Early Nouns in Bilingual Acquisition: a Test of the Separate Development Hypothesis *

Christina Schelletter[1], Indra Sinka[1] and Michael Garman[2]
School of Humanities, The University of Hertfordshire[1],
School of Linguistics & Applied language Studies, The University of Reading[2]

Abstract. In previous work (Sinka & Schelletter 1998) we have addressed the morphosyntactic development of two bilingual children and the issues raised by the opposition between the Single System and the Separate Development hypotheses. Interactions between the two language systems were found to be very rare, consistent with the Separate Development Hypothesis. This is further underlined by the developmental lead-lag pattern evidenced in the emergence of Functional Categories.

More recently, (Sinka, Garman & Schelletter 2000), we have supplemented our investigation of early grammatical development by using a lexical profiling approach to focus on the evidence from the acquisition of main verbs. Results suggest that the lead-lag order of development for the two languages in each child is the same as for the grammatical system, and each bilingual child appears to be developing the system of main verbs independently for each language, although there are some commonalities across languages, e.g. in the development of the verb 'to be'.

In this paper we extend the lexical profiling approach to the analysis of noun vocabulary. We look at the general characteristics of types and tokens, and then consider more fine-grained analysis of the nouns used by each child in terms of grammatico-semantic categories. The findings will be discussed in relation to the Separate Development Hypothesis.

1. *Introduction*

Previous analyses (Schelletter, Sinka & Garman 1999, Garman, Schelletter & Sinka 1999) have established an order of development for functional categories (FCs) in our longitudinal bilingual data on Sonja, acquiring German/English between 1;11-2;8, and Maija, acquiring Latvian/English, 1;3-1;11: Latvian leads German, and English is the lag language for each child, consistent with the degree of morphological marking in each language. We have interpreted these results in relation to the nature of early grammatical development.

* This research was supported by the ESRC, Grant No. R000222072.

We have also explored possible links between the lexicon and syntax (Sinka, Garman & Schelletter 2000), and specifically, whether the emergence of the grammatical categories of Tense/Agreement (T/A) is associated with appropriate vocabulary development, in the form of the category of main verbs. The framework for that investigation is the lexical profiling approach developed in Garman (1995), which provides a format for displaying quantitative and qualitative aspects of the full vocabulary of expressive language samples. Among other measures, the lexical profile provides type-token ratios (TTRs) for these form classes, as well as the more traditional TTR for the whole vocabulary. To control for effect on TTR of sample size (Richards 1987), this profile uses fixed (and rather small) size samples that are realistic within the clinical linguistic context in which the profile procedure has been developed - 250 continuous word tokens.

This approach has revealed that the lead-lag order of development for the two languages in each child is the same for both grammatical and lexical systems. A further finding is that each of our bilingual children appears to be developing the system of main verbs independently for each language, although there are some commonalities across languages, e.g. in the development of the verb 'to be'. Most importantly, the class of main verbs is shown to be non-uniform in the type-token characteristics of its members, calling into question the assumption that main verbs are all open-class. Indeed, it is further suggested that the findings challenge the traditional equation of set-theoretic concepts such as 'open'- 'closed' classes of vocabulary with distributional-linguistic concepts such as 'auxiliary'- 'main verb'.

In this paper we extend the lexical profiling approach to the analysis of noun vocabulary, which at first sight appears to be typical of an open lexical class in showing generally high TTR values, with little evidence of a frequency differential within the class, and in being more independent of morpho-syntax, and more dependent on situation and topic, than the class of main verbs. We look at the general characteristics of types and tokens, and then consider more fine-grained analysis of the nouns used by each child in terms of grammatico-semantic categories. The findings are then discussed in relation to the Separate Development Hypothesis (De Houwer 1990).

If noun vocabulary is not 'plugged in' to morphosyntactic development, and if it is true that the lead-lag developmental ordering is indeed, as we have argued, based on morphological marking, it would follow from this that noun deployment should be much less likely to show the lead-lag developmental effect between the three languages.

2. *Method*

2.1 Data

Tables 1a and 1b show the sampling distribution for complete and intelligible utterances. Sessions were targeted on specific participants/languages, and relatively few non-target or mixed utterances were elicited. For Sonja, practical difficulties meant that sampling in English was less continuous through the observation period than for German. It should be stressed that each child had equivalent, and considerable, exposure to each language.

Table 1a. Sonja, English and German utterances by sessions

Age	English sessions			German sessions		
	English	German[a]	Mixed[b]	German	English[a]	Mixed[b]
1:11				75	9	3
2;0				226	13	18
2;1				177	5	16
2;2	113	13	17	191	7	14
2;3	150	8	6	207	7	12
2;4				264	10	2
2;5				274	0	5
2;6	229	7	7	180	0	3
2;7	247	4	20	196	0	2
2;8	110	0	4			
Total	849	32	54	1790	51	75

Table 1b. Maija, English and Latvian utterances by sessions

Age	English sessions			Latvian sessions		
	English	Latvian[a]	Mixed[b]	Latvian	English[a]	Mixed[b]
1;3	132	11	1	123	6	0
1;4	131	53	4	128	2	0
1;5	96	40	5	199	13	3
1;6	167	81	3	314	16	8
1;7	150	19	5	201	4	2
1;8	180	39	0	202	2	0
1;9	153	52	7	207	5	2
1;10	221	17	7	252	2	1
1;11	251	3	3	205	0	4
Total	1481	315	35	1831	50	20

[a]non-target language utterances

[b]utterances showing lexical or syntactic mixing

2.2 Procedure

Data were analysed in the Systematic Analysis of Language Transcripts programme (SALT: Miller & Chapman 1993).

We first established cuts of the data samples that each consisted of 250 word tokens, starting from the first word, and finishing with the nearest word to 250 that concluded an utterance. Inflected forms are treated as tokens of a single type. So are alternative realisations of words such as yes, until and because. We excluded words in mazes, but included words from incomplete or partially unintelligible utterances.

We then tagged all nouns, including those whose inflectional ending was incorrect (missing or the wrong form) but whose stem form was not in doubt. We entered all tagged nouns into a spreadsheet for calculation and display of types and tokens.

3 *Results*

3.1 General vocabulary measures - all types

We look first at the most general vocabulary measure, the type-token ratio. Since each sample is 250 word tokens in length, we present in Fig. 1 the dependent measure of the number of all word types per sample, for each language and each child.

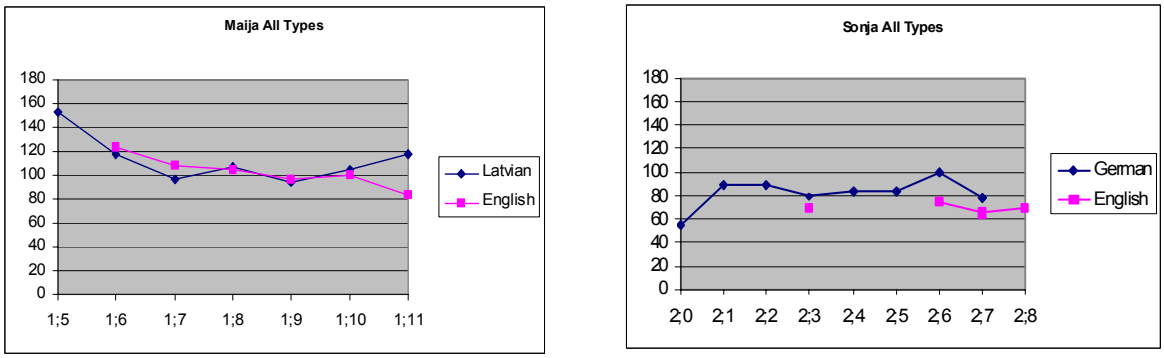

Fig.1. The number of all word types per 250 tokens, for each language of each bilingual child.

It is notable how similar the two languages are for each child, on this measure, and how distinct the children are. The mean TTR for all samples for Maija's

Latvian is 0.45, while for her English it is 0.41. By contrast, Sonja's mean TTR is 0.33 for German and 0.28 for English. It is clear that the measure is relatively stable for each language and child across the period of observation. We thus have a robust difference in vocabulary deployment between the two children.

Such a global measure of vocabulary, however, potentially obscures considerable variation among constituent word classes, and our purpose in this paper is to examine the noun class in more detail.

3.2 Noun vocabulary: all noun types

Fig. 2 presents the number of all noun types per 250 all-word tokens, by sample. It should be noted that these are not a direct reflection of noun class TTRs, since that would require plotting all noun types over all noun tokens. The proportion of noun tokens in the samples is a dependent variable of some interest, but in this paper we are are concerned with other issues.

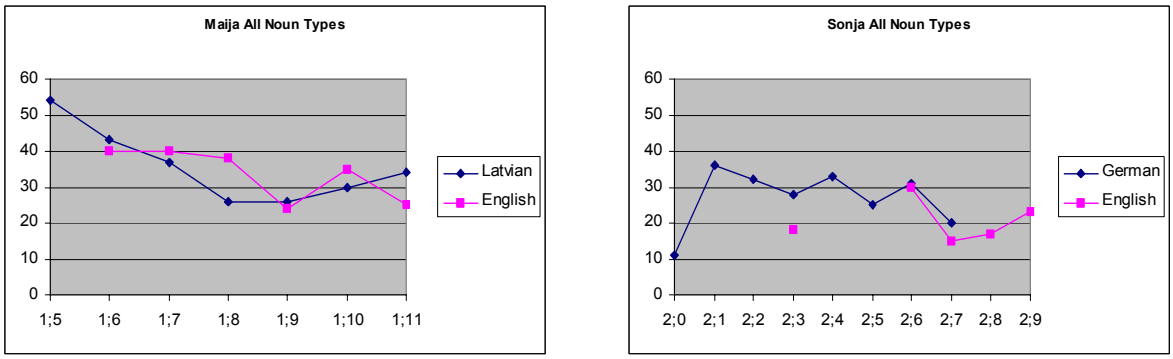

Fig. 2. The number of all noun types per 250 tokens, for each language of each bilingual child.

For Maija, noun types constitute about one third overall of all word types, with a suggestion of a developmental trend from higher to mid range values in the period of observation. Sonja shows a rather different pattern in the earlier samples, but the two children show a degree of convergence in later samples.

3.3 Noun vocabulary: all noun tokens

Next, we look at the noun tokens per 250 all-word tokens: see Fig.3. The Latvian and English data show impressive agreement in Maija's data; and as far as comparisons are possible in Sonja's records, there are some suggestive indications. Moreover, each child shows broadly similar patterns over the

period of observation, with higher values gradually decreasing to mid-range, though not continuously.

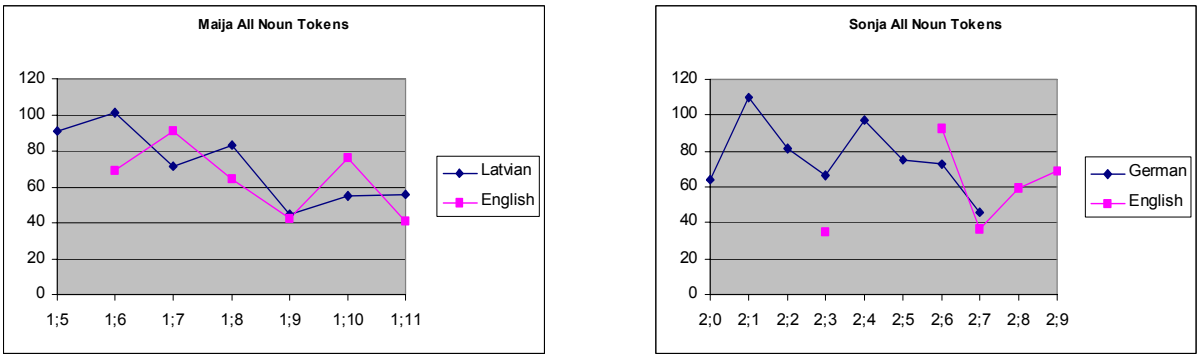

Fig. 3. The number of all noun tokens per 250 all-word tokens, for each language of each bilingual child.

3.4 Noun vocabulary: all nouns, types/tokens

Now we are in a position to look at TTR values for the noun class. Nouns are traditionally regarded as forming an open class, and hence are expected to have relatively many types with few tokens, contributing to a high proportion of types to tokens. See Fig. 4 for Maija.

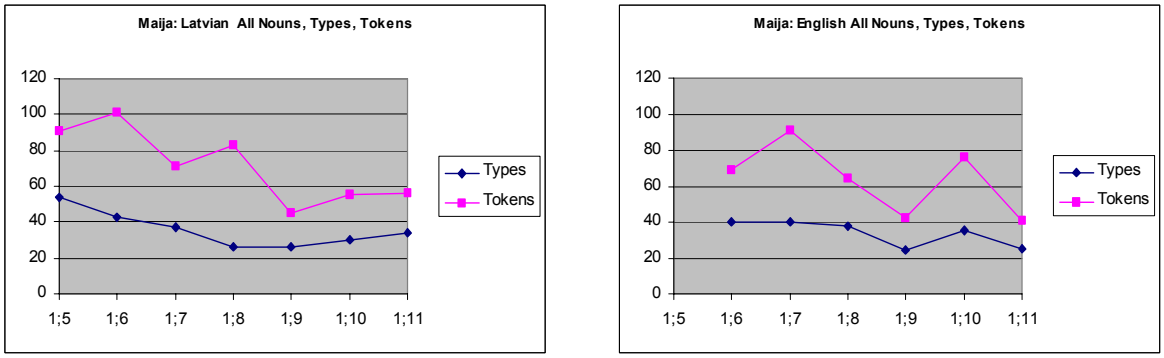

Fig. 4. All noun types and tokens, for Maija's Latvian and English.

We need to bear in mind that the TTR values are based on variable numbers of tokens (see our remark above on the dependent nature of this variable), but we feel that there are good reasons for looking at the data this way, as long as the baseline tokens are kept in view.

For this reason, we do not present just TTR values, but put the noun types and tokens in relation to each other, by sample, for each language and child. The mean TTR value is actually lower than expected, at 0.50 for Latvian, and is

only slightly higher in English at 0.53. Developmentally, tokens in each language generally decrease, although not continuously; in Latvian, types decrease initially, then stabilise, in English they decrease at 2 points, 1;9 and 1;11, reflecting sample-specific effects. Overall, the two languages show a similar pattern.

Sonja's data is presented in Fig. 5. The mean TTR value is even lower than for Maija, at 0.36 in German and 0.38 in English. Thus we observe again a pattern of considerable between-child differences, and within-child similarities between each of the languages.

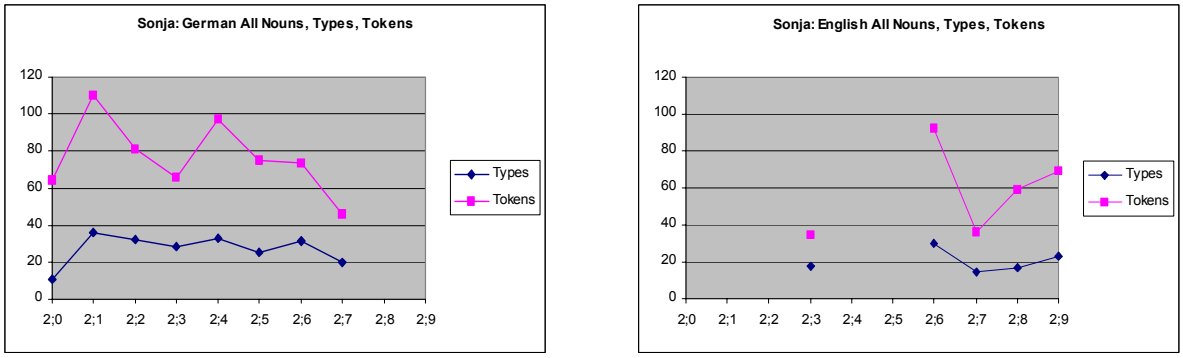

Fig. 5. All noun types and tokens, for Sonja's German and English.

Developmentally, Maija's data shows a non-continuous decrease in types and tokens over the period of observation, and a tendency for tokens to decrease in relation to types over time. This yields higher TTR values in the later samples, as we might expect: the noun class is becoming more truly open, on this measure. Sonja's German shows a similar pattern from the sample at 2;1. The English data is rather fragmentary, but appears to be complex, showing convergence followed by divergence.

The surprisingly low overall noun class TTR values in each language are most naturally interpreted as reflecting a compromise between higher and lower frequency noun types within the class. In this respect, an obvious possibility that we should consider is that we have inadvertently contributed to this state of affairs by including two very different types of nouns in the analysis thus far - proper and common. Proper nouns are thought to be used a lot by young children, because they have clear unique referents; so it is possible that, in the particular situations of the samples, where the same participants may be called by name a number of times, proper nouns exhibit higher token frequencies overall than common nouns. This suggests in turn that noun class development and deployment may be more truly measured by considering just common nouns (see also Bates, Bretherton & Snyder 1988). We address this issue in

the next section.

3.5 Common nouns, types/tokens

Accordingly, we excluded all proper nouns from the samples and recalculated the data. For brevity, we shall present just the counterpart to the types and tokens analysis of Figs 4 and 5, here presented in Figs 6 and 7 respectively.

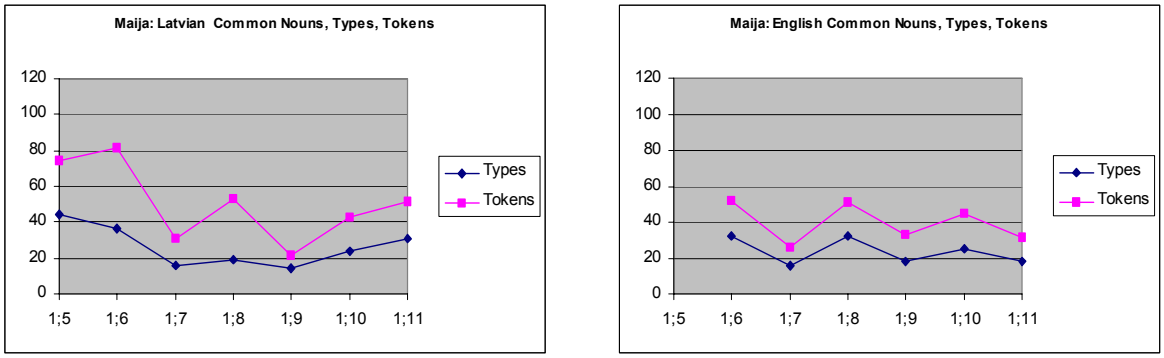

Fig. 6. Common noun types and tokens, for Maija's Latvian and English.

It can be seen that the types and tokens values are closer in Fig. 6 for each language than in Fig 4. Nevertheless, the overall TTR vaules are still quite close: 0.54 for common nouns in Latvian (against 0.50 for all nouns), and 0.59 for English (against 0.53 for all nouns). It would seem that including proper nouns in the analysis is not responsible for lowering the TTR values significantly in Maija's data.

Sonja's data is presented in Fig. 7. The German and English noun types are reduced by an average of 7 per sample - but particularly in sample 2;6.

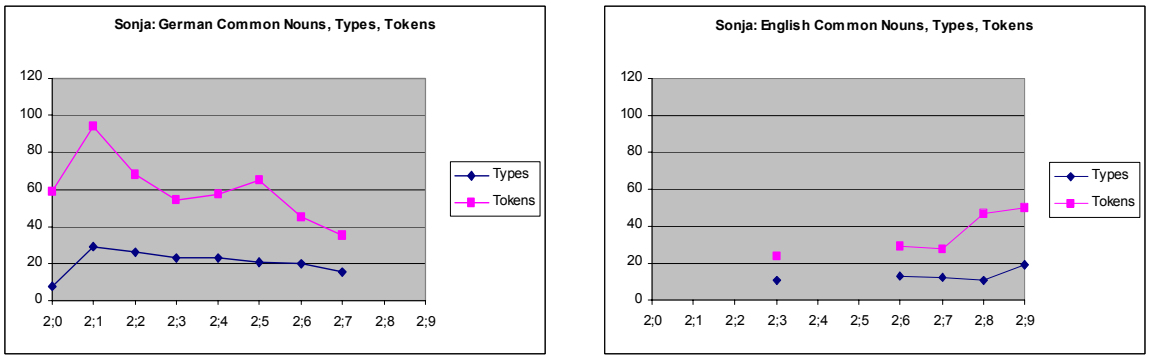

Fig. 7. Common noun types and tokens, for Sonja's Latvian and English.

The German noun tokens are reduced by about one quarter on average,

while English noun tokens are reduced by more than one third on average. In spite of these differences, the effect of removing proper nouns from the analysis has no effect on German TTR, which remains at 0.36 for all nouns and common nouns; likewise, there is also no effect on overall English TTR, which stays at 0.38 for all nouns and common nouns.

3.6 English one

There is one further issue to note here: the existence in the English noun class of certain relatively frequent noun types such as *thing* and *one*, as in phrases like *this thing* and *that one*. This is certainly an important feature of adult conversational data in English, and it has also been detected in some child language data that we have looked at. In the present data sets, *thing* is notable by its absence, however, possibly indicating that it is not characteristic of early noun vocabulary. As for the pro-noun *one*, while it is used by both Maija and Sonja, it is not paralleled by any corresponding element in either the Latvian or the German data, for the very good reason that neither of these languages has such a form. We have therefore excluded it from our analyses here. We have to bear in mind that the English TTR values would be reduced slightly if it were included.

On the basis of the analyses presented so far, we conclude that there is robust evidence of rather mixed TTR values within the central noun class, in each of the languages controlled by these bilingual children. In order to investigate the nature of the mix, we now look at the internal composition of the noun classes.

3.7 Noun frequency profiles

We can show the complex nature of the noun class for each child and language by means of noun frequency profiles that are derived as follows. For each sample, the nouns are ordered by token frequency, from highest to lowest. For all the samples for each language and child, the average token frequency values are then calculated for each of the first 10 frequency ranks. The first 10 ranks are chosen because they contain most of the variance. The result is the noun class frequency spectrum for the language-child pairing. Fig 8 shows the averaged frequency spectra for Maija's Latvian and English noun classes (this and subsequent analyses are carried out on common nouns only).

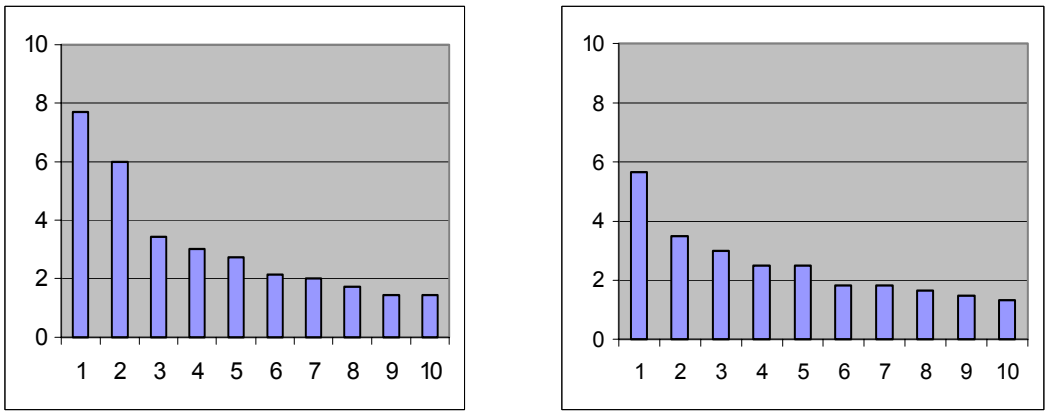

Fig. 8. Common noun frequency spectra averaged across all samples, for Maija's Latvian and English.

We can see that Latvian common nouns show a greater frequency differential among their first 10 ranks than do the English; but if English noun-use of *one* were incorporated, the English frequency differential would be greater, and more in line with the Latvian.

Sonja's data shows a similar pattern, as can be seen from Fig. 9. The steepness of the gradient is particularly apparent in the first two ranks in the German data.

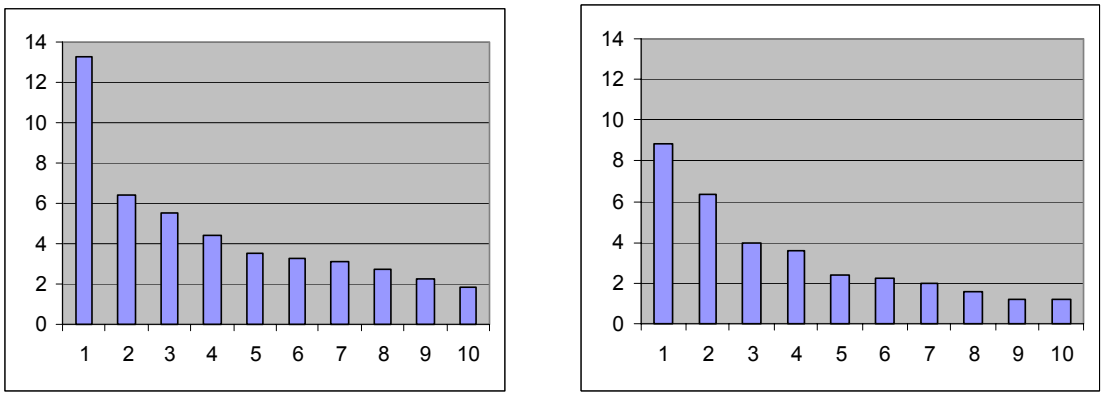

Fig. 9. Common noun frequency spectra averaged across all samples, for Sonja's German and English.

The frequency profiles we have looked at in this section reflect the result of considerable variation across samples. Inspection of these (which we do not have space to illustrate here) reveals variation from samples that have virtually flat profiles with low token values for all types (the typical pattern of the open word class), to those that have markedly steep gradients. This variation appears to show no developmental pattern, but rather the operation of little-understood contingent factors on noun deployment in individual situations.

A further feature of these profiles is that they reflect frequency ranks rather than particular nouns, and therefore they cannot tell us whether it is the case that the same nouns will be the most frequent, or among the upper frequencies, in more than one sample. They are therefore highly abstract constructs, and we now ask: What nouns inhabit the higher frequency ranks?

3.8 Individual nouns in the higher frequency ranks

For this phase of the analysis, we are interested in whether certain nouns can be identified across samples as high frequency nouns in the child's vocabulary, at least for certain stages of development. The traditional view is that noun vocabulary is highly sensitive to topic and situation, as opposed to close class elements, which are much more determined by the grammatical properties of the language and hence relatively invulnerable to topic and situation effects. If there are any stable high frequency nouns, we would want to know whether they share any common characteristics, such as reflecting the child's favourite semantic categories, and whether they are common to each language, or even each child. For this purpose, we look just at the nouns in the first four frequency ranks, since these ranks contain the high frequency nouns that are most likely to enable us to answer our questions. Accordingly, Tables 2-5 present the data for each child-language pairing.

Table 2 shows the Latvian nouns in ranks 1-4. In this and the other tables, each distinct noun type within a rank is set on its own line, so that repeating types can be easily detected. We note that *kāja*, 'leg' appears at Rank 1 in samples 1;5 and 1;6, but that thereafter no instances of predictable high frequency nouns are found. The same item turns up at Rank 4, sample 1.10, but otherwise each rank and sample contains its own noun types.

We further tested the nouns for semantic category by classifying them according to the categories of the MacArthur Communicative Development Inventories (CDI; Fenson, Dale, Reznick, Bates, Thal & Hartung 1994, Dale & Fenson 1996). The result is that there is no detectable effect of semantic class: the upper ranks are heterogeneous in respect of semantic classes.

We carried out a similar analysis for Maija's English, in Table 3, with similar results. Only *book* is found in more than one sample, at Rank 4, and there is considerable semantic heterogeneity among the nouns in the ranks.

We also tested the nouns for similarity of form across the two languages: it is possible that one determinant of a 'favourite' noun in early vocabulary is the degree to which it can be used in either language. In this respect, we expected the Latvian-English pair to provide Maija with a smaller range of 'Euro-stock'

items than Sonja had from the German-English pair. The nouns that have corresponding forms in either language are marked with an asterisk; we prefer not to use the term 'cognate' since children might trade on accidental form similarities. The Latvian data has *album* at Rank 1, sample 1;7, and *mikrofons* at Rank 4, sample 1;6, and neither of these has its corresponding English form in Table 3. Conversely, Table 3 has a number of candidates: at Rank 1 *baby*, sample 1;7; Rank 2, *giraffe*, sample 1;9, *cat*, sample 1;11; Rank 3 *kangaroo*, sample 1;8, *crocodile* sample 1;11, and Rank 4 *apple*, sample 1;8 and *baby*, sample 1;11. None of these has its corresponding form in Table 2, however. We conclude that form similarity is not a determinant of noun frequency for Maija after all.

Table 2. Maija, Latvian nouns, ranks 1-4, by sample.

	1;5	**1;6**	**1;7**	**1;8**	**1;9**	**1;10**	**1;11**
Rank 1	kāja 6 *leg*	kāja 8 *leg*	*album 16	amur 16 *hammer*	burtinu 3 *letters*	berni 7 *children*	punktiņš 5 *dot, spot*
Rank 2	cūciņa 5 *piglet*	kurpites 7 *shoe*	virins 5 *man*	pilite 13 *duck*	maja 3 *home*	podzinas 5 *butter*	bumbiņa 4 *ball*
Rank 3	gulbīši 4 *swans*	kreklins 6 *shirt*	cepure 2 *hat*	bildites 4 *picture*	meteli 2 *coat*	gaismina 3 *light*	papīrīts 3 *paper*
Rank 4	cuku 4 *pig*	*mikrofons 6	pucites 2 *owl*	biksem 2 *trousers*	rokasspradziti 2 *bracelet*	kāja 3 *leg*	ugunsdzēsēju 3 *fireman*

Table 3. Maija, English nouns, ranks 1-4, by sample

	1;5	1;6	1;7	1;8	1;9	1;10	1;11
Rank 1	-	flower 3	*baby 5	clock 5	seesaw 9	glasses 8	book 4
Rank 2	-	hand 3	car 3	blueballoon 4	*giraffe/s 3	book 4	*cat 4
Rank 3	-	nose 3	pig 3	*kangaroo 4	bit 2	house 3	*crocodile 3
Rank 4	-	piglet 3	book 2	*apple 3	book 2	pussycat 3	*baby 2

Sonja's data is presented in Tables 4 and 5. In Table 4, *miau* crosses samples 2;1-2;2-2;3 at Rank 2, and *bett* crosses samples 2;3-2;4, at Rank 3, and is also found at Rank 4 in sample 2;7. Apart from these, however, each rank and sample shows its own noun types, and there is considerable semantic heterogeneity.

Table 4. Sonja, German nouns, ranks 1-4, by sample

	2;0	2;1	2;2	2;3	2;4	2;5	2;6	2;7
Rank 1	*pizza 39	*creme 19	bauch 15	*junge 7	babyfrisur 6	tier 8	*teddy 7	löffel 5
Rank 2	messer 6	*miau 10	*miau 9	*miau 7	*baby 5	stift 8	*wauwau 2	mütze 4
Rank 3	gabel 6	*keks 9	*mikrofon 5	*bett 5	*hamster 5	zahnpasta 6	*bett 4	*junge 4
Rank 4	*buggy 4	deckel 6	gabel 4	tisch 4	mädchen 5	*brumbrum 5	gummibärchen 4	*bett 3

Nouns that show form similarities with English nouns are quite numerous: at Rank 1, *pizza*, *creme*, *teddy*; at Rank 2 *miau*, *baby*, *wauwau*; at Rank 3 *mikrofon*, *bett*, *hamster*; and at Rank 4 *buggy*, *brum-brum* and *bett* (again). The English data in Table 5 shows that *baby*, *bett* and *buggy* are matched by their corresponding English forms.

Table 5. Sonja, English nouns, ranks 1-4, by sample

	2;0			**2;3**	**2;6**	**2;7**	**2;8**	**2;9**
Rank 1	-			*baby 10	*baby 7	boat 5	*baby 13	turn 9
Rank 2	-			*bed 5	*buggy 4	bean 5	lorry 13	breakfast 5
Rank 3	-			*shoe 4	*chocolate 3	boy 4	eyes 5	bike 4
Rank 4	-			milk 4	icecream 2	tea 4	sugar 5	chair 3

Conversely, the English data shows in addition the forms *shoe* and *chocolate* at Rank 3, and *milk* and *tea* at Rank 4, which have corresponding forms in German but which are not found in Table 4. We conclude that semantic categories do not predict noun occurrence by rank in the German or English data, and that form similarity is not a determinant of noun deployment in Sonja's data.

4. *Summary and Conclusions*

We may summarise as follows:

- Nouns as a lexical class are less 'plugged in' to grammar than are main verbs
- Because of this, no lead-lag effect between the languages in this study emerges in respect of noun development
- Averaged over time, nouns exhibit a complex frequency profile, with a fairly steep gradient between the highest 4 frequency ranks
- As a result the noun TTR value is lower than would be expected for an open class
- However, their frequency profiles are highly variable between samples

- This suggests nouns are particularly susceptible to sampling factors such as topic and situation
- There are considerable within-child similarities in noun development between the two languages controlled by each child
- There are considerable differences between the children in noun development, as reflected in TTR values
- There are no stable frequency characteristics that attach to nouns across samples
- They appear therefore to be selected by communicative demands
- They are selected independently in each language
- They do not reflect semantic preferences
- They are not influenced by the existence of cross-linguistic form similarities.

We conclude that the broad inter-language similarities observed in each child are (at least) consistent with the Separate Development Hypothesis, in that each child brings her own development style to each of the languages she controls. More compelling support is found in the lack of predictably frequent individual nouns, and in the lack of any strategy that relies on the use of nouns that are similar in form between the languages. Each of these observations tends to the view that that noun development is separately controlled for each language;

We also conclude that the lack of a lead-lag effect for nouns underlines the importance of this effect in the analysis of functional categories and main verbs; i.e. its presence there is not just the result of some general property of our data, since in that case it would be expected to appear with noun development also. It is rather the result of a specific property, which functional categories and main verbs share, and which nouns do not.

References

Bates, E., Bretherton, I. & Snyder, L. (1988). *From first words to grammar: individual differences and dissociable mechanisms*. Cambridge: C.U.P.

De Houwer, A. (1990). *The acquisition of two languages from birth: a case study*. Cambridge: C.U.P.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers* **28**. 125-127.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Hartung, J. (1994). Variability in early communication development. *Monographs of the Society for Research in Child Development*, 59, serial no. 242.

Garman, M. (1995). Lexical profiling of spontaneous speech in children and adults. In P. Kerswill, R. Ingham, Y. Huang & L. Shockey (eds), *Working papers in linguistics*, Vol 2, University of Reading, 1-22.

Garman, M., Schelletter, C. & Sinka, I. (1999). Three hypotheses on early grammatical development. In Perkins, M. & Howard, S. (eds) *New directions in language development and disorders*. New York: Kluwer/Plenum. 73-84.

Miller, J. & Chapman, R. (1993), *Systematic analysis of language transcripts.* Madison, WI: University of Wisconsin.

Richards, B. (1987). Type-token ratios: What do they really tell us? *Journal of Child Language* **12**. 201-209.

Schelletter, C, Sinka, I. & Garman, M. (1999). Latvian/English and German/English bilingual acquisition: new light on Universal Grammar. *Proceedings of the first international symposium on bilingualism*, Vigo: University of Vigo, October 1997.

Sinka, I., Garman, M. & Schelletter, C, (2000). Early verbs in bilingual acquisition: a lexical profiling approach. In R. Ingham & P. Kerswill (eds), *Working papers in linguistics*, Vol 4, University of Reading, 175-187.