Lund University, Dept. of Linguistics and Phonetics
Working Papers 53 (2008), 61–79

Lexical diversity and lexical density in speech and writing: a developmental perspective

Victoria Johansson

Introduction

Literature about early, pre-school lexical development often mentions *vocabulary development*. As an example, the reader of *Handbook of child language* (Fletcher and MacWhinney 1995) is referred to 'vocabulary development' when looking up the term 'lexical development'. The same term is used by e.g., Dromi 1999 in her overview of early lexical development, and the index in David Crystal's *The Cambridge encyclopedia of language* (1997) refers to 'vocabulary' from the index entry 'lexicon'.

This article will compare two measures that often have been used to describe later lexical development: *lexical diversity* and *lexical density*. Lexical diversity is a measure of how many different words that are used in a text, while lexical density provides a measure of the proportion of lexical items (i.e. nouns, verbs, adjectives and some adverbs) in the text. Both measures have the advantage of being easy to operationalise, and also practical to apply in computer analyses of large data corpora. Further, both lexical diversity and lexical density have been shown to be significantly higher in writing than in speaking (Ure 1971, Halliday 1985). One conclusion from this could be that the two measures are interchangeable, and that we will encounter a similar developmental pattern independent of the measure used for describing lexical development.

It is, however, theoretically possible that a text has high lexical diversity (i.e. contains many different word types), but low lexical density (i.e. contains many pronouns and auxiliaries rather than nouns and lexical verbs), or, vice versa, that a text has low lexical diversity (i.e. the same words or

phrases are repeated over and over) but high lexical density (i.e. the words that are repeated are nouns, adjective or verbs).

Lexical diversity is often used as an equivalent to *lexical richness* (e.g., by Daller, van Hout & Treffers-Daller 2003). However, Malvern et al. 2004 begin their book about lexical diversity with discussing the difference between lexical diversity and lexical richness, stating (along the lines of Read 2000) that the lexical diversity measure is only one part of the multidimensional feature of lexical richness. Other factors proposed by Read are lexical sophistication, number of errors, and lexical density (Read 2000). I side with Read and Malvern et al.; neither lexical diversity nor lexical density is the one and only measure. However, both measures are easily accessible and easy to apply to corpora of different kinds. No doubt they also provide important insights into the texts, and as long as the measures are not used as the only way to judge a text qualitatively, they are very useful.

The aim and outline of the study

This study focuses on developmental patterns in terms of the measures lexical diversity and lexical density. I will examine whether these measures are sensitive to genre (narrative vs. expository) and modality (writing vs. speaking). Another goal is to investigate to what extent the two measures are correlated.

The article starts with a theoretical background on the two measures, followed by a presentation of the data, then moves on to statistical analyses presented measure by measure, age group by age grup, and ends with a general discussion and a conclusion.

Lexical diversity

The more varied a vocabulary a text possesses, the higher lexical diversity. For a text to be highly lexically diverse, the speaker or writer has to use many different words, with little repetition of the words already used.

The type-token ratio

The traditional lexical diversity measure is the ratio of different words (types) to the total number of words (tokens), the so-called type-token ratio, or TTR (e.g., Lieven 1978, Bates, Bretherton & Snyder 1988). A problem with the TTR measure is that text samples containing large numbers of tokens give lower values for TTR and vice versa. The reason for this is that the number of word tokens can increase infinitely, and although the same is true for word

types, it is often necessary for the writer or speaker to re-use several function words in order to produce one new (lexical) word. This implies that a longer text in general has a lower TTR value than a shorter text, which makes it especially complicated to use TTR in developmental comparisons, e.g., between age-groups, where the number of word tokens often increase with age. Gayraud 2000 compares TTR and the number of word tokens and shows that although the number of word tokens increases substantially with speaker/writer's age, the TTR drops.

One consequence of this is that TTR is only possible to use when comparing texts of equal length. In spite of this, TTR is still used for comparing text production, for instance between children's texts, or between various groups with language impairment. For instance, TTR is part of the *SALT* (*Systematic Analysis of Language Transcripts*) programs, a set of computer programs developed by Miller and Chapman in order to quantify developmental aspects of speech for typically as well as atypically developing children (Miller & Klee 1995).

A variant of the TTR measure is the so-called *index of Guiraud*. This measure uses the square root of TTR. Other proposed variants are *Advanced TTR* and *Guiraud Advanced*, for instance used by Daller et al. 2003.

Vermeer 2000 discusses TTR and various other measures, and their use in both first and second language acquisition. She concludes her discussion with proposing that lexical richness can be more successfully measured by exploring the degree of difficulty for the words in a text, as measured by their frequency in everyday life.

Theoretical vocabulary

Other ways around the TTR-problem have been proposed and used. One is the so-called *theoretical vocabulary* (see e.g. Broeder, Extra & van Hout 1986). The principle behind this measure is to pick a number of words (e.g., 100 words) from a text at random, and calculate the number of word types in the sample. The theoretical vocabulary takes into account all possible ways of choosing 100 words from the text. In this way, one can compare texts of different lenghts, with the only restriction that the shortest text limits the maximal number of random words to be picked.

Johansson 1999 uses theoretical vocabulary for comparing spoken and written expository texts between a group of Swedish university students and 12-year-olds. In this case the program *Vocab* (developed by Leif Grönqvist, Department of Linguistics, Göteborg University) was used for calculating

theoretical vocabulary. The result shows that the lexical diversity is higher in writing than in speech for both the adults and the 12-year-olds. The adults have higher lexical diversity than the 12-year-olds. *Vocab* was also used by Wengelin 2002 to compare written texts in various genres from three populations: a group of adult controls, a group of congenitally deaf adults, and a group of adults with reading and writing difficulties. The adult controls had higher diversity than the other groups. Some of the written texts had spoken equivalents, and Wengelin was able to show that the control group had a greater difference between their spoken and written texts than the group with reading and writing difficulties.

VocD

In order to compare texts of different lengths, a measure independent of sample size is required. One such measure is the *D* measure developed by Brian Richards and David Malvern (Richards & Malvern 1997, Malvern et al. 2004, MacWhinney 2000). The *D* measure is based on the predicted decline of the TTR, as the sample size increases. This mathematical curve is compared with empirical data from a text sample. For calculating *D*, information from the whole text sample is used (the minimum length of the text is 50 words, however). A higher value of *D* indicates higher lexical diversity, and thus a richer vocabulary. The *D* measure is implemented in the most recent versions of *CLAN* (MacWhinney 2000), under the name *VocD*. The measure *VocD* is described at length in Malvern et al. 2004, with many examples and references to previous studies about lexical measures.

Although Malvern and Richards claim that *VocD* permits comparisons between texts of unequal length, not everybody is convinced that the text length factor is completely eliminated by using *VocD*. The *D* measure has been criticised, for instance by Daller et al. 2003, who instead prefer the index of Guiraud.

Malvern and Richards' *D* measure is severely criticised by McCarthy and Jarvis 2007 for not being insensitive to text lengths. McCarthy and Jarvis compare *D* to 13 alternative methods for measuring lexical diversity. They conclude that *D* (or *VocD*) performs better than most alternatives, but that there are better options. However, another conclusion is that the length of texts one wants to compare should determine which measure one uses, since some measures are more effective within certain ranges. Their analysis shows that *D* is the second best of all measures within the text length of 100–400 word tokens, which is also what is claimed in Malvern et al. 2004. McCarthy and Jarvis 2007:483 finish by questioning "whether a single index has the capacity to encompass the construct of lexical diversity".

Strömqvist et al. 2002 used *VocD* to compare spoken and written expository and narrative texts produced by adults from four countries. The results show strong differences between speech and writing, where writing has a much higher lexical diversity. However, a conclusion from this study is that one should be careful when using the measure to compare data from different languages. The morphological structure of the language highly influences the outcome of the comparison.

The definition of lexical diversity in this article

To conclude, there are several ways to compare lexical diversity between texts of different lengths. In spite of some criticism, *VocD* seems to be the most accurate instrument to use. For the calculations of lexical diversity below, I will consequently use the measure *D*.

Lexical density

Lexical density is the term most often used for describing the proportion of content words (nouns, verbs, adjectives, and often also adverbs) to the total number of words. By investigating this, we receive a notion of *information packaging*; a text with a high proportion of content words contains more information than a text with a high proportion of function words (prepositions, interjections, pronouns, conjunctions and count words).

Various variants of lexical density have been proposed. A popular 'minor variant' is to calculate the *noun density*, the number of nouns divided by the total number of tokens in the text. Other options are for instance verb or adjective or adverb types per total lexical words. Various options are described and discussed in Wolfe-Quintero, Inagaki & Kim 1998.

Introducing the concept of *lexical density*, Ure 1971 distinguishes between words with lexical properties, and those without. According to Ure, items that do not have lexical properties can be described "purely in terms of grammar" (p. 445), meaning that such words (or items) possess a more grammatical-syntactic function than the lexical items. Lexical density is then defined as the total number of words with lexical properties divided by the total number of orthographic words. The result is a percentage for each text in the corpus. Ure concludes that a large majority of the spoken texts have a lexical density of under 40%, while a large majority of the written texts have a lexical density of 40% or higher. One remark here is that these numbers

ought to be highly language dependent – a language with more bound morphology would probably show a higher proportion of lexical items.

In a later article, Ure defines lexical density as "the proportion of words carrying lexical values (members of open-ended sets) to the words with grammatical values (items representing terms in closed sets). Since *all* words have grammatical values, this is a part : whole relation" (Ure & Ellis 1977:207).

Ure and Ure & Ellis correctly maintain that the matter of lexicality is important when discussing the concept of lexical density. Traditionally, nouns, verbs and adjectives are the three word classes considered to have lexical properties (although this is not stated clearly in Ure 1971 or Ure & Ellis 1977). Often these items are called *content words* or *open class words* (because of the possibility to easily include new members of the class – while the more grammatical parts of speech are called *closed classes*, since new prepositions or pronouns seldom enter the language).

The concept of lexical density is developed, and further refined by Halliday 1985. He points out the importance of discriminating between lexical items and grammatical items. An item may consist of more than one word. Thus, Halliday counts *turn up* as one lexical item, while Ure 1971 counts it as one lexical item (*turn*) and one grammatical item (*up*). A lexical item is by Halliday defined as an item that "function[s] in lexical sets not grammatical systems: that is to say, they enter into open not closed contrasts" (Halliday 1985:63). The lexical item is part of an open set, that can be contrasted with a number of items in the world. A grammatical item, on the other hand, enters into a closed system, according to Halliday. Characteristic for the grammatical system is that the (word) classes belonging to it have a fixed set of items, where it is impossible to add new members.

According to Halliday, child language gives evidence for the existence of two classes, one with lexical and one with grammatical items. In the beginning of their linguistic development, children often construct sentences where all grammatical items are missing. Halliday further emphasises that there is a continuum from lexis into grammar, and that there are – and always will be – intermediate cases. For instance, he claims that English prepositions and certain classes of adverbs are on the borderline between lexical and grammatical items. The adverbs that he gives as examples are the modal adverbs, such as *always* and *perhaps*. When comparing e.g. speech and writing, the important thing is to be consistent in drawing the line between

'lexical' adverbs and 'grammatical' adverbs, but it matters less where the line is drawn.

The definition of lexical density given by Halliday is thus "the number of lexical items, as a proportion of the number of running words" (Halliday 1985:64). The difference between Halliday's and Ure's definitions of lexical density is that Halliday counts some adverbs as lexical items.

The definition of lexical density in this article

This article follows Halliday's definition of lexical density. Thus, grammatical adverbs are included in the closed class items, while non-grammaticalised adverbs (including all adverbs derived from adjectives) are counted as lexical items. In our data, lexical density was calculated by dividing the number of lexical items by the total number of words in each text.

Data

To compare lexical diversity and lexical density in a developmental perspective, I have used material from the Swedish part of an international study on developing literacy, the so-called *Spencer project*[1] (for more details on data collection, see Berman and Verhoeven 2002, or Johansson 2008). The Spencer study aimed at investigating the development of literacy in both speech and writing in two different genres: narrative and expository. The Swedish data consist of 316 texts distributed evenly on written and spoken narrative and expository texts. Four age groups participated in the study: 10-year-olds (4th-graders), 13-year-olds (7th-graders), 17-year-olds (11th-graders), and adults (university students with at least 2 years of university education, during which they had produced at least one major paper). All participants were monolingual Swedish speakers[2], with no known reading or writing difficulties. Each group consisted of 20 persons, except the adult group which had only 19 members. The text length range was 50–650 words.

After watching a wordless elicitation movie showing scenes from a school-day (e.g., from cheating, fighting, bullying, stealing), the participants were asked to produce four texts each. The experimental tasks were balanced

[1]The project was supported by the Spencer Foundation Major Grant for the Study of Developing Literacy to Ruth Berman. Apart from Sweden, six other countries participated: Israel, Netherlands, France, Spain, Iceland and California, USA.

[2]'Monolingual speaker' here means that both parents had Swedish as their first language, and that Swedish was the main language used both at home and at school. At the time of the recording, all subjects had at least started to learn English in school, however, and some of the participants in the adult group might have spent long time abroad.

for order. The text types and the topic for each task were as follows (with the elicitation question rephrased):

- *Spoken narrative (NS)*: Tell me about one time when you helped somebody in/was helped by somebody out of a predicament.
- *Written narrative (NW)*: Write about one time when you helped somebody in/was helped by somebody out of a predicament.
- *Spoken expository (ES)* (i.e. a speech): Give a speech, where you discuss the problems you just saw in the film. Don't describe the film, but instead say something about the cause of the problems, and possible solutions.
- *Written expository (EW)* (i.e. an essay): Write an essay where you discuss the problems you just saw in the film. Don't describe the film, but instead say something about the cause of the problems, and possible solutions.

Correlating lexical diversity and lexical density

Before exploring each lexical measure individually, a correlation test will give a hint on whether or not the two measures are connected in the data. Not surprisingly, given that both measures have been proposed to show lexical development, there proved to be a highly significant correlation between lexical diversity and lexical density ($r = 0.733, p < 0.01$).

Overall patterns of age, modality and genre

After stating that lexical diversity and lexical density are correlated in the data, multivariate ANOVA was used to explore overall patterns of age, modality and genre for each lexical measure.

To summarise the results below, the general effects were significant for almost all factors, including an interaction of genre, age and modality.

To investigate the main effects of genre and modality and the interactions between these factors, a within-subject factor test was used, while a between-subjects test was used to look for main effects of age. Table 1 shows an overview of the results of the *post hoc* tests.

Lexical diversity: Multivariate analyses

Multivariate analyses of lexical diverstity show a significant main effect of genre ($F(1,69) = 4.236, p < 0.05, \eta^2 = 0.058$), of modality ($F(1,69) = 333.805$, $p < 0.01, \eta^2 = 0.829$), and of age ($F(3,69) = 3302.206$, $p < 0.01, \eta^2 = 0.702$).

Table 1. Results of the *post hoc* comparisons between lexical diversity and lexical density.

Lexical Measure	*Subset 1*	*Subset 2*	*Subset 3*
Lexical diversity	10-year-olds 13-year-olds	17-year-olds	Adults
Lexical density	10-year-olds 13-year-olds	17-year-olds Adults	

A significant interaction of modality and age is also found ($F(3,69) = 11.664, p < 0.01, \eta^2 = 0.336$), as with genre and modality ($F(3,69) = 3.363$, $p < 0.05, \eta^2 = 0.128$). However, there is no significant interaction of genre and age group.

Tukey's *post hoc* analyses show no significant difference between the two youngest age groups (10-year-olds and 13-year-olds), but a significant difference between the two youngest age groups and the two oldest ones. Further, there was a significant difference between the two oldest groups, in that the adults had higher lexical diversity than the 17-year-olds (cf. the subsets from the *post hoc* tests in Table 1).

Lexical density: multivariate analyses

Multivariate analyses of lexical density show a main effect of modality ($F(1,75) = 651.744$, $p < 0.01$, $\eta^2 = 0.897$), and of age ($F(3,37) = 20.215$, $p < 0.01, \eta^2 = 0.447$), but no effects of genre.

Further, a significant interaction of genre and age is found ($F(3,75) = 4.181, p < 0.01, \eta^2 = 0.143$), and of modality and age group ($F(3,75) = 3.811$, $p < 0.05, \eta^2 = 0.132$), but there is no interaction between genre and modality.

Tukey's *post hoc* analyses show no significant differences between 10-year-olds and 13-year-olds, or any significant differences between 17-year-olds and adults. However, there is a significant difference between the two younger age groups on the one hand, and the two oldest age groups on the other (cf. the subsets from the *post hoc* test in Table 1).

Conclusion: overall patterns of age, modality and genre

Table 1 shows the homogeneous subsets from the *post hoc* tests, and summarises in that way the differences between the lexical measures. 17-year-olds and adults differ significantly for the lexical diversity measure, but their texts appear to be equally lexically dense. Thus, the progression of development is more outstreched when we use lexical diversity.

One conclusion is that although the correlation test shows a strong correlation betwen the lexical measures, and the multivariate analyses show a significant main effect of age, the developmental pattern varies depending on the lexical measure of investigation.

Both measures showed a modality effect, in that they are significantly higher for the written discourses. This confirms the results from Ure 1971. However, we only find an effect of genre for lexical diversity; the lexical density measure seems to be indifferent to genre.

From this follows that although the lexical measures are correlated, we might get different insights depending on which measure we look at. Compared with lexical density, lexical diversity proved to be more genre sensitive, as well as more sensitive to development.

Comparing text types within each age group

In the following, I will compare each measure within each text type (narrative written, narrative spoken, expository written or expository spoken) as well as within each age group.

A multivariate ANOVA will be used to compare the differences for each text type within each age group, with the aim of investigating genre and modality differences within each age group. If such differences can be established, a paired sample *t*-test will be used to find differences between pairs within a factor, e.g., differences between expository spoken texts and narrative spoken texts.

Lexical diversity

Table 2 shows the means of lexical diversity broken down by age group and text type, and Figure 1 illustrates this graphically. There is a trend for lexical diversity to increase with age, and the striking difference between speech and writing for all age groups, independent of text type, is salient in the figure.

10-year-olds

The 10-year-olds show a significant effect of genre ($F(1,14) = 5.355$, $p < 0.05$, $\eta^2 = 0.277$), in that the expository texts are more lexically diverse than the narrative ones. The highest lexical diversity is found in the written expository texts, and the lowest lexical diversity is found in the spoken narrative texts.

Table 2. Lexical diversity: means broken down by age group and text type.

Text Type	*10-year-olds*	*13-year-olds*	*17-year-olds*	*Adults*
NS	38.39	36.18	50.17	61.11
NW	61.72	66.52	103.07	107.05
ES	47.23	42.51	54.02	81.12
EW	67.50	64.03	98.54	99.15

Furthermore, there is a significant effect of modality ($F(1,14) = 34.748$, $p < 0.01$, $\eta^2 = 0.713$), so that the written texts have higher lexical diversity than the spoken ones.

A paired sample *t*-test shows no significant differences between the two spoken genres or the two written genres.

13-year-olds

The 13-year-olds show a significant effect of modality ($F(1,18) = 141.876$, $p < 0.01$, $\eta^2 = 0.887$), but no effect of genre. This means that the 13-year-olds have higher lexical diversity in their written texts, but there is no difference between narrative written and expository written texts.

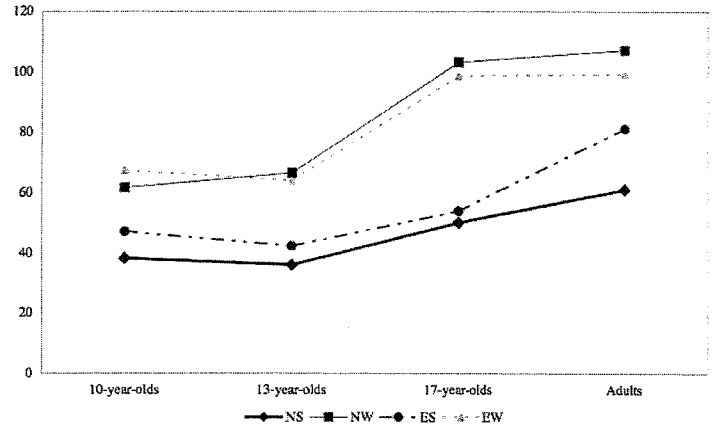

Figure 1. Lexical diversity broken down by age group and text type.

However, a paired sample *t*-test shows a significant difference between the narrative spoken texts and the expository spoken texts ($t(19) = 2.698$, $p < 0.05$). Thus, for the spoken texts there is a genre effect, where the expository spoken texts are more lexically diverse.

17-year-olds

The 17-year-olds show patterns similar to the 13-year-olds'. Thus, there is a significant effect of modality ($F(1,19) = 132.124$, $p < 0.01$, $\eta^2 = 0.874$), but no significant effect of genre. This means that the 17-year-olds have higher lexical diversity in their written texts, but there is no difference between narrative and expository texts.

Adults

Like the younger age groups, the adults also show a significant effect of modality ($F(1,18) = 86.502$, $p < 0.01$, $\eta^2 = 0.828$). Again, there is no significant effect of genre. The written texts thus have higher lexical diversity than the spoken texts, but there is no difference between narrative and expository texts. A paired sample *t*-test shows that the adults, just like the 13-year-olds, have a difference between their spoken narrative texts and their spoken expository texts ($t(18) = 3.378$, $p < 0.01$); the expository spoken texts have higher lexical diversity.

Lexical density

Table 3 presents the means of lexical density broken down by age group, and text type. Figure 2 gives a graphic overview of the same data. Just as for lexical diversity, the graph of lexical density show a difference between the spoken and the written texts, independent of genre. We also find a trend for lexical density to increase with age, although the trend seems less salient than for lexical diversity (cf. Figure 1).

10-year-olds

The 10-year-olds show a significant difference of modality ($F(1,19) = 127.360$, $p < 0.01$, $\eta^2 = 0.870$), in that the written texts have higher lexical density than the spoken ones. However, there are no genre effects.

13-year-olds

The 13-year-olds show a significant modality effect ($F(1,19) = 171.839$, $p < 0.01$, $\eta^2 = 0.900$), with the highest lexical density in the written texts.

Table 3. Lexical density: means broken down by age group and text type.

Text Type	*10-year-olds*	*13-year-olds*	*17-year-olds*	*Adults*
NS	00.23	00.24	00.25	00.27
NW	00.32	00.33	00.36	00.39
ES	00.23	00.20	00.25	00.28
EW	00.32	00.31	00.39	00.39

Furthermore, there is a genre effect, where the narrative texts have higher lexical density than the expository texts ($F(1,19) = 9.942$, $p < 0.01$, $\eta^2 = 0.344$). Thus, the narrative written texts show the highest lexical density, while the lowest lexical density is found in the expository spoken texts. A paired sample *t*-test shows a significant difference between the narrative spoken texts and the expository spoken texts ($t(19) = -7.730$, $p < 0.001$). Thus, there is a genre effect for the spoken texts, where the narrative spoken texts are more lexically dense than the expository ones.

17-year-olds

The 17-year-olds show a significant effect of modality ($F(1,19) = 183.290$, $p < 0.01$, $\eta^2 = 0.906$), where – again – the written texts have higher lexical density than the spoken texts. There are no effects of genre.

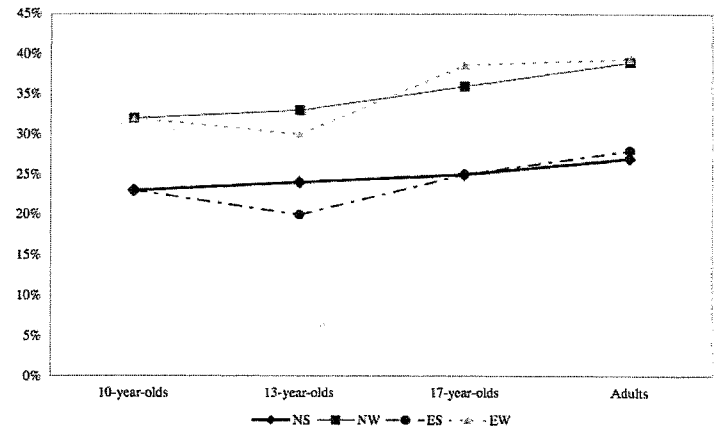

Figure 2. Lexical density broken down by age group and text type.

Adults

The adults have a significant effect of modality ($F(1,18) = 173.284, p < 0.01$, $\eta^2 = 0.906$), where the lexical density is higher in the written texts. No genre effects are found.

Conclusion: lexical measures within each age group

After comparing the text types within each age group, some conclusions can be drawn. First, the 10-year-olds show a modality difference in all tests, but no genre differences. Thus, this group seems to be highly sensitive to the modality, but able to adapt their lexicon less to genre.

The 13-year-olds, on the other hand, is the odd group here; they show both modality and genre differences. The modality difference is not so difficult to explain; independent of age group a more diverse and dense language is generally required in writing compared to speaking, due to the decontextualised conditions in writing. In the same way it is problematic to produce a lexically diverse or dense text in speech; repetitions are necessary (which makes the diversity lower), and pronouns are both more adequate and easier accessible in speech (which makes the density lower).

More interesting are the genre effects found in the 13-year-olds' texts. Their spoken expository texts have higher lexical diversity than the narrative equvalents, but the spoken narrative texts have higher lexical density than the equivalent expository spoken texts. This means that more content words are used in their spoken narratives, but that the vocabulary is more varied in the spo-ken expositories. One factor which would increase diversity, but decrease density is an extensive use of pronouns (such as 'they', 'we', 'I', or *man* ('one', the Swedish generic pronoun) to express degrees of generalisation, in combination with consecutive conjunctions ('because', 'so that', 'therefore') to express connections between problems and solutions. Another factor is that the elicitation of the expository task invited to a more context-bound discourse; all subjects knew that the experiment leader had seen the elicitation movie. In addition they were explicitly told not to describe the movie. Together, these factors may invoke a more extensive use of pronouns, and thereby decrease lexical density, while the diversity remains higher than in the narratives since the variation of pronouns is great.

The 17-year-olds show a modality difference for all the lexical measures, but no genre differences at all. In the light of the 13-year-olds' pattern we could interpret the results so that the 17-year-olds increase their number of lexical items in the expository texts, and in that way even out genre differences for lexical density here. The genre effect of lexical diversity can be explained by the 17-year-olds' immense use of fillers and empty phrases; they are making strong efforts not to be silent during their spoken expositories. On the other hand, this decreases the lexical diversity substantially.

The adults show, like all groups, modality differences, but the genre differences in this group are especially interesting in a developmental perspective. To resume, at the upper end of the developmental scale, the adults use a more varied vocabulary in their spoken expository texts than in the spoken narrative ones. One conclusion is that the adults are able to use knowledge acquired and practised in writing, also when they speak, and that this is most noticeable in the cognitively more demanding expository genre.

Comparing text types between age groups

Following strong indications of significant differences between genres, between modalities, and finally between age groups, ANOVA will be used to examine how the differences distribute over age in each text type. Thus, for each lexical measure I will look for differences in age groups in each text type (Expository spoken, Expository written, Narrative spoken and Narrative written). Since each participant only wrote one text of each text type, an ordinary ANOVA can be used to compare the four age groups with each other. Results from Tukey's *post hoc* tests will be used to explore significant age group differences within a text type.

Lexical diversity: text types and age

Figure 1 showed that lexical diversity increased with age. This is confirmed by the findings previously presented. The results from Tukey's *post hoc* tests presented in Table 4 show how the age groups can be divided into homogeneous subsets in the various text types. The table indicates that there are no differences between 10-year-olds and 13-year-olds. Nor do 17-year-olds and adults differ in the written conditions.

However, the spoken conditions show a more outstretched developmental pattern. The 17-year-olds use a more varied vocabulary in the narrative spoken texts than the youngest age group, indicating that the familiarity of that text type facilitates a more lexically diverse text production. The adults find it even more easy to vary their lexicon in their spoken narratives. Notable is also that the 17-year-olds are not more lexically diverse in their

Table 4. Lexical diversity: results of the *post hoc* comparison, presented for each text type separately.

Text type	*Subset 1*	*Subset 2*	*Subset 3*
NS	10-year-olds 13-year-olds	17-year-olds	Adults
NW	10-year-olds 13-year-olds	17-year-olds Adults	
ES	10-year-olds 13-year-olds 17-year-olds	Adults	
EW	10-year-olds 13-year-olds	17-year-olds Adults	

spoken expositories than the 10-year-olds and the 13-year-olds, while the adults outrule them all.

Lexical density: text types and age

If we end the analysis by looking at how the age groups divide into homogeneous subsets for the text types in lexical density we achieve a more complicated pattern.

Table 5 gives an overview of the results of the *post hoc* tests for lexical density. It shows that the narrative spoken texts are equally dense for all age groups. Thus, the adults use the same proportion of content words as the 10-year-olds! In the narrative written condition the pattern is more stretched-out, indicating that the 17-year-olds and the adults differ from the 10-year-olds. Further, the adults differ from the 13-year-olds.

For the expository spoken texts, again, the adults are outstanding. They use more content words than the other age groups. In the written expository texts, however, we find no difference between adults and 17-year-olds, indicating that the 17-year-olds can compete with equally lexically dense texts in writing, but not in speech. As has been proposed before, one explanation might be that the adults take time to think before they formulate their spoken texts, while the 17-year-olds repeat the same phrases while thinking, decreasing the ratio of content words to the total number of word tokens.

Conclusion

This study has shown that although both lexical density and lexical diversity can be used to account for modality differences and developmental

Table 5. Lexical density: results of the *post hoc* comparisons presented for each text type separately.

Text type	*Subset 1*	*Subset 2*	*Subset 3*
NS	10-year-olds 13-year-olds 17-year-olds Adults		
NW	10-year-olds 13-year-olds	13-year-olds 17-year-olds	17-year-olds Adults
ES	10-year-olds 13-year-olds	13-year-olds 17-year-olds	Adults
EW	10-year-olds 13-year-olds	17-year-olds Adults	

differences, a closer analysis where both measures are used on the same material reveals that they are not interchangeable.

Interesting enough, for both measures, there is no age difference between the 10-year-olds and the 13-year-olds. In the same way, we do not find differences between 13-year-olds and 17-year-olds for all text types. This indicates that although there is an age factor involved in the increase of lexicon (independent of measure), these patterns will not always be salient if we do not look at a long term development. One should be careful not to use these measures alone when comparing texts produced by children with small age differences.

Another conclusion is that we perceive a more noticeable developmental trend for lexical diversity than for density. This suggests that lexical diversity is a better measure to use for detecting differences between age groups.

Finally, much development takes place between the last years in high school, and the university. The main differences, independent of measure, have been found in the spoken conditions between the adults and the other age groups. I would like to propose that the adults' more extensive use of written language (both reading and writing) have given them a vocabulary platform, which facilitates not only their written language, but also have high influence on their spoken productions. The 17-year-olds are in many ways able to compete with the adults in writing (when it comes to a varied, lexical dense vocabulary), when the time constraints of speech is removed, but this varied vocabulary is less accessible in writing.

References

Bates, E., I. Bretherton & L. Snyder. 1988. *From first words to grammar: individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Berman, Ruth A. & Ludo Verhoeven. 2002. 'Cross-linguistic perspectives on the development of text-production abilities: speech and writing'. *Written Language and Literacy* 5, 1-44.

Broeder, Peter, Guus Extra & Roeland van Hout. 1986. 'Measuring lexical richness and diversity in second language research'. *Polyglot* 8, 1-16.

Crystal, David. 1997. *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

Daller, Helmut, Roeland van Hout & Jeanine Treffers-Daller. 2003. 'Lexical richness in the spontaneous speech of bilinguals'. *Applied Linguistics* 24 (2), 197-222.

Dromi, Esther. 1999. 'Early lexical development'. In Martyn Barrett (ed.), *The development of language*, 99-131. Hove: Psychology Press.

Fletcher, Paul & Brian MacWhinney. 1995. *The handbook of child language*. Oxford: Blackwell.

Gayraud, Frédérique. 2000. *Le développement de la différenciation oral/écrit vu à travers le lexique*. Université Lumière - Lyon 2. Science du Langage.

Halliday, M. A. K. 1985. *Spoken and written language*. Geelong Vict.: Deakin University.

Johansson, Victoria. 1999. 'Word frequencies in speech and writing: a study of expository discourse'. In Ravid A. Aisenman (ed.), *Working papers in developing literacy across genres, modalities, and languages*, vol. I, 182-98. Tel Aviv: Tel Aviv University Press.

Johansson, Victoria. 2008. *A developmental study of text writing*. PhD Thesis. Lund University.

Lieven, E. V. M. 1978. 'Conversations between mothers and young children: individual differences and their possible implication for the study of child language learning'. In N. Waterson & C. E. Snow. *The development of communication*, 173-187. Chichester: Wiley.

MacWhinney, Brian. 2000. *The CHILDES Project: tools for analyzing talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, David, Brian Richards, Ngoni Chipere & Pilar Durán. 2004. *Lexical diversity and language development: quantification and assessment*. New York: Palgrave Macmillan.

McCarthy, Philip M. & Scott Jarvis. 2007. 'vocd: a theoretical and empirical evaluation'. *Language Testing* 24:4, 459-88.

Miller, Jon F. & Thomas Klee. 1995. 'Computational approaches to the analysis of language impairment'. In Paul Fletcher & Brian MacWhinney (eds.), *The handbook of child language*, 545-572. Oxford: Blackwell.

Read, John. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, Brian J. & David Malvern. 1997. *Quantifying lexical diversity in the study of language development*. Reading: Faculty of Education and Community Studies.

Strömqvist, Sven, Victoria Johansson, Sarah Kriz, Hrafnhildur Ragnarsdóttir, Ravid Aisenman & Dorit Ravid. 2002. 'Toward a crosslinguistic comparison of lexical quanta in speech and writing'. *Written language and literacy* 5, 45-67.

Ure, Jean. 1971. 'Lexical density and register differentiation'. In G. E. Perren & J. L. M. Trim (eds.), *Applications of linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, 443-452. Cambridge: Cambridge University Press.

Ure, Jean & Jeffrey Ellis. 1977. 'Register in descriptive linguistics and linguistic sociology'. In Oscar Uribe-Villegas (ed.), *Issues in sociolinguistics*, 197-243. The Hague: Mouton.

Vermeer, Anne. 2000. 'Coming to grips with lexical richness in spontaneous speech data'. *Language Testing* 17, 65-83.

Wengelin, Åsa. 2002. *Text production in adults with reading and writing difficulties*. Department of Linguistics, Göteborg University.

Wolfe-Quintero, Kate, Shunju Inagaki & Hae-Young Kim. 1998. *Quantifying lexical diversity in the study of language development*. Hawaii: Second Language Teaching & Curriculum Center. University of Hawaii.