plot with specific (bundles of) linguistic variables. In incidental cases that may be the case, like with component g and the polite V-pronoun, but the overall pattern exhibited by the linguistic variables is a continuum, not a set of discrete clusters.

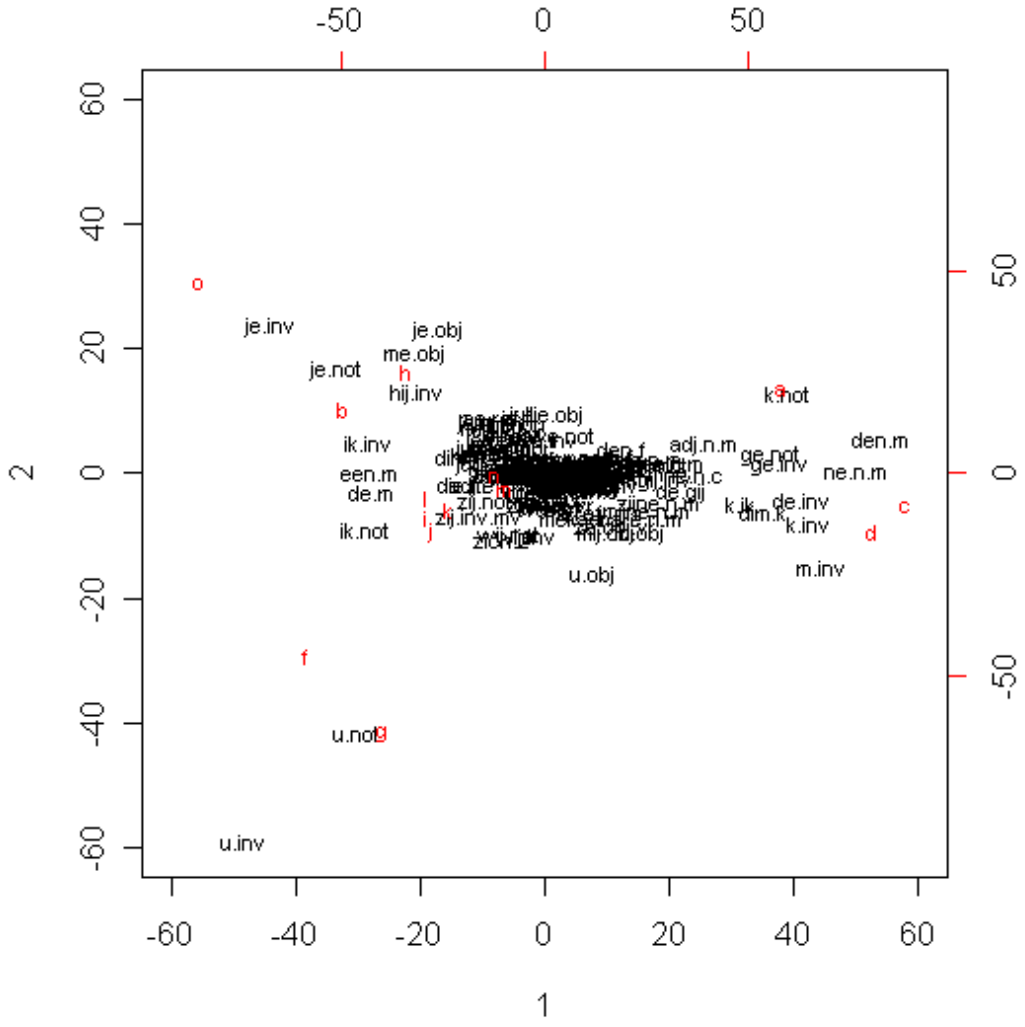

Figure 2
Correspondence analysis of linguistic variables over register

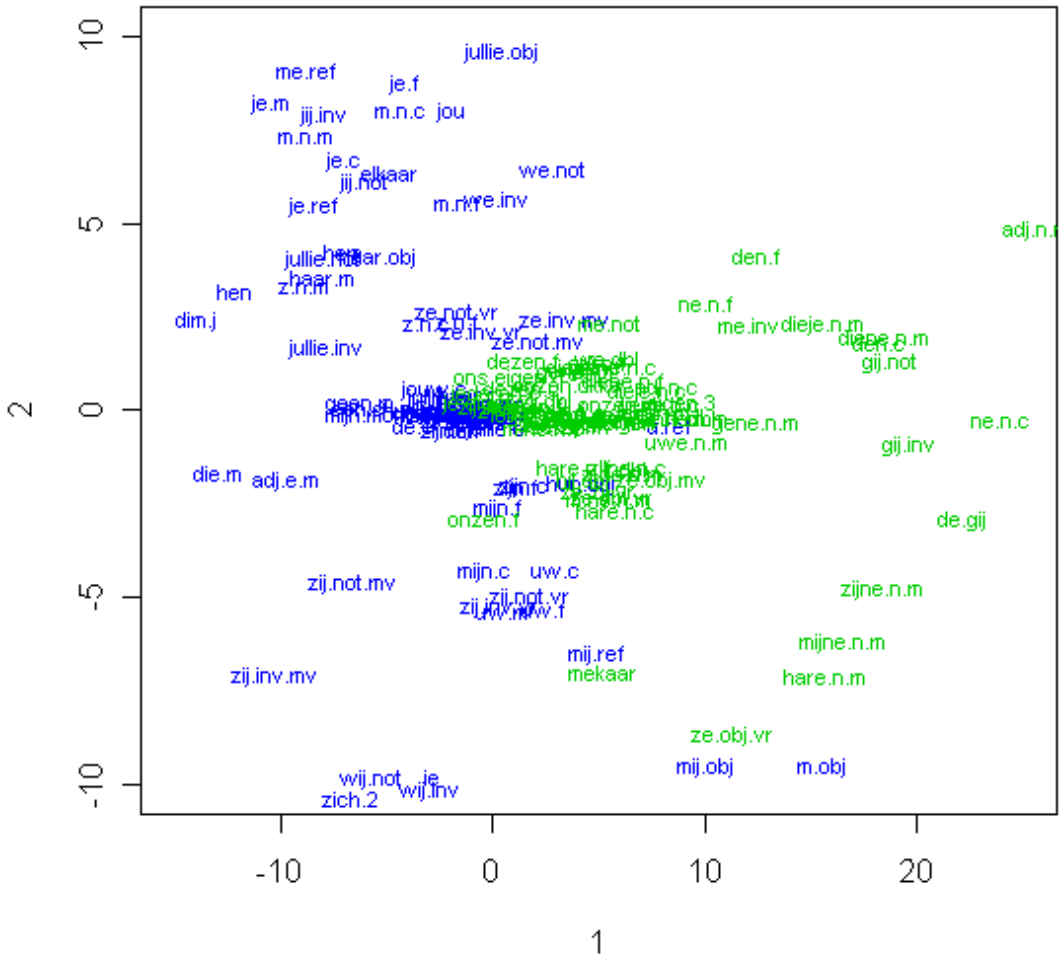

Figure 3
Zooming in on the central part of Figure 2

We may conclude, then, that lects are much more fuzzy than a traditional conception of the notion 'linguistic system' would suggest. If an operational definition of 'lect' is one in which lects are characterized by bundels of co-occurring features, then the CGN components are no lects, and the overall systematicity of the language system is lower than we have been learned to assume. There are three alternative interpretations of the plots that need to be considered and that might lead to a revision of the conclusion that we have just reached.

It is tempting to say, in the first place, that the structural fuzziness we see in the plots is merely 'a statistical effect': after all, even if we think in terms of systems (bundles of co-occurring phenomena) as typical for a certain socio-communicative configuration (a socially distinctive group of speakers, a communicative situation, a combination of both etc.), we could still easily accept that not all the individual language users realize that bundle to the same extent. For instance, if A, B, and C are the specific elements of such a lectally characteristic bundle, and if [x] is the socio-communicative configuration for which it is typical, we could say that the

combination of A, B, C, and [x] defines a lect. But we would not assume that all the usage acts performed in [x] exhibit A, B, and C. If [x] is a local dialect situation, some speakers would still speak the standard language in that situation, i.e. would not realize A, B, and C. In that sense, the situation in [x] would be somewhat fuzzy: we can identify a lect typical for that particular socio-communicative situation, but it would not be selected for use by all speakers appearing in that situation.

However, that would still imply that A, B, and C behave in the same way. The reflection in the plot would be that the bundle A, B, C is slightly more removed from [x] than if situation [x] featured only speakers of the 'pure dialect' A, B, C, but A, B, and C would still occur together in the graphical representation. The situations that we see reflected in the plots we have actually considered are different: although there is a lot of co-occurrence, some linguistic phenomena, if not all, clearly fall outside of any bundle.

But couldn't we then say, in the second place, that the fuzziness we find in the plots results from the fact that speakers may choose to realize a given lect to a certain degree? As in the previous interpretation, the choices language users make would involve the selection of a lect (a bundle of linguistic phenomena), but rather than choosing in a wholesale manner for the lect in question, the choice would be selective: some speakers might realize some elements, while other speakers might realize others. So, instead of the propositions 'all users in situation [x] opt for the bundle A, B, C' or 'some users in situation [x] opt for the bundle A, B, C', we now consider the interpretation 'some users in situation [x] opt to some extent for the bundle A, B, C' (or 'some users in situation [x] opt for some of the elements in the bundle A, B, C').

However, if we take a 'systems first' rather than 'individual elements first' approach, the choice for either A, B, or C as a partial instantiation of the lect would be random: it would amount to the same thing whether you select the combination AB, BC or AC as a selective, partial, mitigated realization of the lect; all combinations would be equally probable as an instantiation of the lect. As such, A, B, and C would again occur together in the graphical representation, because their distance to sociocommunicative situation [x] would be identical. That does not happen, though: we get clear indications in the plots that some elements are more typical for certain sociocommunicative situations than others.

Finally, we could formulate the objection that the plots we have so far considered do not actually live up to the theoretical question that we started off with: by looking at one lectal dimension only - the stylistic dimension of register variation - our lectal analysis is not yet multidimensional, as we initially suggested it should be. Couldn't it be the case that the fuzzy pattern that we find in the data is clarified when we consider not just the CGN components, but all the lectal dimensions that we introduced? Wouldn't we then find more lectal structure if we included more lectal variables?

There are actually two ways of doing this: we can look for broader lectal entities, when two lectal points coincide, or we can have a look at more fine-grained lectal entities, by breaking down the CGN components that we have so far considered.

The first approach is illustrated by Figure 4: we plot a correspondence analysis of all linguistic variables over all lectal variables, and when two lectal variables cluster together, we check whether that clustering could point to a 'super-lect' combining both. In Figure 4, for instance, brab (the Brabant province) and lim (the Limburg province) appear close to each other: from the point of view of colloquial Belgian Dutch, they seem to behave more or less identically, in

contrast with the other regions. Conversely, the position of West Flemish (wvl) as an outlier in the upper left corner is not so surprising if we look at the linguistic variables involved: the *je*-pronouns that surface in the neighbourhood of wvl (but that are also, as we have seen, characteristic of standardized speech) are indigenous to many West Flemish dialects.

In other cases, however, the fact that two lectal variables co-occur does not necessarily point to a combination of both, but rather to an overlapping. The co-occurrence of component h (classroom lessons) and occC (academics) is not surprising: who would do the teaching anyway? In the same way, it is not surprising that component f (interviews) teams up with occH (people working in media and entertainment). The fact that co-occurrences in a plot like Figure 4 may be interpreted in such different ways indicates that it may be necessary to look at combinations of lectal variables in yet another way. And in any case, the overall picture in Figure 4 does not exhibit much more lectal structure than in the previous figures.

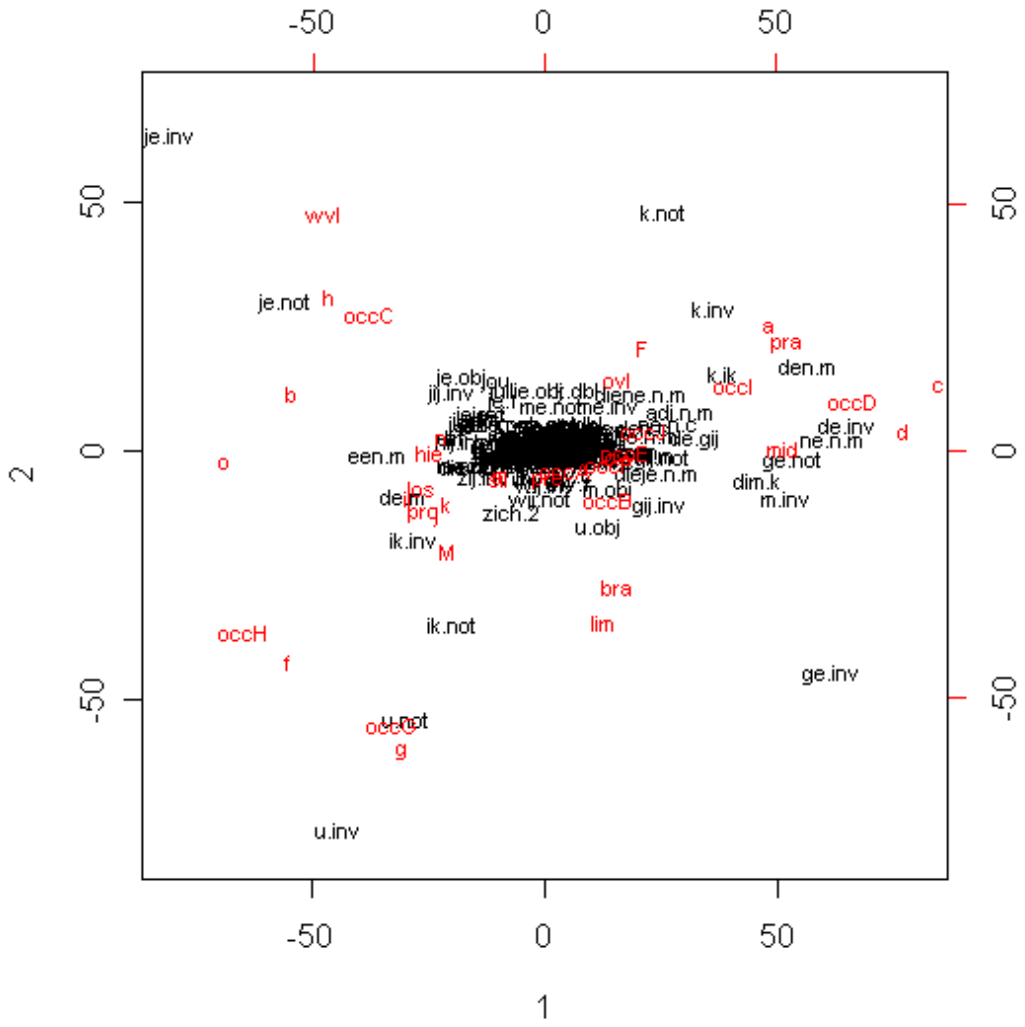

Figure 4
Correspondence analysis of linguistic variables over lectal variables

A different approach, then, is illustrated by Figure 5 and Figure 6. Instead of looking at single lectal points or clusters of such points, we look at the interaction of different lectal dimensions.. For instance, insead of considering each generation separately, we consider generations in combination with another factor. In the figures that serve as an illustration of this approach, we consider the interaction of generation and gender. The lectal point pro.M, for instance, refers to the men in the protest generation. The logic of the argument will be clear: if we start from a more fine-grained lectal architecture by taking into account such interactions of lectal dimensions, does the overall structure of the variation become more systematic?

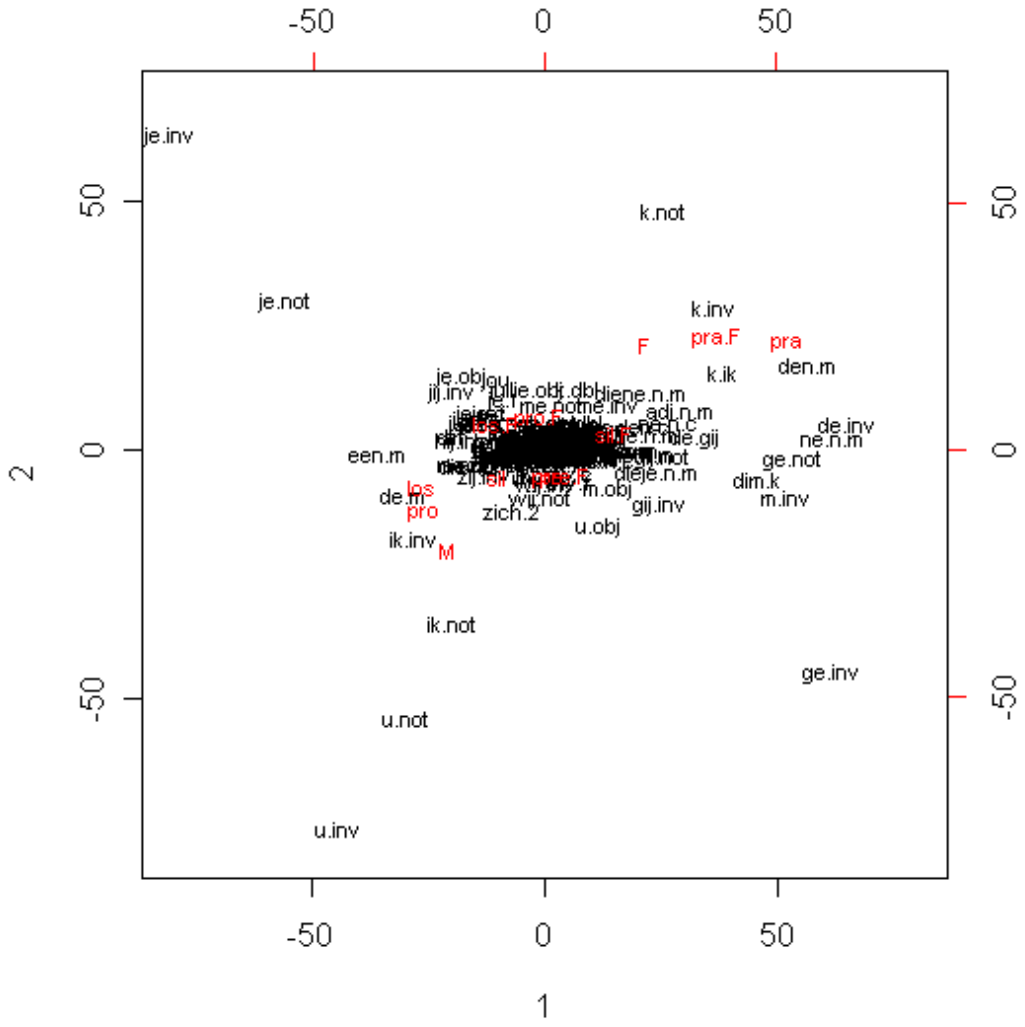

Figure 5
Correspondence analysis of linguistic variables over lectal variables, defined by the interaction of generation and female gender

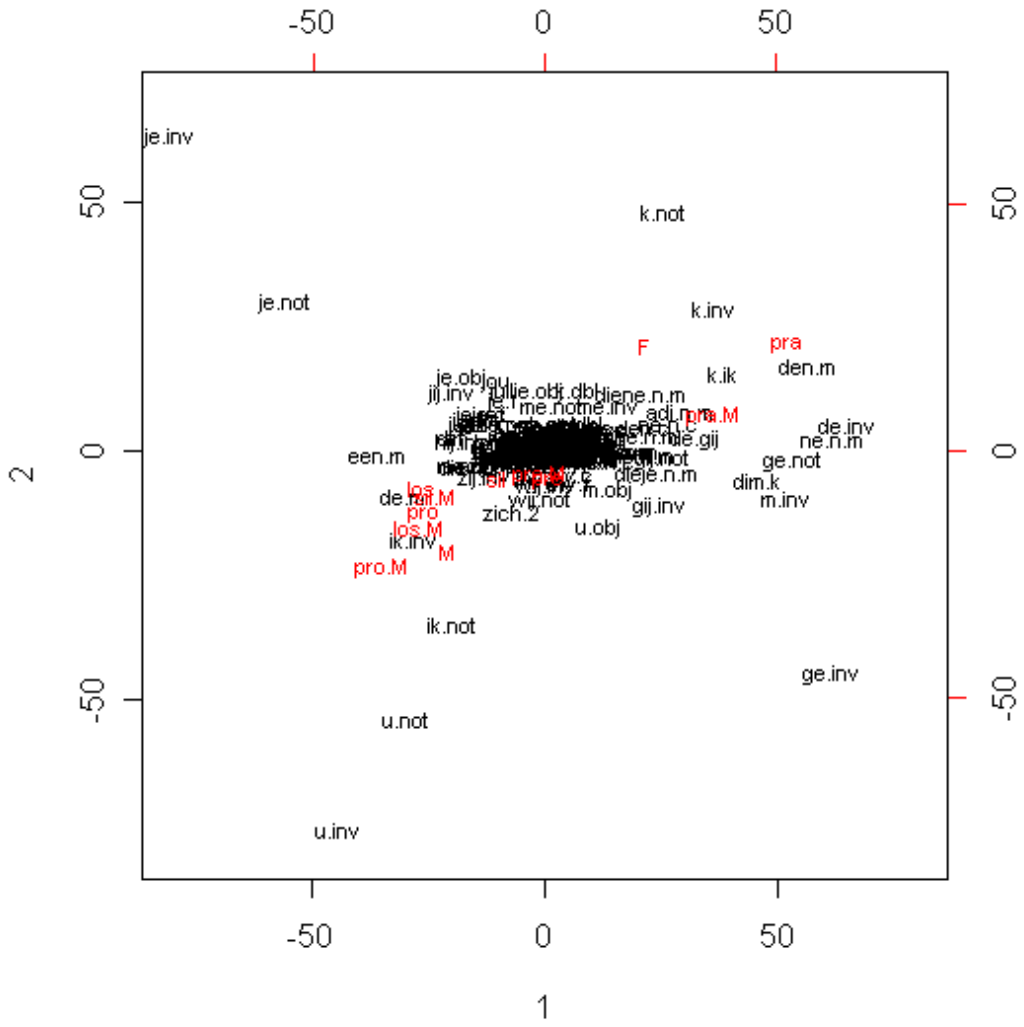

Figure 6
Correspondence analysis of linguistic variables over lectal variables, defined by the interaction of generation and male gender

The figures are not very promising in that respect, however. Some tendencies do become apparent, such as the fact that younger women (pra.F) are situated more to the colloquial right hand side of the figure than younger men, or the fact that men in the protest generation (pro.M) and the lost generation (los.M) are more represented towards the formal, public sphere - the lower left hand side of the figure - than other groups. The overall structure of the plot, however, is as heterogeneous as the others we have seen. Obviously, there are many lectal interactions of this type that could be illustrated, but in all cases (as may be checked in Plevoets 2008), the global picture of heterogeneity remains the same.

5 Conclusion

So, it seems we have to conclude that a language system, once you start looking at it from a usage-based, social point of view, is much less structured than a traditional structuralist conception of language would have us expect. Linguistic phenomena do not occur in clearly distinguishable bundles that correspond in a straightforward way with lectal dimensions, and as such, *the primary unit of lectally structured variation is not language systems, but individual language phenomena*. The lectal structure of the linguistic system is not built up (entirely) from strictly co-occurring phenomena, but rather from phenomena that may each have their own lectal distribution.

Being typical for a certain sociocommunicative situation is a graded matter, but that inevitably implies a difference between the various linguistic forms that we would think of as constituting a lect: describing lects implies describing differences in lectal status between the linguistic phenomena constituting the lect - that is to say: it implies an 'individual elements first' perspective. Or, to put it more simply: to the extent that we wish to talk about lects at all, lects have prototype structure (Kristiansen 2003).

From the point of view of historical linguistics as represented by Schuchardt and Schmidt, and probably also from a common sense perspective, that may not be an entirely surprising conclusion. But from a theoretical perspective, the consequences are important enough. Both from the Saussurean and the Chomskyan tradition, we have learned to think of languages as self-contained systems. But how systematic are the systems? A radical usage-based approach would seem to do away with the notion of system altogether. Arguing against such a radically restrictive attitude, Boye and Harder (2007) sketch a dialectical conception of the language system, which both arises from and interacts with the level of usage: they argue 'that language is indeed based on actual, attested usage, but that it rises above attested instances in providing the speaker not only with actual usage tokens but also with a structured potential that is distilled out of previous usage' (2007: 572). Along the lines of Harder (2003), this potential should be seen as a social fact.

The present paper, then, argues that such a social turn in the conception of the language system has two consequences that are not receiving due attention in Harder's formulation. First, *a social conception of linguistic structure entails a variationist conception of the linguistic system*: instead of one homogeneous system, we have to think in terms of a cluster of lectal systems, each of them partial with regard to what we would normally consider to be 'the' language. Second, these lectal systems themselves are not homogeneous, but consist of linguistic phenomena that may be more or less typical for the lect in question: *lects have prototype structure*.

References

Barlow, Michael and Suzanne Kemmer (eds.). 2000. Usage-based Models of Language. Stanford, Calif.: CSLI Publications.

Becker, Henk. 1992. Generaties en hun kansen. Amsterdam: Meulenhoff.

Benzécri, Jean-Paul. 1992. Correspondence analysis handbook. New York: Dekker.

Boye, Kasper and Peter Harder. 2007. Complement-taking predicates: usage and linguistic structure. Studies in Language 31: 569-606.

Brugmann, Karl and Bertold Delbrück. 1886-1916. Grundriß der vergleichenden Grammatik der indogermanischen Sprachen. Straßburg: Trübner Verlag.

Bybee, Joan L. 2001. Phonology and Language Use. Cambridge: Cambridge University Press.

Bybee, Joan L. 2006. Frequency of Use and the Organization of Language. Oxford: Oxford University Press.

De Sutter, Gert. 2005. Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen. PhD Thesis, KU Leuven.

De Sutter, Gert, Dirk Speelman and Dirk Geeraerts. 2008. Prosodic and syntactic-pragmatic mechanisms of grammatical variation: the impact of a postverbal constituent on the word order in Dutch clause final verb clusters. International Journal of Corpus Linguistics 13: 194-224.

Geeraerts, Dirk. 2001. Everyday language in the media. The case of Belgian Dutch soap series. In Matthias Kammerer, Klaus-Peter Konerding, Andrea Lehr, Angelika Storrer, Caja Thimm and Werner Wolski (eds.), Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik Herbert Ernst Wiegand zum 65. Geburtstag gewidmet 281-291. Berlin / New YorkBerlin / New York: Walter de Gruyter.

Geeraerts, Dirk. 2002. The scope of diachronic onomasiology. In Vilmos Agel, Andreas Gardt, Ulrike Hass-Zumkehr and Thorsten Roelcke (eds.), Das Wort. Seine strukturelle und kulturelle Dimension. Festschrift für Oskar Reichmann zum 65. Geburtstag 29-44. Tübingen: Niemeyer.

Geeraerts, Dirk. 2005. Lectal data and empirical variation in Cognitive Linguistics. In Francisco Ruiz de Mendoza Ibañez and Sandra Peña Cervel (eds.), Cognitive Linguistics. Internal Dynamics and Interdisciplinary Interactions 163-189. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk and Stefan Grondelaers. 1995. Looking back at anger. Cultural traditions and metaphorical patterns. In John Taylor and Robert E. MacLaury (eds.), Language and the Construal of the World 153-180. Berlin/New York: Mouton de Gruyter.

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

Goossens, Jan. 2000. De toekomst van het Nederlands in Vlaanderen. Ons Erfdeel 43: 3-13.

Greenacre, Michael. 1984. Theory and applications of correspondence analysis. London / New York: Academic Press.

Grondelaers, Stefan, Dirk Geeraerts, Dirk Speelman and José Tummers. 2001. Lexical standardisation in internet conversations. Comparing Belgium and The Netherlands. In Josep M. Fontana, Louise McNally, M. Teresa Turell and Enric Vallduví (eds.), Proceedings of the First International Conference on Language Variation in Europe 90-100. Barcelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, Unitat de Investigació de Variació Lingüística.

Grondelaers, Stefan, Dirk Speelman and Dirk Geeraerts. 2008. National variation in the use of er "there". Regional and diachronic constraints on cognitive explanations. In Gitte Kristiansen and René Dirven (eds.), Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems 153-203. Berlin/New York: Mouton de Gruyter.

Harder, Peter. 2003. The status of linguistic facts: Rethinking the relation between cognition, social institution and utterance from a functional point of view. Mind and Language 18: 52-76.

Harder, Peter. Forthcoming. The social turn in Cognitive Linguistics. Berlin/New York: Mouton de Gruyter.

Hopper, Paul. 1998. Emergent grammar. In Michael Tomasello (ed.), The New Psychology of Language: Cognitive and Functional Approaches to Language Structure 155-175. Mahwah, N.J.: Lawrence Erlbaum.

Itkonen, Esa. 2003. What is Language? A Study in the Philosophy of Linguistics. Turku: Åbo Akademis tryckeri.
Kövecses, Zoltán. 2005. Metaphor in Culture: Universality and Variation. Cambridge; New York: Cambridge University Press.
Kristiansen, Gitte. 2003. How to do things with allophones: Linguistic stereotypes as cognitive reference points in social cognition. In René Dirven, Roslyn Frank and Martin Pütz (eds.), Cognitive Models in Language and Thought: Ideologies, Metaphors, and Meanings 69-120. Berlin; New York: Mouton de Gruyter.
Kristiansen, Gitte and René Dirven (eds.). 2008. Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems. Berlin/New York: Mouton de Gruyter.
Oostdijk, Nelleke. 2002. The design of the Spoken Dutch Corpus. In Pam Peters, Peter Collins and Adam Smith (eds.), New Frontiers in Corpus Research 105-112. Amsterdam: Rodopi.
Palmer, Gary B. 1996. Toward a Theory of Cultural Linguistics. Austin, Tex.: University of Texas Press.
Paris, Gaston. 1888. Les parlers de France. Lecture faite à la réunion des sociétés savantes le 26 mai 1888. Revue des patois gallo-romans 2: 161-175.
Plevoets, Koen. 2008. Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische veschijnselen uit het gesproken Nederlands. PhD Thesis, KU Leuven.
Plevoets, Koen, Dirk Speelman and Dirk Geeraerts. 2008. The distribution of T/V pronouns in Netherlandic and Belgian Dutch. In Klaus Schneider and Anne Barron (eds.), Variational Pragmatics 181-209. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Schleicher, August. 1850. Die Sprachen Europas in systematischer Übersicht. Bonn: H.B. König.
Schmidt, Johannes. 1872. Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar: Hermann Böhlau Verlag.
Schuchardt, Hugo. 1885. Über die Junggrammatiker. Gegen die Lautgesetze. Berlin: Oppenheim.
Sinha, Chris. 2007. Cognitive linguistics, psychology and cognitive science. In Dirk Geeraerts and Hubert Cuyckens (eds.), The Oxford Handbook of Cognitive Linguistics 1266-1294. New York: Oxford University Press.
Speelman, Dirk, Stefan Grondelaers and Dirk Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. Computers and the Humanities 37: 317-337.
Tomasello, Michael. 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, Mass.: Harvard University Press.
Tummers, José. 2005. Het naakt(e) adjectief. Kwantitatief-empirisch onderzoek naar de adjectivische buigingsalternantie bij neutra. PhD Thesis, KU Leuven.
Tummers, José, Kris Heylen and Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. Corpus Linguistics and Linguistic Theory 1: 225-261.
Tummers, José, Dirk Speelman and Dirk Geeraerts. 2004. Quantifying semantic effects. The impact of lexical collocations on the inflectional variation of Dutch attributive adjectives. In Gérald Purnelle, Cédrick Fairon and Anne Dister (eds.), Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles 1079-1088. Louvain-la-Neuve: Presses Universitaires de Louvain.
Van Gijsel, Sofie. 2007. Sociovariation in lexical richness. A quantitative, corpus linguistic analysis. PhD Thesis, KU Leuven.
Van Gijsel, Sofie, Dirk Speelman and Dirk Geeraerts. 2008. Style shifting in commercials. International Journal of Pragmatics 40: 205-226.
Verhagen, Arie. 2005. Constructions of Intersubjectivity: Discourse, Syntax, and Cognition. Oxford: Oxford University Press.
Wang, William S.Y. 1969. Competing changes as a cause of residue. Language 45: 9-25.
Wang, William S.Y. (ed.) 1977. The Lexicon in Phonological Change. The Hague: Mouton.
Zlatev, Jordan. 2005. What's in a schema? Bodily mimesis and the grounding of language. In Beate Hampe (ed.), From Perception to Meaning: Image Schemas in Cognitive Linguistics 313-342. Berlin/New York: Mouton de Gruyter.