Locating lexical richness : a corpus linguistic, sociovariational analysis

Sofie Van Gijsel[1], Dirk Speelman[1] & Dirk Geeraerts[1]

[1]KU Leuven - RU Quantitative Lexicology & Variational Linguistics (QLVL)

Blijde Inkomststraat 21 - 3000 Leuven – Belgium

Sofie.VanGijsel@arts.kuleuven.be, Dirk.Speelman@arts.kuleuven, Dirk.Geeraerts@arts.kuleuven.be

Abstract

This article discusses a quantitative analysis of the distribution of lexical richness, using a corpus of spoken Dutch (*CGN*, Schuurman et al., 2003). Lexical richness measures have been a concern both in applied linguistics (see e.g. Read, 2000) and in the context of word frequency distributions (Baayen, 2001). In applied linguistics, lexical tests mostly focus on child language acquisition or on the extent of vocabulary acquisition of (typically L2) language users, while word frequency distributions statistically model the vocabulary in (a collection of) longer texts. Yet, relatively little research has been conducted that specifically attempts to investigate the distribution of vocabulary use from a *sociolinguistic*, *variationist* perspective, integrating the effect of extralinguistic parameters, such as register or region, in a multivariate analysis. For this analysis, the type-token ratio's (TTR's) of equally sized and relatively small texts chunks, sampled from the different corpus components, are analyzed. It will be shown that the *register* of the texts strongly determines their lexical richness, while the effect of the other factors analyzed (viz. 'sex', 'educational level' and 'region') is less strong. Therefore, a more in-depth analysis of the different registers is proposed. Analyses of subsamples per *part of speech* reveal significant differences in the distribution of the TTR's over the registers, especially for nouns. The results point at the influence of thematic and stylistic effects on the lexical richness of a text. It will be argued that it is necessary to locate these effects, in order to allow for a more fine-grained analysis of the sociovariational distribution of lexical richness.

Keywords : lexical richness, corpus linguistics, variationist linguistics, Dutch

1. Introduction

This paper discusses an analysis of the sociovariational distribution of lexical richness. A corpus linguistic, quantitative methodology is proposed, analyzing the CGN (*Corpus of Spoken Dutch*, Schuurman et. al., 2003). A number of lexical richness analyses have already been suggested, both in applied linguistics and in the field of word frequency distributions. Yet, the present analysis takes a rather different perspective, in that it aims to assess the effect of a number of extralinguistic parameters (such as 'region' or 'register') on lexical richness. A stratified corpus sample of equally sized text chunks is constructed, and the lexical richness of the samples is evaluated using a type-token ratio measure (TTR). A multivariate analysis is then performed to assess the effect of a number of extralinguistic parameters on the TTR. This analysis reveals that the lexical richness of the texts is strongly determined by the factor

[1] The research was supported by a Ph.D. grant (aspirant) for Sofie Van Gijsel from the Fund for Scientific Research–Flanders.

'register'. In order to explore this effect, the lexical make-up of the different registers has to be analyzed in greater detail. For the different registers, subsamples per part of speech are constructed, the analyses of which give indications for a more precise location of the lexical richness effects attested.

This paper is structured as follows : in the next section, a brief overview of lexical richness research is given, pointing at differences with the present study. In Section 2, the corpus and the sampling method are discussed in more detail. Section 3, 4 and 5 present and discuss the statistical analyses performed. In Section 3, the results of a global linear analysis are discussed (3.1.). Next, the results for additional multivariate analyses are presented, zooming in on the different corpus registers and dimensions (3.2.). In Section 4, for the different registers, analyses on additional subcorpora per part of speech are discussed. Finally, Section 5 presents the (preliminary) conclusions and indicates further research steps.

2. Measuring lexical richness

As mentioned, lexical richness has already been studied both in applied linguistics and in the context of word frequency distributions. In *applied linguistics*, a number of tests have been developed for measuring the lexical usage of children or second language learners (for an overview, see for example Read, 2000). Lexical richness measures are used to assess the lexical proficiency level of a child or student, comparing their lexical richness with an external reference point. The most widespread tests are based on the concept of *vocabulary diversity*, which is evaluated using a 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, this ratio is highly text length dependent : the longer a text is, the lower the TTR will automatically be (see for example Arnaud, 1984). Therefore, a number of adjustments have been proposed, including the *Mean Segmental TTR* (MSTTR), as proposed by Engber (1995), which calculates the mean TTR of consecutive text sections of equal length. Other transformations, which all attempt to reduce the influence of the token size, include the *Index of Guiraud*, the *Index of Herdan* and *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, dubbed parameter *D*. Although some researchers report favorably 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.

In the context of *word frequency distributions*, lexical richness has also been studied. Interestingly, Tweedie & Baayen (1998) and especially Baayen (2001) have shown that all mathematical TTR transformations proposed (including the Index of Guiraud, Herdan and Uber) are text length dependent. More specifically, they are unable to capture the specific structure of the lexicon, which is characterized by a *Large Number of Rare Events* (LNRE) : while a small number of words is very frequent, the majority of words occurs only a few times, even in large token samples. 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, 2005, for a more detailed description of these models).

While both strands of research raise interesting issues and caveats with respect to analyzing lexical richness, the analysis discussed in this article takes a different perspective. As said, we will attempt to chart the distribution of lexical richness from a *sociolinguistic*, *variationist* point of view. Thus, unlike research in applied linguistics, the lexical variety in adult mother tongue language is investigated. Also, instead of determining the lexical richness of a language user (e.g. a student of English) *text-internally*, *sets* of texts are compared, according to a number of extralinguistic characteristics. On the other hand, it is also not attempted to statistically model the word frequency distribution in the corpus. Rather, we would like to *directly compare* subsamples of the corpus, attempting to locate differences in lexical richness between sociolinguistically delineated groups of speakers. Therefore, at this point of the investigation, a simple TTR-measure is proposed. In order to reduce the text-length dependency, the TTR's of equally sized corpus samples, which are picked according to the extralinguistic parameters selected, are calculated. After a number of preliminary tests with different token lengths, we decided to operationalize our analysis on text chunks of 1350 tokens.[2] A more detailed description of the sampling method is given in 3.2.

3. The Corpus of Spoken Dutch (CGN)

3.1. Corpus description

The corpus analyzed is the *Corpus of Spoken Dutch*, release 1 (*Corpus Gesproken Nederlands* or *CGN* ; Schuurman et al., 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). Additionally, the corpus is structured by underlying dimensions, viz. 'spontaneous vs. prepared speech' and 'dialogues vs. monologues'. Table 1 gives an overview of the corpus contents. The corpus is further 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 factors 'sex' and 'educational level' (split up in speakers with and without a higher educational degree) are taken into account. An overview of the corpus is given in Table 1 :

Comp	**Description**	**spont vs. prep**	**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 with local interface)	spont	dial
f	Interviews/ discussions/debates (broadcast)	prep	dial

[2] It should be noticed that the measure used is somewhat akin to the MSTTR, as equally sized text chunks are analyzed. Yet, the MSTTR measure, as employed in child language acquisition research, is used on short language samples, typically containing 30 to 100 tokens. Secondly and more importantly, the MSTTR measures the TTR text-internally, while in this study, it is attempted to compare sets of texts, organized according to a number of sociovariational dimensions.

g	(Political) discussions/debates/ meetings (non-broadcast)	spont	dial
h	Lessons recorded in the classroom	spont	dial
i	Live 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[3]

3.2. Corpus sampling

As explained above, 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 higher educated (eduHigh) men (sex1) in Flanders (regioFl). In order to have a fairly sized set of subcorpora, the aim was to sample five subcorpora for each of these combinations. Yet, due to the uneven make-up of the corpus, it was not always possible to obtain five 1350 token samples. In total, a set of 526 subcorpora was constructed.[4] 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
...					
compbN1eduHighsex1ttr.txt	b	N1	eduHigh	sex1	30.74
...					
compoFleduLowsex2ttr.txt	o	vl	eduLow	sex2	40.22

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

[3] 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.
[4] The analysis presented here is performed on word forms. Yet, a parallel analysis on lemma's yielded 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. This is in line with the similarities between word frequency distributions for lemma's and word forms, described by Baroni (2005, to be published). All further analyses in this paper are based on word forms.

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, consisting of 526 subcorpora, a multiple linear regression is performed. The dependent variable is the TTR, while the extralinguistic factors selected function as independent variables. Thus, the linear model proposed is the following :

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

Table 3 presents the output of the linear regression analysis. This regression analysis and all further statistical analyses described in this paper are implemented using the R package.[5] 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).

Components	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 the dataset (analysis based on word forms ; n = 526)[6]

The results indicate that the statistical model is very significant ($p < 0.001$) and that a large proportion of the variation is explained (R-squared = 0.82). The p-values of the different factors in the model indicate that all CGN components, with the exception of component b (interviews with teachers of Dutch) and h (classes), are significant with respect to the reference value, which is component a (face-to-face conversations). With regard to the other extralinguistic factors, we only find a significant effect for 'sex', with a lower TTR for women, as the estimate indicates. 'Region' and 'educational level' are not significant. Apparently, *register variation*, as represented by the different corpus components, is a determining factor for the TTR of the samples. Interpreting the estimates of the different components, it is clear 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). It

[5] The R package for statistical computing and graphics is freely available from http://www.r-project.org.
[6] The significance codes are the following: $p < 0.001$: ***, $p < 0.01$: ** and $p < 0.05$: *.

could be hypothesized that the 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). In general, the more formal, prepared components, such as for example news items or speeches (component k and m), have higher TTR's than the informal, dialogic ones, such as conversations and casual interviews with Dutch teachers (component a and b).

The question arises whether an effect of the other extralinguistic factors (region and educational level) can be found in a more fine-grained analysis of the different registers. Also, it would be interesting to locate the effect of the factor 'sex' more precisely. Therefore, additional 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 are discussed in the next Section.

4.2. Linear analysis per component and per dimension

The linear regressions presented here follow the same regression model as the global analysis, although now, the dataset is limited to either one dimension or to one component.[7] The reference values are similar to those in the global linear regression (viz. 'regN1' for region, 'eduHigh' for educational level and 'men' for the factor sex). Table 4 gives an overview 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 were not significant.[8]

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

These analyses reveal a couple of interesting effects. Most noticeably, these 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 a 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. It is not surprising that the difference

[7] Hence, the model tested was *TTR ~ region + eduLevel + sex*.

[8] For these models, $p > 0.05$, which means that they do not significantly differ from an *intercept only 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.

between the national varieties of Dutch also entails lexical richness effects, which do not exist *within* The Netherlands. Yet, 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 very significant.

An analysis per underlying *dimension* is also performed, combining the corpus components into larger groups. The results, which are not presented here, show that there are more effects of the extralinguistic factors on the TTR in the *spontaneous* and the *dialogic* data than in the prepared and monologic data respectively. More specifically, we again find a lower TTR for women in dialogues and spontaneous speech. Speakers without a higher education also have lower TTR's. These significances disappear in prepared and monologic speech, with the exception of a (rather unexpected) higher TTR for lower educated people in monologues.[9]

In conclusion, both the component and the dimension regressions show interesting results, although we do find some unexpected effects too. 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 determines a large proportion of the lexical richness variation in our dataset. The question arises what it is in the *lexical make-up* of the different registers that causes these significant TTR differences. In fact, it could be hypothesized that *thematic effects* play a role here : the register determines to a large extent the *(variety of the) content or themes* discussed in the texts discussed. 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 aim is to examine whether the TTR's of subcorpora consisting of nouns, which are prototypical content-encoders, differ from the TTR of verbs, adjectives, and especially function words, and, more importantly, if these differences are dependent on the registers. The analysis and results are described in the next Section.

5. Analysis per Part of speech (POS)

5.1. TTR's per Part of speech (POS)

For the analysis per POS, new subcorpora of 1350 tokens are created, selecting only nouns (N), verbs (V), adjectives (A), and function words (Func) respectively.[10] The question we would like to answer by analyzing the TTR's of these samples is twofold. First of all, are there large differences between the TTR's for N, A, V and Func? Secondly, are there differences in the distribution over the 14 registers? Figure 1 summarizes the results, plotting the average TTR of the different POS's (on the vertical axis), for each of the corpus components[11]. 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 lie very close to each other (their respective means are 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 confirm expectations : nouns, adjectives and verbs are open word classes, varying in accordance with the topic at hand, while function words form a closed set of lexical items. Further, the difference between nouns on the one hand and adjectives and verbs on the other

[9] It might be that this is a form of *hypercorrection*, in the sense that in unusual, monologic speech situations, these speakers try harder to use formal, 'ceremonial' language.
[10] The group of function words consists of interjections, articles, conjunctions, pronouns and prepositions.
[11] For component m, which is a very small component, a subcorpus of 1350 adjectives could not be constructed.

hand could very well be that the latter 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 line for the function words shows that there are only small TTR differences between the different registers. In general, the distribution for the verbs and the adjectives is similar, although for a few components, the distances are larger.[12] 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. This is the only register where the TTR for N is *not* significantly different from both A and V ($p > 0.05$ in Welch two-sample t-tests). It is interesting that this effect shows up in this restricted register, typically covering only a limited range of topics. On the other hand, for component k, for example, which contains news texts about a wide range of topics, the distance with the A and V curves is a lot larger.[13]

It can be concluded that breaking up the analysis per POS helps in locating lexical richness effects in texts. We find indications that for nouns, the TTR is more heavily influenced by the thematic content, inherent in the different registers, than for adjectives, verbs, and especially function words. As will be discussed in Section 5, a further analysis of the content-dependency may be envisaged on the basis of a keyword analysis. In view of the importance of the POS effects, in the next paragraph, an additional analysis is first discussed, considering the *density* of the different parts of speech. The use of a particular POS is 'dense' if the POS is (relatively) frequent in a text. As will be explained, the density is measured by calculating the number of raw text tokens needed to collect a sample of 1350 tokens of the POS.

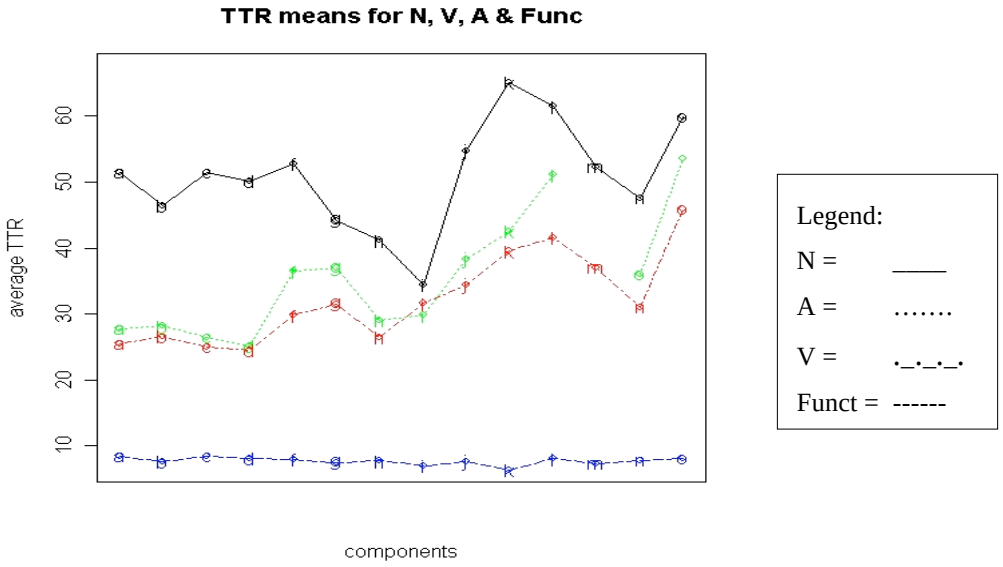

Figure 1 : Average TTR of the CGN components per POS (Nouns, Verbs, Adj and Function words)

[12] Most noticeable in this respect is the relatively high average TTR for the adjectives as opposed to the verbs in component l (viz. 51.04 vs. 41.6 ; $p > 0.05$). This component contains 'evaluative informative texts' (viz. commentaries, reviews, and columns). A closer look at this component reveals that 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'), resulting in a high TTR for adjectives.

[13] With regard to component k, it could be remarked that, even though there seems to be very little variation between the TTR's of the *function words*, this register has a significantly low TTR ($p < 0.01$). It makes sense that news items, which are well prepared and highly informational, contain a smaller array of function words.

5.2. POS density and lexical richness

In the previous paragraphs, it was shown that an analysis of subcorpora consisting of one POS reveals interesting lexical richness effects. Yet, a further question concerns the *density* of the parts of speech. For example, in order to construct a subcorpus of 1350 adjectives, *how many tokens* (of raw text) do we have to sample? Do we find differences in density between the parts of speech, and, more importantly, how are these differences related to the POS analysis in 4.1.? To answer these questions, for each of the POS-subcorpora, the original *text length* is *reconstructed*. Thus, we calculated the total number of tokens sampled for each subcorpus of 1350 POS-tokens. The resulting text lengths are then divided by 1350 (the number of POS-tokens) in order to get easily comparable results, indicating the number of tokens sampled per POS-token.

A first analysis shows that there are significant *differences* between the four *parts of speech* : while the texts needed to sample 1350 POS-tokens are signficantly longer for adjectives than for nouns, those for verbs and especially for function words are shorter. In other words, the density of the function words is highest, followed by the verbs, nouns, and adjectives.[14] These results are in line with expectations : function words are often marked as the most frequent words in texts (see e.g. Burrows, 1987), and (almost) all grammatical sentences need a verb. It also seems logical that texts contain more nouns than adjectives, which often give additional descriptive or evaluative information.

Secondly, in view of the important effect of 'register', shown in 4.1, the differences between the densities for the *components* again have to be investigated. Also, for each of the components, the *interactions* between 'density' and 'lexical richness' should be studied. To take the nouns as an example, is it the case that more lexically rich components also have a higher lexical density? The results are the following. For the *nouns*, all components, except for the telephone dialogues, have a significantly higher density than the conversations (component a). This seems to indicate that a high lexical richness goes hand in hand with a high nominal density. Yet, this interaction is only significant for the news reports (component j), indicating that this prepared, informational register has a fairly 'nouny' style. Further, the interaction for component i shows that in sport commentaries, a low TTR goes together with a relatively high nominal density. This seems to point at the use of a repetitive style : a small array of words is used many times. The densities of the *adjectives* are comparable to the nouns : again, all signficant components have a higher adjectival density than the conversations. Not unexpectedly, in read-aloud texts (o), there is a positive correlation between a high lexical richness and a high adjectival density. The *verbs* show a more mixed picture. There is a signficantly higher verbal density in classroom speech (h) and read-aloud texts (o), while a lower density is found in spontaneaous components such as b (interviews with teachers of Dutch) or i (sport commentaries), but also in the formal, well-prepared components k and m (news and speeches). It is not surprising that the interactions with the TTR's also seem to go in different directions. The telephone dialogues (in c) have a low verbal density, combined with a low TTR. Components i and j (sport commentaries and news reports) also have signficantly low verbal densities, in comparison to their average TTR. This result is probably related to the high nominal density in these registers. In read-aloud speech,

[14] To make this more concrete: one out of every 15.83 tokens is an adjective, one out of 7.99 a noun, one out of 5.84 a verb and one out of 2.48 a function word.

we also find a low verbal density, but this time combined with a very high TTR.[15] Finally, the *function words* show a very different picture than the other POS. Here, as expected, all registers, except for the telephone dialogues (c and d), have a lower density than the conversations. The interactions with the TTR indicate that these informal, conversational components combine a high density of function words with a high TTR, while the formal news component (k) has both a low TTR and a low density for this POS.

It can be concluded that the density of the POS further highlights the specific lexical make-up of the different registers. Generally speaking, for the four POS, it seems to be the case that registers with a high lexical richness also have a fairly high density. Thus, conversational, informal registers, which were shown to have a low TTR, have a high density of function words, but a low density of content words such as nouns or adjectives. Further, there are a number of interesting deviations from this general tendency, which seem related to specific *stylistic aspects* of the registers (e.g. the high nominal density in the sport commentaries). Thus, the lexical typology of the registers seems to be co-determined by both thematic and stylistic effects, that have to be taken into account when locating lexical richness effects.

6. Conclusions & further research steps

The results presented show that the TTR, performed on carefully sampled corpus subsamples 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 results for the other extralinguistic parameters under analysis were less significant, a consistently lower TTR for women was found, both in a global analysis and in split-up analyses for conversational, informal registers. 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.

Further, it was shown that the influence of the *parts of speech* should be acknowledged when locating lexical richness effects. First of all, the TTR is highly dependent on the POS, while for each of the parts of speech, there are also significant differences in the distribution of the TTR between the registers. Secondly, there are some interesting interactions between the *density* and the TTR's of the components. The results of the analyses, especially for the nouns, seem to indicate that the lexical richness of texts is influenced by *topic dependent* as well as *stylistic* effects, which are related to the communicative purpose of the texts.

In consequence, in order to locate lexical richness effects more precisely, in a next step, the thematic and communicative specificity of the texts should be studied in more detail. To this aim, a *keyword analysis* is proposed. First of all, interpreting the (range of) lexical items that are significant in the subcorpora should allow for an analysis of the *thematic multiplicity* of the texts. A comparison of the TTR for the keywords with the TTR of the subcorpora could indicate if the keywords are a strong determiner for the lexical richness found, and if this

[15] This effect may partly be due to the influence of *auxiliaries*, which are counted as verbs, while they could be considered function words. It could be that these verbs are used less in highly prepared speech. The effect of auxiliaries on the TTR should be analysed in more detail.

effect is again dependent on the register of the texts. It would also be interesting to supplement this with a more *qualitative* analysis, to check whether there are differences in the *type* of keywords found in the different subcorpora. For example, do we find more nominal keywords in monologic, prepared data, while in informal dialogues, discourse particles show up as significant keywords? Finally, taking into account the thematic and communicative effects detected, the aim is to re-incorporate the other extralinguistic factors (region, sex and educational level) in the analysis. This should result in a better, more integrated model for measuring the distribution of lexical richness effects.

References

Arnaud, P. J. L. (1984). The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests. In Culhane, T., Klein Bradley, C. and Stevenson, D.K., editors, *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 Lüdeling, A. and Kytö, M., editors, *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).

Burrows, John F. (1987). Word-patterns and story-shapes : The statistical analysis of narrative style. *Literary and Linguistic Computing*, vol. 2(2) : 61-70.

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

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

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

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

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

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*, vol. 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*, vol. 32 : 323–352.

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

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