A Variationist, Corpus Linguistic Analysis of Lexical Richness

Sofie Van Gijsel, Dirk Speelman & Dirk Geeraerts
QLVL, Department of Linguistics
University of Leuven
Sofie.VanGijsel@arts.kuleuven.be, Dirk.Speelman@arts.kuleuven, Dirk.Geeraerts@arts.kuleuven.be

1. Introduction

A number of tests have been developed for measuring the lexical knowledge and use of language learners (see for example Read 2000). Most of these tests focus on child language acquisition or on the extent of vocabulary acquisition of (typically L2) language users from an applied linguistic perspective. Lexical richness measures are thus used to assess the (lexical) proficiency level of the child or student, comparing their lexical richness with an external reference point. Yet, relatively little research has been conducted to investigate the distribution of lexical knowledge from a *sociolinguistic, variationist point of view*. This paper reports on an ongoing PhD project, which attempts to chart the lexical knowledge of adult native speakers, scrutinizing the use of a well-known lexical richness measure, viz. the type-token (TTR; see e.g. Read 2000).

A corpus-driven, quantitative methodology is proposed, analyzing the *CGN* corpus (*Corpus of Spoken Dutch*, Schuurman et al. 2003). This corpus contains linguistic material from two Dutch-speaking communities, viz. The Netherlands and Flanders (the Northern part of Belgium), thus allowing a comparison between these two linguistic societies. Furthermore, the *CGN* is annotated for a number of sociovariational or extralinguistic parameters such as register, educational level and sex. A multivariate analysis will be performed, assessing the effect of these parameters on the distribution of lexical richness. It will be demonstrated that a number of methodological complications have to be taken into account, partly with regard to the somewhat uneven distribution of linguistic material in the corpus, but mostly with regard to the technique for measuring lexical richness. More specifically, it has been shown repeatedly that a simple TTR is text-length dependent (see for example Baayen 2001; Malvern et al. 2004): the longer a text, the smaller the chance that new or different types will be introduced, automatically resulting in lower TTR's for longer texts. In order to reduce the text-length dependency, a stratified sampling method is proposed, dividing the corpus material in equally sized text chunks. A further caveat concerning lexical measures such as the TTR is their thematic dependence (Baayen 2001). In a first attempt to gauge the influence of the topic on the lexical richness measure, an analysis per part-of-speech is performed, testing the difference between TTR's for nouns, which are closely related to the content of the texts, for adjectives and verbs, and for lexically empty function words. Interpreting the results of the multivariate analysis, we will show that lexical richness, as measured by a TTR on text chunks of equal size, is to a very high degree determined by register variation, and we will present some indications that this effect of register may be influenced by the degree of thematic variety in the registers.

The rest of this paper is structured as follows: in Section 2, we briefly discuss some existing lexical richness techniques, focusing on the TTR, and TTR-based measures. In Section 3, the corpus is introduced. Section 4, 5 and 6 discuss the statistical analyses performed. In Section 4, after explaining the sampling method, the results of a global linear analysis are discussed (4.1.). Next, the results for additional multivariate analyses are discussed, zooming in on the different corpus components and dimensions (4.2.). In Section 5, the results for the part-of-speech analysis are given, which will be interpreted as a first key to the content dependency of the lexical measure used. Finally, Section 6 presents the (preliminary) conclusions and indicates further research steps.

2. Measuring lexical richness

As said, lexical richness measures have especially been developed in applied linguistic research. A wide variety of measuring techniques have been proposed, including, for example, lexical density (measuring the amount of content words over the total amount of words in a text; O'Loughlin 1995) or lexical sophistication. The latter starts from the assumption that the more difficult a word is, the less frequent it will be. Thus, lexical sophistication is assessed measuring the proportion of lexical items from a number of frequency bands, which are based on a (typically external) frequency list (e.g. Laufer & Nation 1995). Undoubtedly, the most frequently used lexical richness measure is the type-token ratio (TTR), or a TTR-based measure. Basically, the TTR calculates the number of different words (*types*) over the total number of words (*tokens*) in a text. Yet, in it simplest form, this ratio is highly text length dependent: the longer a text is, the lower the TTR will automatically be (see for example Arnaud 1984). This is a well-known problem, for which a number of possible solutions have been proposed. Interestingly, alternatives for the simple TTR have been a concern both in applied linguistics and in the field of mathematical linguistics. In applied linguistics, adapted measures used include the *Mean Segmental TTR* (MSTTR), as proposed by Engber (1995), where the mean TTR of consecutive text sections of equal length is calculated. Also, a number of transformations have been proposed, such as the *Index of Guiraud*, which measures the amount of types over the square root of the tokens, thus reducing the influence of the token length (e.g. Broeder, Extra & Van Hout 1993). Other TTR transformations include the *Index of Herdan* or *Uber's Index* (see for example Vermeer 2000 for an overview of these measures). A recent measure specifically developed for child language acquisition is the *D-measure* (Malvern et al. 2004), which models the rate at which new words are introduced in increasingly longer text samples, by way of a curve-fitting procedure, which uses one parameter, parameter *D*.

As mentioned, the text-length dependency of the TTR has also been studied in mathematical linguistics, more specifically, in the field of word frequency distribution models. Most notably, Tweedie & Baayen (1998) and especially Baayen (2001) have shown that all the transformations of the TTR proposed so far (including the Indices of Guiraud, Uber and Herdan) are equally text length dependent. As an alternative, Baayen proposes to start from a lexical frequency spectrum, ranking the words in a text according to their frequency of occurrence (viz. the words that occur once, twice, tree times, and so on). To this frequency spectrum, a distribution model is fitted, using

one or more parameters to describe the distribution shape (see also Evert & Baroni, this volume, for a more detailed description of these models).

Since the TTR or a TTR-based measure is the most extensively studied lexical richness measure, we will also scrutinize its usefulness for our purpose. Yet, how should we use this lexical richness measure, and which of the alternatives proposed so far would suit our analysis best? First of all, to take the measures developed in applied linguistics, the 'state of the art' measure seems to be the *D*-parameter. Although some researchers report favourably on the results obtained with this measure (see for example Malvern & Richards 2000 and Silverman & Bernstein Ratner 2002), others are more critical (as for example Jarvis 2002 or Vermeer 2004), showing that the *D*-parameter is not a good alternative for the simple TTR, being equally text length dependent. Further, the *D*-measure typically works on short child language samples, while our aim is to analyse adult mother tongue speech. On the other hand, the mathematical distribution functions developed by Baayen (2001) are not directly applicable, since this research has a different perspective: rather than assessing the vocabulary distribution for a long text or a corpus, attempting to estimate the model parameters to get a fitting distribution function, we would like to be able to directly compare subsamples of one corpus, enabling us to assess the lexical richness of groups of speakers in our corpus. Therefore, at this point of the investigation, we propose to use a fairly simple TTR, which is measured on sampled text chunks of equal token length. It can be noticed that the measure used is somewhat akin to the MSTTR, as equally sized text chunks are analysed. Yet, the MSTTR measure, which is also used in child language acquisition research, works with short language samples, typically containing 30 to 100 tokens. A number of preliminary tests on our corpus materials have shown that short samples (of 150-600 tokens) give less clear results, while measures on the 'range' from 750 up to 1350 tokens perform remarkably better. For longer samples, as of 1500 tokens, the results started to deteriorate again, leading to fewer significances in our statistical models. Therefore, we decided to operationalize our analysis on text chunks of 1350 tokens. A second important difference is that the MSSTR (and, for that matter, most lexical richness analyses in applied linguistics) measures the TTR text-internally, while we attempt to compare sets of texts, organized according to a number of sociovariational dimensions. More details on the sampling method will be given in Section 4; in the next Section, the corpus used is described.

3. The Corpus of Spoken Dutch (CGN)

3.1. Corpus description

The corpus analysed is the *Corpus of Spoken Dutch*, release 1 (*Corpus Gesproken Nederlands* or *CGN*; Schuurman e.a. 2003). This corpus contains 10 million words, 2/3 of which is Dutch spoken in The Netherlands, while 1/3 is Belgian Dutch (as it is spoken in Flanders, the Dutch-speaking, northern part of Belgium). The corpus is structured along 15 register dimensions, ranging from very informal face-to-face conversations (component a) to more formal components, such as lectures and seminars (components m and n) and even read-aloud speech (component o). Furthermore, the corpus is also structured by underlying dimensions, such as spontaneous vs. prepared speech and dialogues vs. monologues. Table 1 gives an overview of the corpus contents. The corpus is also annotated for a number of

extralinguistic factors, three of which are considered here. First, for the factor 'region', we distinguish the central region of the Netherlands (mainly Holland), the rest of the Netherlands, and Flanders. Further, the factor 'sex' and 'educational level' (split up in speakers with and without a higher education degree) are taken into account.

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

Table 1: Overview of the CGN corpus

Since component e (containing business negotiations), only consists of Netherlandic Dutch material, making a comparison between Flanders and The Netherlands impossible, this component was not included in the analysis.

3.2. Corpus sampling

As explained, the lexical richness analysis is performed on equally sized text chunks or 'subcorpora' of 1350 tokens. These subcorpora are sampled for each combination of criteria outlined in 2.1. Thus, for example, one subcorpus could be sampled from component a, spoken by highly educated (eduHigh) men (sex1) in Flanders (regioFl). Ideally, for each of these combinations, five subcorpora would be sampled, resulting in 6750 tokens. Yet, due to the uneven distribution of the corpus, it was not always possible to obtain five 1350 token samples. In total, this sampling method results in 526 subcorpora to be analysed. The following table illustrates the sampling method:

subcorpus	comp	regio	edu	sex	TTR
compaN1eduHighsex1ttr.txt	a	N1	eduHigh	sex1	27.85
compaN1eduHighsex1ttr.txt	a	N1	eduHigh	sex1	30.07
compaN1eduHighsex1ttr.txt	a	N1	eduHigh	sex1	26.59
compaN1eduHighsex1ttr.txt	a	N1	eduHigh	sex1	29.7
compaN1eduHighsex1ttr.txt	a	N1	eduHigh	sex1	30.59
...					
compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	30.74

compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	32.96
compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	28.59
compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	29.41
compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	29.41
...					
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	44.0
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	41.78
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	44.0
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	47.26
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	40.22

Table 2: Illustration of the subcorpora sampled from the CGN corpus

4. Linear Regression analyses

4.1. Global linear regression

As described in the preceding section, the dataset is a stratified sample of subcorpora, each containing 1350 tokens. On this set, containing 526 subcorpora, a multiple linear regression is performed. The dependent variable is the TTR, while the extralinguistic factors, for which the dataset is annotated, function as the independent variables. Thus, the linear model proposed is the following:

$$TTR \sim component + sex + region + eduLevel$$

Table 2 presents the output of the linear regression analysis performed on word forms. This regression analysis and all further statistical analyses described in this paper are implemented using the R package (see http://www.r-project.org/). For the components, which is a factor variable with 14 levels, component a (conversations) is the reference value. For the factor 'sex', 'men' functions as reference value; for 'region', the central region of the Netherlands ('regN1'), is chosen, and for 'eduLevel', the reference value is 'eduHigh' (or speakers with a higher education).

Coeff	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	27.9403	0.4279	65.297	< 2e-16 ***
compb	0.9532	0.6181	1.542	0.12366
compc	-1.5747	0.4924	-3.198	0.00147 **
compd	-1.6772	0.4924	-3.406	0.00071 ***
compf	3.2372	0.5178	6.252	8.59e-10 ***
compg	5.7841	0.5506	10.504	< 2e-16 ***
comph	0.8347	0.5610	1.488	0.13739
compi	5.4131	0.7792	6.947	1.15e-11 ***
compj	7.6465	0.6956	10.993	< 2e-16 ***
compk	16.6581	0.6060	27.491	< 2e-16 ***
compl	11.8009	0.6761	17.454	< 2e-16 ***
compm	7.7570	0.9417	8.238	1.50e-15 ***
compn	6.5075	0.6495	10.019	< 2e-16 ***
compo	12.6548	0.4924	25.702	< 2e-16 ***
regNr	-0.2317	0.2988	-0.775	0.43857
regFl	0.1743	0.2886	0.604	0.54609
eduLow	0.2630	0.2713	0.970	0.33271
women	-0.7928	0.2438	-3.252	0.00122 **

Table 3: Global linear regression model for dataset (analysis based on word forms; n = 526)

First of all, it is important to notice that the global model is highly significant ($p < 0.001$). Also, the R-squared value is 0.82. This value, which measures the proportion of variation in the data that is explained by our model, equally shows that this is a fairly good model. Interpreting the p-values of the different factors in the model, it can be concluded that all CGN components, with the exception of component b (interviews with teachers of Dutch) and component h (classes), are significant with respect to the reference value, which is component a (face-to-face conversations). The estimates (in the second column of the table) show that all significant components have a higher TTR than the face-to-face conversations, with the exception of the telephone dialogues (component c and d). This shows that register variation, as represented by the different corpus components, is a very important factor. Further, the only other extralinguistic factor having a significant effect on the TTR is sex: the model gives a lower TTR to women than to men. Region and educational level are not significant. Finally, it is important to remark that a parallel analysis for lemmas instead of word forms yields very similar results. This was also the case for a number of other tests, not presented here, leading to the conclusion that an analysis on word forms performs equally well for our corpus of adult native speech. In fact, this is in line with what for example Baroni (2005, to be published) notices with regard to the related field of word frequency distributions: plotting the lexical frequency spectrum for both the lemmatized and non-lemmatized BNC corpus, he notices that the distributions are remarkably similar. In the next sections, by default, the results for word forms will be presented.

As mentioned, the results from the linear regression indicate that register variation explains a large part of the TTR variation in the dataset. This dependence of the TTR on the registers is visually demonstrated in Figure 1. This plot shows, on the horizontal axis, the 526 subcorpora, in the order in which they are sampled from the corpus (see also Table 1). Thus, all 1350 token samples of component a are plotted, followed by all subcorpora of component b, c, and so on. On the vertical axis, their respective TTR's are given. The different colors represent the different components (or registers).

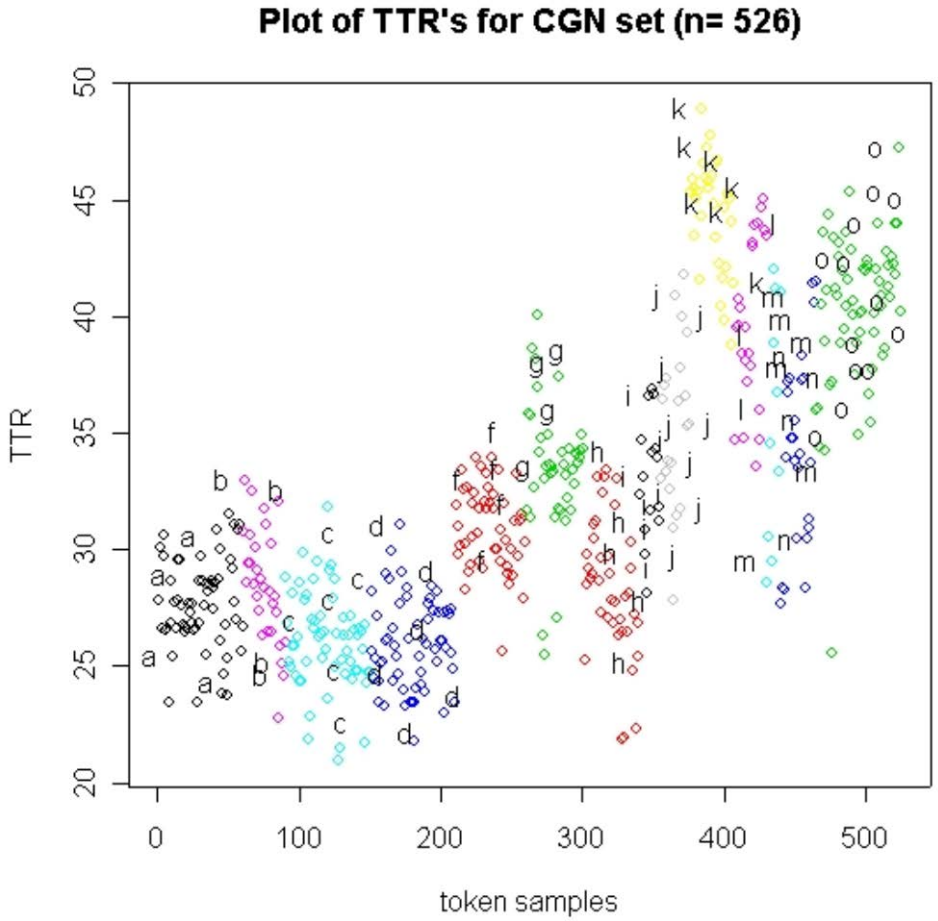

Figure 1: Plot of the TTR's of the CGN set (based on word forms; n = 526)

It is clear that the more informal and conversational components such as a, b, c and d have very low TTR's. Component h (classes) also has low TTR's, although the range is somewhat larger. Tukey tests of significance (which correct for multiple comparisons of the means) show that the TTR's for a, b and h are not significantly different. According to expectations, for the TTR's of component c and d, both consisting of telephone dialogues, a Tukey test equally shows a lack of significance. As mentioned, these components have even lower TTR's than the conversational component a. Although this should be analyzed in more detail, it could be hypothesized that a lower lexical richness in telephone conversations can be explained by a lack of visual interaction between the speakers, which could lead to a more basic use of vocabulary (involving, for example, more repetitions). Also noticeable on the plot is the very high TTR of component k (containing news items), which is indeed significantly higher than any of the other components in a Tukey test. It seems reasonable that news items have a high lexical richness: they consist of formal, well-prepared, mostly monologic speech. It is also possible that the TTR is influenced by the wide range of topics that is typically discussed in news items; a hypothesis which will be explored in more detail in Section 5. Finally, towards the right hand side of the plot, another group of components with high TTR's may be distinguished. For these formal, prepared and monologic registers, Tukey tests show that the TTR's of components m and n (with m containing ceremonial speeches and n containing

lectures and seminars) are not significantly different, while the same goes for component l (columns, reviews, commentaries) and component o (read-aloud speech).

In short, for our corpus, lexical richness, as measured by a TTR on equally sized text chunks, seems to be largely determined by register variation. More informal or conversational components typically have low lexical richness, as measured by the TTR, while the more formal, mostly monologic and prepared registers have high TTR's. The global analysis seems to indicate that the other extralinguistic factors, viz. sex, educational level and region, are not important. In order to find out whether they would nonetheless reveal an effect on the TTR in a more fine-grained analysis, similar linear analyses are performed on each of the CGN components separately, and on the components grouped for the two underlying dimensions (viz. spontaneous vs. prepared and monologues vs. dialogues). The results of these analyses will be discussed in the next Section.

4.2. Linear analysis per component and per dimension

The linear regressions presented here follow the same regression model, although now, the dataset is limited to either one component or to one dimension. The reference values are similar to those in the global linear regression presented above (viz. 'regN1' for region, 'eduHigh' for educational level and 'men' for the factor sex). In Table 3 gives an overview is of the analyses per component, with each row summarizing the results for the regression analysis of the respective component. For each independent factor, the significance is indicated; if the factor is significant, the estimate is also given. It should be remarked that components d, g, i, j, l and o are not listed, since their respective models have p-values higher than 0.05. This indicates that these models do not significantly differ from an intercept only model, which is a model containing only the intercept and none of the independent variables. Consequently, for these models, the p-values and estimates of the independent variables cannot be interpreted.

Component	**regNr**	**regFl**	**eduLow**	**Women**
Compa	n.s.	n.s.	1.07*	-1.45***
Compb	n.s.	-3.61***	/	n.s.
Compc	n.s.	n.s.	n.s.	-1.9***
Compf	n.s.	n.s.	n.s.	n.s.
Comph	n.s.	-2.77*	-2.73*	n.s.
Compk	n.s.	-3.1***	n.s.	n.s.
Compm	/	7.65**	n.s.	n.s.
Compn	n.s.	n.s.	n.s.	n.s.

Table 4: Overview of significances and estimates for the regressions per CGN component

It is clear that again, for the extralinguistic factors, not too many significances are found. Nevertheless, there are some interesting effects. Most noticeably, these stratified analyses permit us to locate the lower TTR given to women in the general model more precisely: the results indicate that it is especially in the most conversational components a and c (viz. conversations and telephone dialogues) that women speech has significantly lower lexical richness. Further research is needed to explain why this is the case. One of the hypotheses could be that women, in

(telephone) conversations, elaborate longer on one subject than men would, resulting in a lower lexical richness for the 1350 token samples. With regard to the factor 'region', there seems to be quite a lot of variation between Flanders (regFl) and the central region of the Netherlands, which is the reference value, while we do not find a single significance between the two Netherlandic regions. Further, it should be said that we also find some rather unexpected results. For instance, a higher TTR for people with a lower education level in conversations (component a) is not immediately expected, although, admittedly, the result is not highly significant. Next, an analysis per dimension is also performed, combining the corpus components in larger groups. For example, as presented in Table 1, the dialogic data are components a-h, while the monologues are component i-o. The results for the analyses per dimension are summarized in Table 4 (the reference values are the same as in previous analyses):

Component	**regNr**	**regFl**	**eduLow**	**Women**
Spontaneous	n.s.	n.s.	-1.5***	-2.02***
Prepared	n.s.	n.s.	n.s	n.s.
Monologues	n.s.	n.s.	2.51**	n.s.
Dialogues	n.s.	n.s.	-0.92*	-1.63***

Table 5: Overview of significances and estimates for the regressions per CGN dimension

First, more variation is found in the spontaneous and the dialogic data than in the prepared and monologic data respectively. More specifically, we find, again, a lower TTR for women in dialogues and spontaneous speech. Also, in these dimensions, speakers with a lower education level have lower TTR's. On the other hand, these effects disappear in prepared or monologic speech, with the exception of the unexpected higher TTR for lower educated people in monologues.

Thus, although both the component and the dimension regressions show some interesting results, a few unexpected effects also appear. Furthermore, the R-squared values of these split-up models are mostly around 0.1, indicating that only a small amount of the variation in the data is explained. Hence, these analyses confirm that the register or component variation is indeed a much more influential factor to explain the lexical richness variation in our dataset. The question arises whether this strong register dependence could be partly attributed to the influence of thematic effects on the TTR. The different registers determine the content or themes discussed in the components. In order to analyse if the variety of themes has an effect on the TTR, a more fine-grained analysis of the different registers is clearly needed. In other words, it should be analysed what it is in the lexical make-up of the different registers that causes these significant TTR differences. In order to explore this question, in a next step, TTR's are calculated per part-of-speech (viz. for nouns, adjectives, verbs and function words separately). The effects of the parts-of-speech on the TTR are then compared, allowing us to examine if, for example, nouns, which are prototypically encoding the theme of the text,undergo more influence of the different components than verbs, adjectives, and especially function words. The analysis and results are described in the next Section.

5. Analysis per Part-of-Speech

For the analysis per part-of-speech (POS), new subcorpora of 1350 tokens are created, selecting only nouns (N), verbs (V), adjectives (A), and function words (Func, grouping interjections, articles, conjunctions, pronouns and prepositions) respectively. The question we would like to answer by analysing the TTR's of these samples is twofold. First of all, are there differences in the TTR's for nouns, adjectives, verbs and function words? If so, what are they? Secondly, it is also important to test whether this effect varies in the different CGN components: is the distribution over the registers similar for the four POS's? Figure 2 plots, on the vertical axis, the average TTR of the different POS's, for each of the corpus components. For component m, which is a very small component, it is not possible to construct a subcorpus of 1350 adjectives.

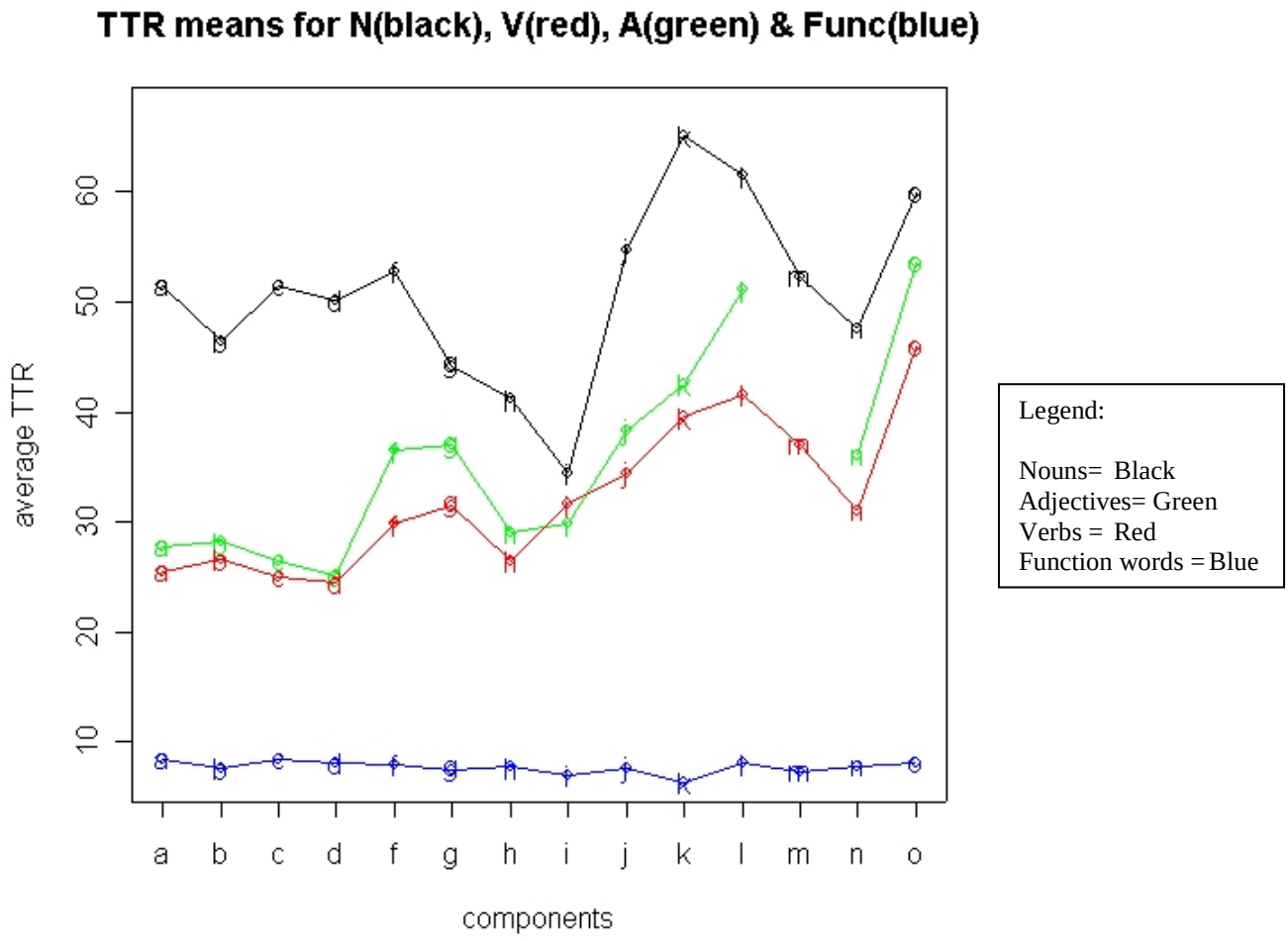

Figure 2: Average TTR of the CGN components per POS (N, V, A and Func)

Comparing the four POS, it becomes clear that the TTR's of the nouns (N) are highest (mean = 51.8), followed by the adjectives (A) and verbs (V), which in turn lie very close to each other (their means are, respectively, 33.48 and 30.3). The function words, on the other hand, behave quite differently, showing very low TTR's (mean = 7. 95). In fact, these results are in line with intuitions: for nouns, speakers can use words of a virtually infinite set of items. Function words, to take the other extreme, form a closed set of words. Adjectives and verbs are also open word classes; yet, the difference with nouns might very well be that they are slightly less dependent on the topic or content of a text. In fact, that brings us to our second question: are the TTR's distributed differently over the components? Looking at the four connecting lines, this indeed appears to be the case. First of all, the oscillating line for the function words shows that there are only small TTR differences over the components. In general, the distribution for the nouns and the adjectives is similar, although for a few components, the distances are larger. Most noticeable is the relatively high average TTR for the adjectives as opposed to the verbs in component l (viz. 51.04 vs. 41.6; $p > 0.05$). In this component, which contains 'evaluative informative texts' (viz. commentaries, reviews, and columns), the high lexical richness characteristic for informative texts is supplemented with typically evaluative and subjective adjectives

(such as *prachtig* ‘gorgeous’, *afschuwelijk* ‘terrible’ or *bijzonder* ‘exceptional’). Finally, for the nouns, the TTR differences between the components are quite large, also in comparison with the curves for adjectives and verbs. A clear example is component i, which contains sport commentaries. While the difference between N’s on the one hand and A’s and V’s on the other is always significant (in Welsh two-sample t-tests), this is not the case for this component. It is interesting that the curves precisely approximate each other in the case of sport commentaries, which typically form a fairly restricted register, with a limited range of topics. This clearly has a strong effect on the behaviour of the nouns, while it affects the adjectives and verbs to a lesser extent, not to mention the function words. This confirms that it might well be the case that nouns are highly sensitive to the thematic content of the components.

To further investigate the behaviour of the POS’s over the components, separate plots per POS are given in Figure 3. On the vertical axis, the TTR of each of the subcorpora (viz. the 1350 token samples) are plotted. Again, different corpus components are represented by different colours.

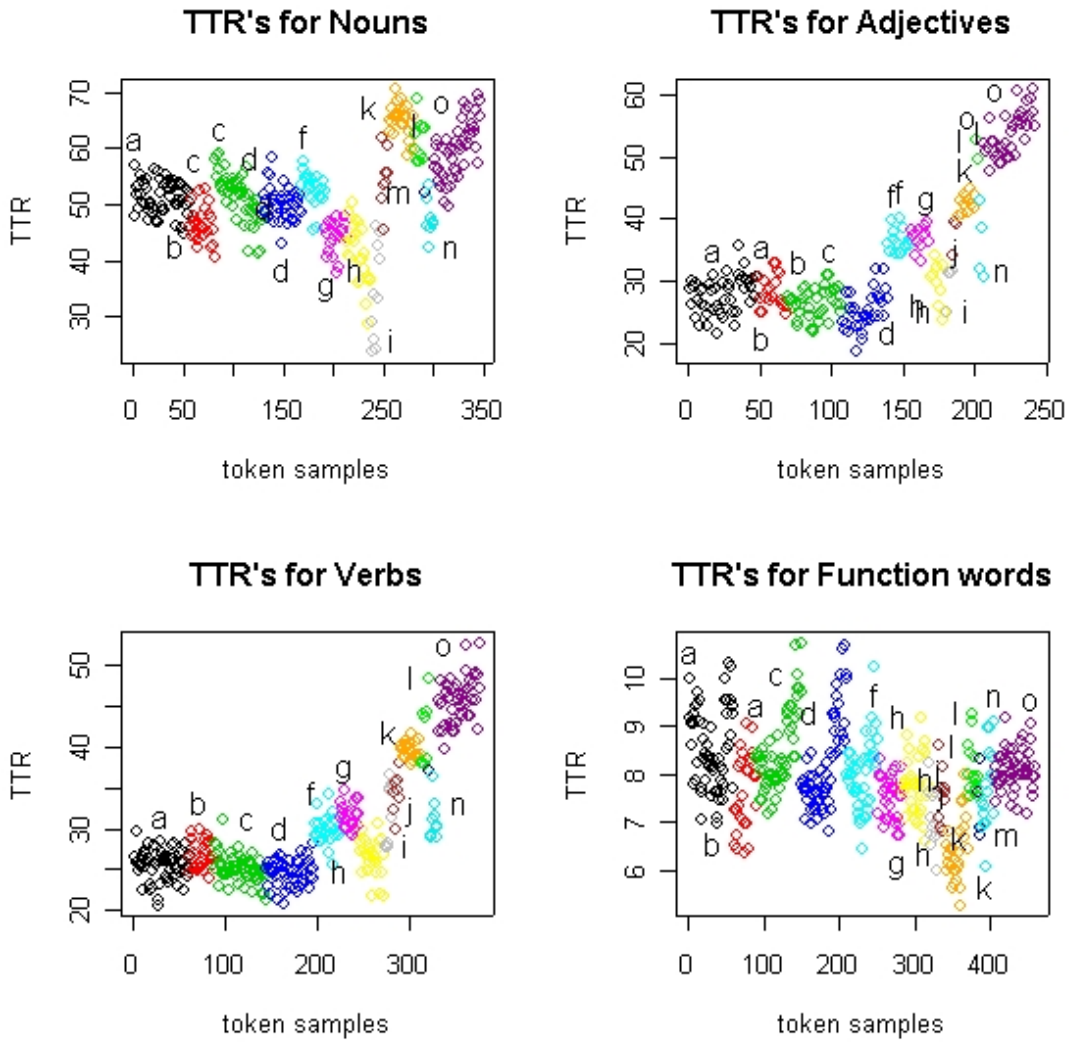

Figure 3: TTR plots for the subcorpora of the four POS

It is clear that the distribution for the adjectives and the verbs is indeed very similar. Also, they are very much alike to the distribution shown in Figure 1, which plotted the global TTR's, for the basic text samples used in the global linear regression. For the nouns, a different image emerges, although this is largely due to two components, viz. i and k (sport commentaries and news items respectively). The explanation given for the lower TTR's for sport commentaries can actually be turned around for component k: contrary to sport commentaries, news items typically cover a broad range of topics, which all acquire a (partially) different vocabulary, giving rise to high TTR's. The plot for the function words shows, as expected, an entirely different, more scattered pattern. Although this plot does not seem to demonstrate any structure at all, there are some significant differences between the components. Most noticeable is, again, component k, which has a significantly lower TTR ($p < 0.01$, with regard to the reference value, component a). It makes sense that news items, which are well-prepared and highly informational, contain a smaller array of function words, such as, for example, interjections. Again, this should be analysed in more detail in further research.

In conclusion, the lexical richness for the N-samples seems more heavily influenced by the thematic content, inherent in the different registers, than the adjectives and verbs, and especially function words. Although this is only a first indication, which requires further analysis, this does show that the dependence of the lexical richness measure on the content of the corpus cannot be underestimated.

6. Conclusions & further research steps

The results presented show that the TTR, performed on carefully sampled corpus subcorpora of equal length, gives consistent results for the corpus under analysis. While the measure used is not highly sophisticated, it does allow us to gauge the influence of a number of sociovariational factors on our corpus data. More specifically, it was demonstrated that register differences, encoded in the different components of the CGN corpus, are the most important factor in explaining the lexical richness differences in the corpus. Components containing more informal, dialogic and/or spontaneous speech typically have lower TTR's than formal, monologic and/or prepared speech. Although the other extralinguistic parameters under analysis were clearly less important, a consistently lower TTR for women was found, both in the general analysis and in split-up models for conversational, informal dialogues. Further, there are indications of a lower TTR for speakers with no higher education, and of more variation between the Netherlands and Flanders than between the two Netherlandic regions under analysis.

Secondly, it was also shown that there are interesting deviations in lexical richness between nouns, adjectives and verbs, and function words. Not only are there systematic differences in the TTR scores, but also, the distribution over the corpus components is different. While the nouns, which typically have high TTR's, are clearly affected by the different registers, this is less so for adjectives and verbs. The subcorpora of function words, which have the lowest lexical richness, seem less dependent on register variation, although there are some significant effects. This variation between the POS is a first indication that the lexical richness measure is influenced by the topics encoded in the different registers or components of the corpus.

In further research steps, the analyses presented here will be replicated on a corpus of written text. Given the strong influence of the registers on lexical richness, it would be interesting to compare the results for a spoken and a written register. A good candidate for this analysis would be the *Condiv* corpus (Grondelaers et al. 2000). This corpus of written Dutch is not lemmatised, but since our results show that the analyses performed on word forms perform equally well, this will not hinder the analysis. Secondly, it would also be interesting to crossvalidate the results obtained here with text-internal lexical richness measures, as used in applied linguistics, or with vocabulary distribution measures such as proposed by Baayen (2001) or Baroni & Evert (2005, this volume). Further, more research is needed with regard to the thematic bias of the TTR, which is especially shown in the register dependency of the TTR's of the noun subcorpora. In fact, an independent measure of thematic multiplicity would help to assess the extent to which the lexical richness measure used is affected by the content of the texts. Such a thematic measure could make use of a keywords method or a LSA-based method. Finally, a detailed content or discourse analysis could complement the present analysis, in order to gain insight in what factors, besides the thematic variety, influence the register variation, allowing, for example, for a more in depth analysis of dialogic vs. monologic texts.

References

Arnaud, P. J. L. (1984) The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests, in T. Culhane, C. Klein Bradley, & D. K. Stevenson (eds.) *Practice and problems in language testing: papers from the International Symposium on Language Testing* (Colchester: University of Essex), 14-28.

Baayen, H. (2001) *Word Frequency Distributions* (Dordrecht: Kluwer Academic Publishers).

Baroni, M. (To appear) Distributions in Texts, in Anke Lüdeling and Merja Kytö (eds.) *Corpus linguistics: An international handbook* (Berlin: Mouton de Gruyter). Available online at http://sslmit.unibo.it/~baroni/publications/hsk_39_dist_rev1.pdf (accessed July 30th, 2005).

Broeder, P., Extra G. and Van Hout, R. (1993) Richness and Variety in the Developing Lexicon, in C. Perdue (ed.) *Adult language acquisition: cross-linguistic perspectives, Vol. I: Field methods* (Cambridge: Cambridge University Press), 145-232.

Engber, C. (1995) The relationship of lexical proficiency to the quality of ESL compositions. *Journal of L2 Writing 4*, *2* 138-155.

Evert, S. and Baroni, M. (2005) Testing the extrapolation quality of word frequency models. *Proceedings of Corpus Linguistics 20*05, Birmingham, UK. Available on-line from http://purl.org/stefan.evert/PUB/EvertBaroni2005.pdf (accessed July 25th, 2005).

Grondelaers, S., Deygers, K., van Aken, H, van den Heede, V. and Speelman, D. (2000) Het ConDiv-corpus geschreven Nederlands. *Nederlandse Taalkunde 5*, 356-363.

Jarvis, S. (2002) Short texts, best fitting curves, and new measures of lexical diversity. *Language Testing, 19*: 1-15.

Laufer, B. and Nation, P. (1995) Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics 16*: 307-322.

Malvern, D. and Richards, B. (2000) Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language* Testing *19*: 85-104.

Malvern, D., Richards, B., Chipere, N. and Durán, P. (2004) *Lexical Diversity and Language Development: Quantification and Assessment* (Palgrave Macmillan).

O'Loughlin, K. (1995) Lexical density in candidate output on direct and semi-indirect versions of an oral proficiency test. *Language Testing 12*, 2: 217-237.

Read, J. (2000) *Assessing vocabulary* (Cambridge: Cambridge University Press).

Schuurman, I., Schouppe, M., Hoekstra, H. and van der Wouden, T. (2003) CGN, an annotated corpus of spoken Dutch. *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, Budapest*, 101-108.

Silverman, S. and Bernstein Ratner, N. (2002) Measuring lexical diversity in children who stutter: application of vocd. *Journal of Fluency Disorders, 27*: 289-304.

Tweedie, F.J. and Baayen, R.H. (1998) How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32*: 323–352.

Vermeer, A. (2000) Coming to grips with lexical richness in spontaneous speech data. *Language Testing 17*, *1*: 65 - 84.

Vermeer, A. (2004) The Relation between Lexical Richness and Vocabulary Size in Dutch L1 and L2 Children, in P. Bogaards & B. Laufer (eds.) *Vocabulary in a Second Language: Selection, Acquisition and Testing.* (Amsterdam: John Benjamins), 173-189.