KOEN PLEVOETS, DIRK SPEELMAN & DIRK GEERAERTS
(LEUVEN)

A corpus-based study of modern colloquial 'Flemish'

1. Background

Historically speaking, Flanders (the Dutch-speaking region of Belgium) did not develop a standard language of its own, but adopted the Dutch standard that already existed in The Netherlands. Due to the strong language policy efforts in the post-war period, this language variety – referred to as the Belgian national standard variety of Dutch – ultimately gained widespread recognition as the common standard for Belgian Dutch (see Jaspaert 1986). Its use, however, remained restricted to the formal and/or written registers, whereas in the colloquial registers, the original Flemish dialects were still being used. This division of labour between standard Belgian Dutch on the one hand and the Flemish dialects on the other witnessed a drastic change from about the mid-1980's onwards, as the use of the dialects for colloquial speech came to be replaced by the so-called 'tussentaal' (literally 'in-between language'). This 'tussentaal' is a supraregional language variety that is highly similar to Belgian standard Dutch in many ways, but that still retains a lot of properties of the – Brabantic – dialects[1]. By consequence, the emergence of the 'tussentaal' can be said to quite typically exemplify a 'standardisation from below'.

2. Research question

Given the intrinsic hybridity of the 'tussentaal' – as the name itself already indicates – the primary question is to what extent does it constitute a uniform language variety. Do the typical characteristics of the 'tussentaal' occur with systematically equal probability, or are some characteristics more frequent and hence more common than others? If the latter is the case, along which dimensions can these differences be accounted for? Previous studies already

[1] Also cf. the contribution by Reinhild Vandekerckhove in this volume

indicated significant differences among the various 'tussentaal'-characteristics on the basis of a single and very specific type of speech situation; see, for example, Van Gijsel 2001 for language in radio and TV advertisements, and Geeraerts 2001 for language in soap series. This paper complements these studies in that it will take several speech situations into account. The objective is to accommodate for the observed (register) variation by looking for some underlying dimensions.

3. Methodology

The methodology will be quantitative and corpus-based. The corpus on which the analysis will be performed is pre-release 5 of the Spoken Dutch Corpus ('Corpus Gesproken Nederlands' – CGN). This corpus is particularly suited for the purpose of this analysis, as it is subdivided into 11 sub-corpora, each sub-corpus containing data from a different type of speech situations. They are the following:

c01: face-to-face conversations
c02: private interviews
c05: public interviews & discussions
c06: discussions, debates & meetings (political)
c07: classroom lectures
c09: sports commentaries
c10: newsreports
c11: (short) news items
c12: prepared commentaries
c13: lectures & speeches
c14: read aloud text

Important to notice with respect to these subcorpora is the fact that they exhibit an inherent structure: Subcorpora 01 to 07, for instance, are types of speech situations that are more dialogic, while from subcorpus 09 onwards the type of speech situations is more monologic.

4. Linguistic variables

The 'tussentaal' involves all sorts of dialectal elements, various aspects of which have already been studied: phonological variation in Van de Velde (1996), and lexicological variation in both Geeraerts, Grondelaers & Speelman (1999) and Grondelaers, Van Aken, Speelman & Geeraerts (2001). For the purpose of this analysis, however, the scope will be narrowed to the

inflectional variation only. The reason for this is the fact that the inflectional characteristics of the 'tussentaal' are commonly considered to be the most predominantly, prototypically 'substandard' ones. These inflectional characteristics, then, can in turn be subdivided into three types: adnominal characteristics, diminutive characteristics, and pronominal characteristics.

The adnominal characteristics of the 'tussentaal' are various determiners and/or attributive elements that are inflected with the dialectal suffix *-e(n)*, in contrast to Belgian standard Dutch where they are not inflected. Table 1 lists a few examples of determiners for both standard Dutch and 'tussentaal', together with a translation:

Standard	Tussentaal	Translation
mijn	*mijn-e(n)*	'my'
elke	*elke-n*	'each'
die	*die-(n)e(n)*	'this'
…	…	…

Table 1: Adnominal variation in Belgian Dutch

The diminutive characteristics of 'tussentaal' also involve a suffixation scheme, but this time in contrast to an existing one in standard Dutch: standard Dutch already has a diminutive system, the so-called J-system; whereas the 'tussentaal' has an alternative one, the K-system. They are listed in Table 2:

Standard	Tussentaal	Translation
bloem-etje	*bloem-eke*	'small flower'
…	…	…

Table 2: Diminutive variation in Belgian Dutch

The most intricate set of 'tussentaal'-characteristics, finally, are the pronominal ones. On the one hand, there are again variants to standard Dutch elements (table 3). They typically occur in post-verbal, enclitic position:

Standard	Tussentaal	Translation
ik	*ekik*	'I'
-ie	*'m*	'he'
…	…	…

Table 3: Pronominal variation in Belgian Dutch

The pronouns of address, on the other hand, even reflect complete alternative (sub-)systems. There exists one system for polite speech, the U-system. For

familiar speech, however, Belgian Dutch speakers can select out of two/three systems, as shown in table 4:

	subject		object
	– inversion	+ inversion	
polite	*u*	*u*	*u*
familiar	*je/jij*	*je/jij*	*je/jou*
	ge/gij	*ge/gij/-de(gij)*	*u*

Table 4: The pronouns of address

Standard Belgian Dutch prescribes the J-system for familiar speech. The alternative, which is therefore often deemed 'substandard', is the G-system (which moreover incorporates a deficient D-system, which is an historical relic from the Flemish dialects; synchronically, however, the distinction between G- and D-system has been blurred). This situation is the outcome of the Belgian language policy: the standard J-system originally belongs to the Northern Dutch dialects as spoken in The Netherlands (and adopted by Flanders). Therefore, it is an exogenic system for Flanders, which renders it necessarily marked for Belgian speakers. The G-system, by contrast, belongs to the Southern Dutch dialects, and consequentially is the endogenic, unmarked system for pronominal address. This paradox concerning the pronouns of address has been frequently commented upon in the literature (see, for example, Vandekerckhove 2004), and will prove particularly interesting, as it will appear in the analysis later on.

5. Analysis

On the basis of the 'tussentaal'-characteristics outlined in the previous section, 80 linguistic variables are selected for the analysis, which are operationalised as the frequency count of one particular form. These do not only contain the substandard forms but also their corresponding counterparts in Belgian standard Dutch. For the analysis at hand, these 80 linguistic variables will be the statistical objects.

The variables for the analysis will be the 11 CGN-subcorpora. As a consequence, the dataset to be analysed is a 80x11-matrix of linguistic objects by register variables, which geometrically amounts to a data-cloud of 80 objects in an 11-dimensional (register) space.

In order to account for the register structure within this data-cloud, the analysis will try to uncover some underlying factors that optimally fit the observed variation (thus constructing some sort of 'colloquialism measure', so

to speak). The technique by which this will be done is the Principal Components Analysis.

Three preliminary transformations are performed on the data-matrix. The first one is the standardisation of the 11 register-variables, i.e. setting both their mean to zero and their variance to one. This is necessary because the 11 sub-corpora are not of equal size, which entails that the frequency counts in one sub-corpus differ from those of another sub-corpus, without this difference, however, being attributable to register variation but merely to sample size. Second, the bare frequency counts are log-transformed. The reason here is to correct for the alleged skewness of word frequencies. The last correction consists in weighting the 80 linguistic objects by their 'surprise-value' log(1/p), the logarithm of the inverse of the relative frequency. This correction is reasonable because some words – like articles, for example – are structurally more frequent than others, these differences again being totally unrelated to register variation. The surprise-value, then, gives the infrequent items a high weight, while a low weight to the more frequent items. This standardised, log-transformed, weighted data-matrix will be the input for the analysis.

The first step in the PCA consists in the computation of the scree-plot, which displays how much of the total variance is explained by each underlying factor (or 'Principal Component'). The scree-plot shown below (Figure 1) points out that the first principal component explains the bulk of all the variance, and that there is not much accumulation in explained variance from the second principal component onwards:

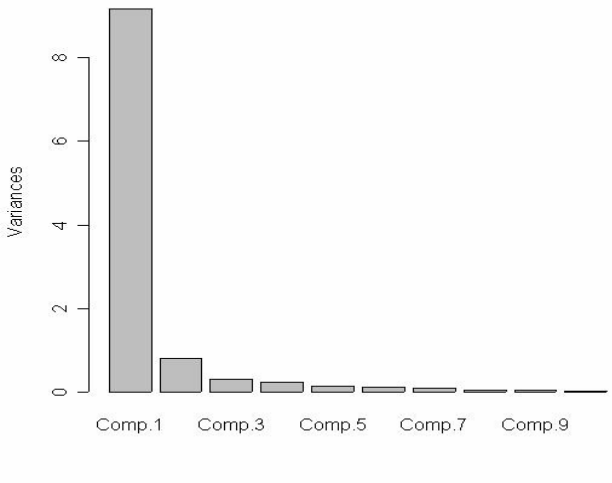

Figure 1: Scree-plot

Consequentially, a structure with the first two principal components can be retained to fit the data.

The next step in the PCA is the computation of the loadings, which are the correlations between the 11 variables on the one hand and the two principal components on the other. Correspondingly, figure 2 shows which of the 11 variables are themselves correlated, and therefore tend to cluster into one register.

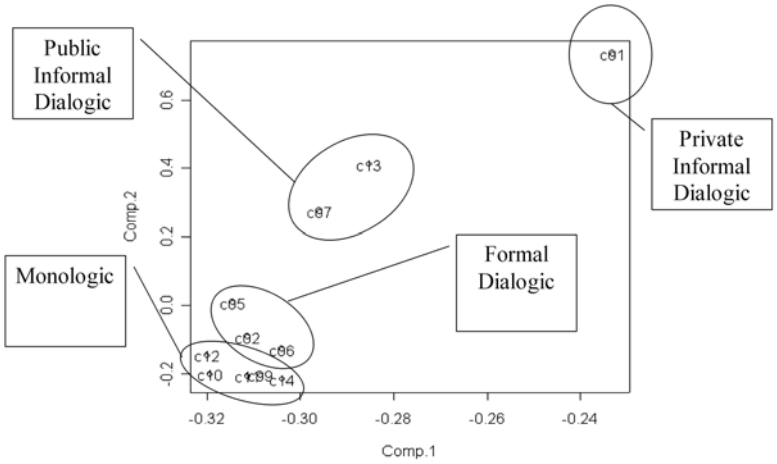

Figure 2: Loadings

Apart from some slight curvature, the loadings display an almost linear trend among the correlations from variables that are negatively correlated with both principal components (bottom-left corner of the plot) to variables that are proportionally more positively correlated with the principal components (top-right corner). As a consequence, the register clusters themselves can be interpreted along these same lines. At the top-right corner of the plot, then, the first subcorpus of the face-to-face conversations (c01) forms a singleton register. More offset, subcorpus 07 of the classroom lectures clusters with subcorpus 13 of the speeches. The difference between this cluster and the face-to-face conversations is that the latter typically consists of two people talking to each other, and hence constitutes a somewhat private type of speech situation. The former cluster, on the other hand, involves speech situations in front of a full and live audience and are therefore more public. A third register is formed by all other subcorpora at the left-hand bottom of the plot. As this cluster contains types of speech situations such as political debates (c06),

and/or newsreels (c11), the – admittedly tentative – interpretation is that these subcorpora reflect formal types of speech situations, whereas the previous two clusters are more informal. After a more thorough inspection of this third cluster, however, a more finegrained structure can be discerned. The three subcorpora 02, 05, and 06 are all at a clear distance from subcorpus 09, 10, 11, 12, and 14, which are in turn completely at the corner of the plot. Aside from subcorpus 13 whose hybrid nature remains unclear for the moment and can therefore only be taken as a plain fact, these are the more monologic speech situations, while the subcorpora in the other three clusters are more dialogic. In conclusion, it can be assumed that there are four registers to be distinguished in Belgian Dutch, which can be discriminated along the oppositions of private-public, informal-formal, and dialogic-monologic; the poles of private, informal, and dialogic furthermore being proportionally more positively correlated with both principal components than their respective counterpoles.

The last step in the PCA is the computation of the scores for the 80 linguistic objects on the space spanned by the two principal components. The first principal component (Figure 3) displays the following structure:

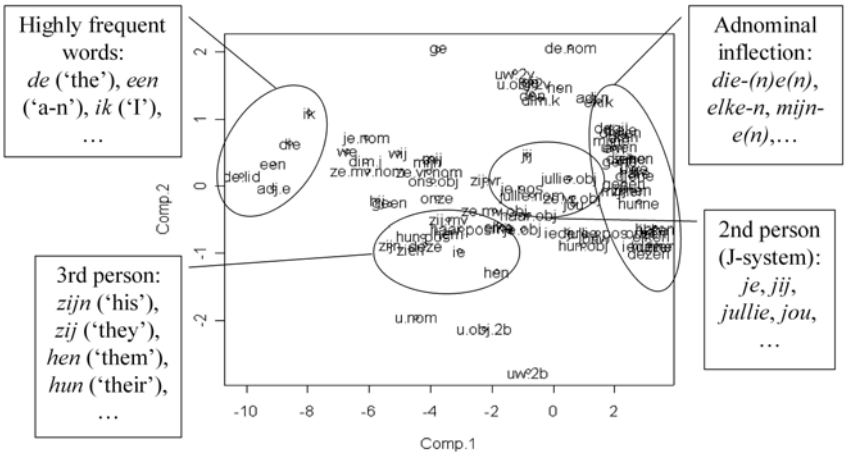

Figure 3: Principal Component 1

On the left-hand side of the plot there is a clearly separate cluster containing articles, the pronoun *I* and so on; in other words, elements that are highly frequent and very common. Somewhat more to the right, there is a cluster with all the third person pronouns. These are still common but slightly less so than the articles. Even more to the right, there is a cluster with 2nd person pronouns, and more particularly those belonging to the J-system. As has

already been mentioned, this is an exogenic and marked system of address for Belgian Dutch, and is therefore not so common in usage. Finally, on the right-hand side of the plot, where the scores become positive, there is the highly dense cluster of the adnominal elements from the 'tussentaal'. Again, these elements are marked and restricted in usage, only this time not because they would be exogenic – on the contrary, they are endogenic – but because they are substandard. Taking into consideration, then, the inherent ambiguity in the use of the concept of 'markedness' – the J-pronouns are marked because they are exogenic, while the adnominal elements are marked because they are substandard – the first principal component can be concluded to form a range from elements that are neutral and common to all registers to elements that are restricted to the more colloquial registers (where the 'tussentaal' is used).

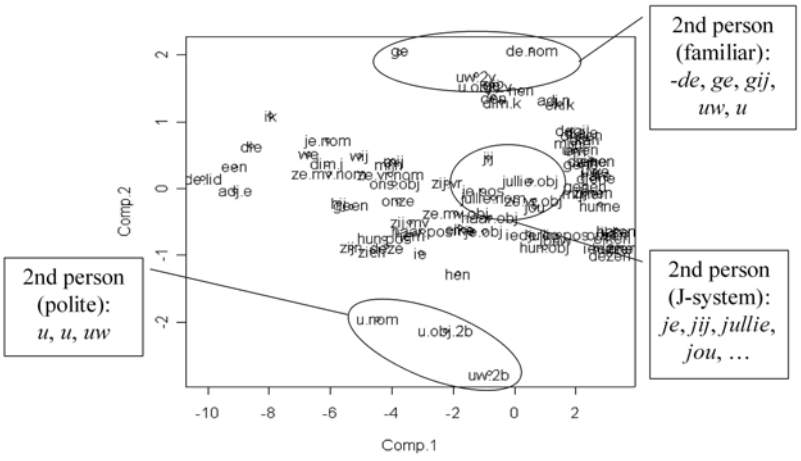

Figure 4: Principal Component 2

Although the second principal component accounts for much less variance (according to the scree-plot), it exhibits some interesting structural differences concerning the pronouns of address that are not visible on the first principal component alone. At the negatively scored bottom of figure 4, for instance, the 2nd person U-pronouns for polite speech are clustered separately from all other variables. Next, the cluster of the J-system has already been pointed out in the middle of the plot. Located at the positively scored top of the plot, finally, are the pronouns of the G-system, which has been stated as the unmarked system of familiar address for Belgian Dutch. The second principal component, then, ranges from unmarked forms for familiar speech (of the 'tussentaal') over marked ones to forms for polite speech. As a consequence, it will become clear now why a geometrical space of two dimensions is retained, and not just of

one. As has been pointed out, the bulk of the register variation can be accomodated for by a range from neutral/common to marked/restricted. On top of that stylistic axis, however, a range of conversational variation between familiar and polite speech can be identified that is not reducible to the stylistic first axis. The interpretation of the two register dimensions – and the location of the 'tussentaal'-characteristics on it (at the top-right corner) – concludes the analysis of the scores.

6. Conclusion.

By means of Principal Components Analysis, two register dimensions can be distinguished that together define four registers for Belgian Dutch. The four registers are to be discriminated along the oppositions private-public, informal-formal, and dialogic-monologic speech, with the poles of private, informal, and dialogic being proportionally more positively correlated with the two principal components than their respective opposite poles (public, formal, and monologic). The scores of the linguistic objects on the two principal components reveal the 'tussentaal'-characteristics to be located at the top-right corner of the space, and hence by association to be typically used in speech situations that are private, informal, and/or dialogic. The horizontal first principal component that accounts for the bulk of the register variation can be interpreted as a stylistic range from words common to all speech situations (left) to words that are restricted to the colloquial ones only (right). On top of that, the vertical second principal component accomodates for the conversational range from polite speech (bottom) to familiar speech (top). As neither type of variation is reducible to the other, it must be concluded that the 'tussentaal'-characteristics do not occur with systematically equal probability over various registers. There are indications, in sum, that the 'tussentaal' is not a uniform language variety.

References

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.* Berlin and New York: de Gruyter, 281–291.

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat.* Amsterdam: Meertens Instituut.

Grondelaers, Stefan, Hilde Van Aken, Dirk Speelman and Dirk Geeraerts. 2001. Inhoudswoorden en preposities als standaardiseringsindicatoren. De diachrone en synchrone status van het Belgische Nederlands. *Nederlandse Taalkunde* 6: 179–202.

Jaspaert, Koen. 1986. *Statuut en structuur van standaardtalig Vlaanderen*. Ph.D. dissertation. Leuven: UP.

Vandekerckhove, Reinild. 2004. Waar zijn je, jij en jou(w) gebleven? Pronominale aanspreekvormen in het gesproken Nederlands van Vlamingen. In Johan de Caluwe, Georges de Schutter, Magda Devos and Jacques van Keymeulen (eds.), *Taeldeman, man van taal, schatbewaarder van de taal.* Gent: Vakgroep Nederlandse Taalkunde-Academia Press, 981–993.

Van de Velde, Hans. 1996. *Variatie en verandering in het gesproken Standaard-Nederlands (1935–1993).* Ph.D. dissertation. Nijmegen: UP.

Van Gijsel, Sofie. 2001. *Taalgebruik in reclame: Een variationele en functionele analyse.* Leuven: unpubl. M.A. thesis.