Cognitive Science 33 (2009) 75–103

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2008.01004.x

Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech

Sabine Stoll,[a] Kirsten Abbot-Smith,[b] Elena Lieven[a,c]

[a]*Max Planck Institute for Evolutionary Anthropology, Leipzig*
[b]*School of Psychology, University of Plymouth*
[c]*School of Psychological Sciences, University of Manchester*

Received 19 September 2006; received in revised form 14 March 2008; accepted 18 April 2008

Abstract

This study investigates the child-directed speech (CDS) of four Russian-, six German, and six English-speaking mothers to their 2-year-old children. Typologically Russian has considerably less restricted word order than either German or English, with German showing more word-order variants than English. This could lead to the prediction that the lexical restrictiveness previously found in the initial strings of English CDS by Cameron-Faulkner, Lieven, and Tomasello (2003) would not be found in Russian or German CDS. However, despite differences between the three corpora that clearly derive from typological differences between the languages, the most significant finding of this study is a high degree of lexical restrictiveness at the beginnings of CDS utterances in all three languages.

Keywords: Child-directed speech; Language learning; Russian; German; English; Lexical restrictedness; Lexical strings; Learning strings

1. Introduction

A wide range of studies has shown that infants are capable of distributional learning in the first year of life with the factors to which they are sensitive changing as they develop. However, though the role of the language that children hear has always been central to theoretical arguments about how language can be learned, we know rather little how the input is in fact structured, especially for languages other than English. Following Chomsky, many

Correspondence should be sent to Sabine Stoll, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. E-mail: stoll@eva.mpg.de

have claimed that children could not learn the underlying structure of language from what they hear (Chomsky, 1959; Crain, 1991; see also Pinker, 1989). Two main reasons were suggested for this, first that the input is chaotic and sometimes ungrammatical and, second, that the surface form of utterances does not provide a guide to their underlying structure. Research in the generativist tradition has consequently assumed that children use innate syntactic representations to construct the grammar of the language that they are learning. However, all learnability theorists have of course assumed that children must in addition use the mechanism of distributional learning when learning the particular grammatical categories of their language (e.g., Pinker, 1989). This is because languages differ in the range of categories they manifest (for instance, some languages do not have a clear adjective category) and in the boundaries of these categories. The goal of the present study is to investigate how ''chaotic'' the input that children receive really is. To do this we analyze the input of three typologically different languages. Since distributional learning depends on identifying patterns, we analyse the extent to which the input of these three languages with different grammatical structures contains repeated strings of words. In this introduction, we first briefly review the role of distributional learning in early language development before summarizing the typological differences between the three languages under study: Russian, German, and English.

We know that by the time infants are about 7 months old, they can identify words they have heard in isolation in strings of speech (Jusczyk, 1997), and by 11–12 months of age they have been shown to be able to discriminate patterns in strings of artificial syllables (Gómez & Gerken, 1999). By about 14–16 months, German infants are able to categorise a novel word following a determiner as a ''noun'' in a preferential head-turn procedure (Höhle, Weissenborn, Kiefer, Schulz, & Schmitz, 2006). These developing abilities depend on the distributional probabilities of a number of interacting features in the input, for instance, prosodic structure, number of syllables, and phonetic weight. Modeling studies have contributed greatly to our understanding of the interacting role of these factors by varying the nature of the input to the model and seeing the difference this makes to the kinds of clustering that emerges. Although earlier studies were usually done with English input (Brent & Siskind, 2001; Redington, Chater, & Finch, 1998), there have been more recent studies that have examined child-directed speech corpora in languages other than English. This is obviously important to do since English is a strictly word-ordered language and it is possible that while this could give rise to putative categorization from English input, languages with much freer word order or more morphology would not yield similar results. Two recent studies suggest that this is not the case (Monaghan, Christiansen, & Chater, 2007; Chang, Lieven, & Tomasello, in press). However, it is important to note that in both cases the models were learning from more than one type of distributional information. In the Monaghan et al. study, the model learned to distinguish open from closed class words and nouns from verbs in English, Dutch, French, and Japanese using phonological cues and distributional associations between successive words. However, there was an interaction between these two types of cues for the different languages which showed that when distributional cues were less reliable, phonological cues were stronger. The study by Chang et al. (in press) was based on Chang's two-pathway model, (Chang, 2002). In this study one

measure was of the overall order of words in the utterance and the second was of the immediate adjacencies between words. It was found that, depending on the characteristics of the language being analyzed, overall order played a more important role in more analytic languages such as English, whereas both measures were equally important in more synthetic languages such as Estonian, Hebrew and Hungarian. Thus, the distributional information available in the input can cluster word categories with high statistical success and not only for English, but this works much better when a number of factors are combined.

It has also been suggested that distributional learning might be used not just to learn grammatical categories but also to approximate grammatical rule-learning (e.g., Lieven, Pine, & Baldwin, 1997; see also Perruchet & Pacteau, 1990). A vast body of research on artificial grammar learning has found that adult humans are indeed capable of implicitly learning quite complex systems through what is essentially distributional learning (e.g., Cleeremans, Destrebecqz, & Boyer, 1998). Infants and even nonhuman primates have been found to show similar capabilities (e.g., Gómez & Gerken, 1999; Hauser, Weiss, & Marcus, 2002). A number of models have shown how distributional learning over a range of simpler sentences might enable further parsing of the input in terms of more complex syntax (Lewis & Elman, 2001; Mintz, 2003).

One theoretical framework in which distributional learning is also assumed to play a crucial role in the learning of higher-order grammatical structures is the constructivist approach to child grammatical acquisition (e.g., Lieven et al., 1997). In this approach children are argued to initially represent grammatical constructions in terms of ''frames'' consisting of partially rote-learned lexical material and ''slots,'' where particular categories have been abstracted. As an example, a child might initially learn English wh-questions by learning lexically based frames such as *What does PERSON ACTION?* where *what does* is a lexically specific frame, the slot *PERSON* allows any term referring to a person to be inserted, and the slot ''ACTION'' allows a range of action verbs to be slotted in.

There is both experimental and naturalistic support for this constructivist proposal (e.g., Lieven, Behrens, Speares, & Tomasello, 2003; see Tomasello, 2003 for a review; Wilson, 2003). In regards to the acquisition of wh-questions, for example, 4-year-old English children have been found to make many more noninversion errors when asking questions beginning with *What do…?* than with *What does…?* (Ambridge, Rowland, Theakston, & Tomasello, 2006). This indicates that English-speaking children are not initially learning *WH-word AUXILIARY SUBJECT... ?* as an ordering of abstract categories of elements, but rather they are learning patternings of particular words such as *what* and *does*. Thus, it may be the case the children are able to break into syntax without needing innate syntactic representations.

The next question is why certain lexical strings end up as ''frames'' and others as ''slots.'' Part of the solution may be simply the relative token frequency of particular lexical ''chunks'' such as *what do…* and *what does….* Another part of the solution may be that this and other frames remain fairly constant in the input, whereas the items in the slot show a relatively high type frequency. There is some empirical support for this possibility. Children's rate of auxiliary inversion errors in wh-questions has found to be determined at least in part by the frequency of particular wh-words-auxiliary combinations (e.g., *what can ...?, where*

is ...?) in the linguistic input (e.g., Rowland, 2007; Rowland & Pine, 2000). In addition, when 2½-year-olds hear the frame *He's VERBing it* with a number of different verb types, they are much better able to extend this construction to a novel verb in a posttest than if they hear a variety of noun phrases in the subject position (e.g., Childers & Tomasello, 2001). This fits with proposals that, if there are salient perceptual cues associated with particular grammatical constructions, this may assist children to acquire the constructions more quickly (e.g., Brooks, Braine, Catalano, Brody, & Sudhalter, 1993). It also fits with Casenhiser and Goldberg's (2005) study where a novel construction in which the main verb appeared with highly skewed frequencies was easier to learn. Many constructions do appear with these skewed verb frequencies—Goldberg (2006) gives the example of *give* in the double object dative. Thus, type frequency in the verb position is important—otherwise the child might learn a lexically specific formula, but if, in addition to this variation, one verb takes the lion's share, it may help in the learning of the construction together with its basic meaning.

Thus, the data from language acquisition (at least in English) would lead to the expectation that increased lexical repetition within particular grammatical constructions in the input might ease the process of acquisition. The majority of previous studies on child-directed speech have, however, tended to either look at the input in terms of particular vocabulary items (e.g., DeVilliers, 1985; Hoff, 2003; Naigles & Hoff-Ginsberg, 1998) or have looked at syntax at an abstract (non–lexically based) level (e.g., Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002; Newport, Gleitman, & Gleitman, 1977; Wells, 1981).

One exception to this is the study by Cameron-Faulkner, Lieven, and Tomasello (2003). They analyzed the CDS of 12 English-speaking mothers first into broad categories of syntactic construction types and then into ''frames'' comprising the first one, two, or three words of the mothers' utterances. A frame was defined as any initial word or sequence of words that was repeated four or more times in the speech of one mother, for instance ***You —*** was a one-word frame in intransitives, ***I think —*** was a two-word frame in complex constructions, and ***That's a —*** was a three-word frame in yes-no questions. Cameron-Faulkner et al. (2003) calculated what proportion of the mothers' utterances these frames accounted for. The main analyses in this study were within construction types: thus, 20 lexically specific strings accounted for 67% of all questions, eight for 77% of copulas, and six for 53% of imperatives. These three categories together with "fragments" made up 76% of all utterances addressed to the children (confirming the previously cited studies showing that English CDS contains a heterogeneous set of constructions).

Thus, the input English-speaking children receive contains so much lexical repetition that it can clearly be described in terms of ''slot-and-frame'' patterns. This might help the child to learn a number of aspects of her language. First, Fernald and Hurtado (2006) recently took some of the frames found by Cameron-Faulkner et al. (2003) and showed that 18-month-old infants were faster to interpret words if they were preceded by a frame (e.g., *Look at the* DOGGIE!) than if they were not (e.g., *Look!* DOGGIE!). Second, Mintz (2002) found that adults will categorize words from an artificial language into categories based on their occurrence in frames, which would imply that this might also hold for grammatical categories for children. Third, Saffran (2001) found that adults (and to some

extent 6- to 9-year-olds) can use predictive relationships between items in an artificial language to learn something approximating phrase structure rules. Finally, Bannard and Matthews (2008) in a study that compared children's repetition speed and accuracy for highly frequent four-word strings in their input with four-word strings matched for individual word frequencies, found that the children repeated the frequent strings faster and with fewer errors, suggesting that the frequent strings may indeed be represented in the child's system as wholes.

Findings like those of Saffran (2001) translate well into thinking about languages such as English, in which for example certain words such as ''the/a'' predict a following noun 90% of the time (e.g., Thorpe & Fernald, 2006). Further examples of this predictability given by Saffran (2001) include the fact that English prepositions predict following noun phrases (but not the reverse) and English transitive verbs predict a following noun phrase (but not the reverse).

The question is whether this degree of predictability in the input is restricted to English or whether we also find similar levels in other, typologically different languages. To test this we chose two other languages with varying degree of word order and morphology. Thus, in the current study, we examine German and Russian CDS to see the degree to which they show any linearly based lexical restrictiveness at the beginnings of utterances.

1.1. Typological characteristics of English, German, and Russian

English has very restrictive word order and very little morphology. It might be that the high degree of linearly based lexical restrictiveness found in English CDS (Cameron-Faulkner et al., 2003) is not reflected in the CDS to children learning typologically different languages. For instance, more word-order possibilities for particular constructions, richer inflectional morphology, and high rates of argument omission might all reduce the degree of linearly based lexical restrictiveness at the beginnings of utterances.

Russian is a language with case- and gender-marking, considerable morphology, and, in addition, the copula is not expressed in present tense. More significant for the present study, word order in Russian is syntactically flexible and largely determined by pragmatics; thus, there are few constructions with syntactically determined word order. For instance, in wh-questions, the wh-word can remain in situ and thus wh-questions can show much more variable word order than they can in English (Rojina, 2004). These word-order variations lead to many more possible word-order variants in Russian than in English or German. While a sentence like:

(1) I saw my brother.

allows no alternative word order in English, in Russian grammar, all 24 possible word orders for this sentence would be regarded, in principle, as grammatically correct. The possibility of subject omission adds another six variants to this sentence, allowing 30 possible orderings in all. Admittedly, such word-order variation is not possible for every sentence in Russian, but in general word-order variation is much less syntactically determined and more

pragmatically determined than in German or English. This should then affect the likelihood of finding lexical frames at the beginnings of Russian utterances.

In terms of syntactic description, German lies somewhere between English and Russian on a number of dimensions; it has more word-order variants than English. As a translation for example 1 above, German has three grammatically permissable word orders:

(2) (a) Ich habe meinen Bruder gesehen. *(lit: I have my brother seen)*
(b) Meinen Bruder habe ich gesehen. *(lit: my brother have I seen)*
(c) Gesehen habe ich meinen Bruder. *(lit: seen have I my brother)*

In fact the SVO order in 2a is more common in German CDS than is the OVS order shown in 2b. As an example of this, Dittmar, Abbot-Smith, Lieven, and Tomasello (2008) analyzed CHILDES data of spontaneous speech by German mothers to six monolingual normally developing children (see Szagun, 2004). They found that in 22% of transitives in German CDS the object precedes the subject as in 2b. It is also very common for the subject to follow the verb as in example 2c above and 3 below.

(3) Gestern habe ich meinen Bruder gesehen *(lit: yesterday have I my brother seen).*

A number of morphological differences across the three languages could also give rise to differences in the degree of lexical restrictiveness. English has a definite and an indefinite article, which do not vary for gender, case, or number. German marks gender, number, and case on both. English and German nouns are usually accompanied by these articles. Thus, the more rigid word order and smaller range of article forms in English suggests that the equivalent sentences might be more likely to have predictable initial strings than in German (compare examples 4 and 5):

(4) That's a X.
(5) Der/Die/Das ist ein/eine X. *(lit: that.$_{masc}$/that.$_{fem}$/that.$_{neut}$ is a.$_{masc}$/$_{neut}$/a.$_{fem}$ X)*

Russian has no articles but syntactically allows more word-order variability, so this could also give rise to a more variable set of utterances, again reducing the likelihood of finding frames. Thus, if speakers also use the word-order variants that are permissible in German and Russian syntax, these differences should be even stronger.

As well as reducing the degree of lexical specificity at the beginnings of utterances, these grammatical facts of the languages should also have an effect on the length of any repeated strings that are found. Thus, the Russian equivalent of examples 4 and 5 above is 6 below:

(6) Eto X. (or *X eto*, in certain pragmatic circumstances)

If this is found to be an initial frame in Russian CDS, the absence of determiners and of the copula in present tense will lead to a much shorter initial string than will be the case in either English or German.

These grammatical differences between the languages can lead us to two, somewhat opposed hypotheses. No one disputes that there are many grammatical differences between

the languages of the world and that this should lead to variation in the kinds of patterns that children hear. But researchers often look at the descriptive grammar of a language (i.e., what it is in principle possible to say) and derive assumptions about what children will hear, concluding from this which elements will therefore be easy or difficult to learn. We will call this approach the ''grammar-based'' hypothesis. On the basis of differences between the three languages in word-order restrictiveness and the amount of inflectional morphology, this would predict that we will find less repetition at the beginning of utterances in German than in English and even less in Russian.

However, we do not know whether actual usage, particularly in speech to children, in fact shows all grammatically permissible variants in equal proportions. Thus, the language-usage situation (the pragmatics of interacting with young children) may in fact result in certain lexical items and word orders predominating and this may skew the types of structures and elements children hear when they are starting to learn language. We will call this the ''usage-based hypothesis.'' In the present study we explore how the grammar of a language interacts with what mothers actually want or need to say to very young children by comparing CDS in Russian and German to CDS in English.

We compared the degree of lexical restrictiveness in the initial one to three words of all the utterances across each mother's whole corpus. Then, to find out whether there are language specific differences in the length of frames, we compared this for the three languages. We analyzed initial strings using the method introduced by Cameron-Faulkner et al. (2003) because (1) it allows us to compare our results with the results of this previous study and (2) because it has been shown that children pay more attention to sentence-initial than to sentence-medial words (e.g., Seidl & Johnson, 2006). We chose to analyze frames up to a maximum length of three words. In eyeballing the data prior to analysis this seemed to be the maximum number of words for which a comparison of the three languages was still useful, as will become apparent in the Results section below.

2. Method

2.1. Data

The data for this study were taken from longitudinal corpora of Russian, English, and German. All the children in the three corpora were normally developing and monolingual (age range 1;8–2;6), growing up in urban environments. The data were collected during the children's interactions with their caretakers, mostly in free play. The English data are part of the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001). We used the CDS of 6 of the 12 dyads analyzed by Cameron-Faulkner et al. (2003). The data were partially recoded and reanalyzed for the present study. Cameron-Faulkner et al. analyzed a mean of 1,400 maternal utterances per dyad (range 1,143–1,736). To match this for Russian and German, 1,400 utterances were coded per mother. The Russian data were taken from four mother–child dyads from the Stoll corpus (three boys and one girl; Stoll, unpublished data). For German we used six mother–child dyads (five girls and one boy) from the Szagun

corpus on CHILDES (Szagun, 2004). Partially unintelligible utterances, attention getters, and communicators were not coded.

2.2. Coding

As in Cameron-Faulkner et al. (2003), tag complements were coded as statements and utterances starting with conjunctions (''and,'' ''or,'' or ''but'') were coded as starting with the subsequent word. Vocatives were ignored, wherever they occurred in the utterance. Utterances that continued after interruption were coded only for the second part. If two utterances were transcribed on the same line, without any syntactic connectors, only the first utterance was coded.

The Russian data were coded by a native Russian speaker and the German by a native-German speaking English Linguistics Masters student. Points of difficulty were discussed by all coders and authors in weekly meetings. To check reliability, 200 utterances from each German- and Russian-speaking dyad were coded by a second coder, the first author. This amounts to about 14% of the data. Agreement was very good for both languages with a Cohen's kappa of .93 for German and of .89 for Russian. The English data was recoded by the main coder for German after she had coded the German data.

The goal of the frame analysis was to find out whether the three languages behave differently with respect to lexical specificity at the beginnings of utterances. To do this we analyzed repeated lexical strings at the beginning of utterances and the proportions of utterances that begin with these ''frames.'' These were coded for each mother separately. We followed Cameron-Faulkner et al. (2003) in using a criterion of four exemplars of the same initial one to three words in one mother's speech. They justified this on the grounds that a criterion of at least four exemplars in 1,400 utterances represented a ''reasonably stringent criterion within the sampling limitations that are always a feature of this research'' (p. 849).

Table 1 provides an example. If a particular mother's corpus contains *That's a girl, That's a dog, That's a flower*, and *That's your pen*, this is coded as a ***That's —*** a two-word frame.[1] If we then find another utterance, *That's a lorry*, a ***That's a***, a three-word frame is coded, and the ***That's —*** frame drops out together with ***That's your pen***. It would require another three utterances starting with either ***That's —*** or ***That's your —*** before the ***That's your pen*** utterance could be coded as belonging to a frame. Because we counted contracted forms separately, ***That's —*** is a ''two-word'' frame and

Table 1
Coding frames: an example

Utterances	Frame
1. That's a girl	That's —
2. That's a dog	
3. That's a flower	
4. That's your pen	
5. That's a lorry	That's a —

That's a — is a "three-word" frame. Equally, for German, ***Jetzt geht's —*** (literally "Now it goes—") is a "three-word" frame.[2] The number of frames and the proportion of utterances that they account for were calculated for each mother.

We were interested in whether there was a tendency for mothers to use the same frames, so if the same frame was present for at least half the mothers in each sample (i.e., three mothers for English and German and two for Russian), we identified it as a ''core frame'' (Cameron-Faulkner et al., 2003). When calculating the number of utterances accounted for by core frames, utterances falling into a core frame were counted only when they had contributed to a frame for a particular mother. The measure of core frames gives us an indication of the overlap between mothers in the utterances that they are using.[3] Finally we coded frames for whether they consisted of one, two, or three words and compared this across the three languages.

3. Results

3.1. Analysis 1

3.1.1. All frames

The left-hand group on Fig. 1 shows, for each mother, the amount of input data that could be accounted by either one-, two-, or three-word linearly based sentence-initial frames such as *Do you want…?* or *I've…* or *Did…?*. Sentence-initial frames account for a mean of 70% in Russian, 80% in German, and 86% in English of the mothers' utterances. Of course, this is only likely to help the language-learning child, if the number of frames accounting for this data is not disproportionately large. That is, the degree of sentence-initial predictability would not be very high, if the child was hearing thousands of frames. In fact, this is not the case. The mean number of frames per mother is given in Table 2. Here it can be seen that

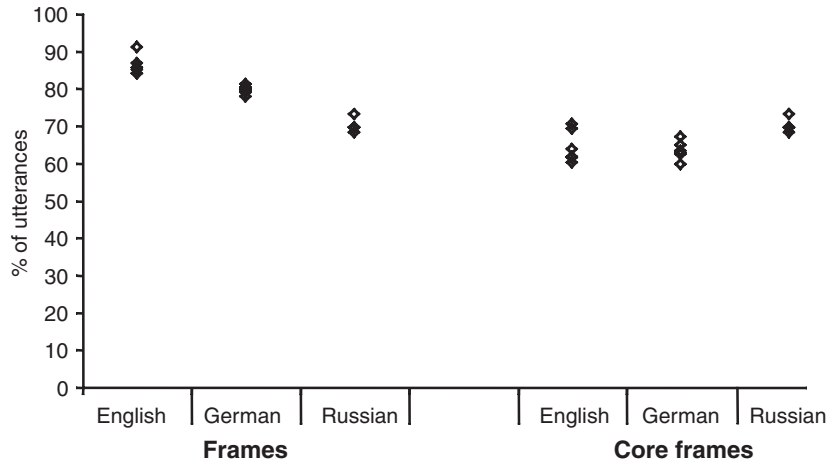

Fig. 1. Percentage of utterances of individual mothers accounted for by frames and core frames.

Table 2
Mean number of frames and core frames per mother

	Frames per Mother	Core Frames per Mother
English	143	88
German	97	58
Russian	78	45

there are 78 frames on average per Russian mother, 97 per German mother, and 143 frames per English mother. This is certainly not a large number when one considers estimates that 2-year-old children are learning between 3 and 10 words every day (e.g., Carey, 1978). In terms of precise degrees of predictability, it is probably easiest to consider these numbers in relation to our total sample. Given that a mean of 1,400 utterances was coded for each mother, each maternal frame therefore accounts for 12–13 utterances per Russian mother and 8–9 utterances for each English mother. Thus, these Russian and German mothers are similar to the English mothers in using a very high degree of linearly based sentence-initial repetitiveness at the beginnings of utterances to their children.

To make sure that these results are not produced by a small number of very highly frequent frames, Table 3 presents the data broken down for the number of items going into frames and the percentage of the mothers' utterances that this accounts for. First, there is consistency across mothers in the numbers of utterances falling into frames and the percentage of that mother's utterances that they account for. Second, while frames with 4–10

Table 3
Number of utterances in frames and percentage of utterances accounted for

Utterances Contributing to Frame	Percentage of Utterances Accounted for by Frames						
	Mother 1	Mother 2	Mother 3	Mother 4	Mother 5	Mother 6	Mean
English							
4–10	40.2	45.3	47.2	41.8	52.6	43.7	45.1
11–20	29.4	17.3	24.1	15.5	12.5	29.1	21.3
21–30	1.4	8.7	4.5	11.1	11.3	3.3	6.7
>30	13.0	19.8	9.3	18.3	9.1	8.0	12.9
German							
4–10	32.0	34.2	29.0	36.4	34.0	31.1	32.8
11–20	16.3	12.2	14.2	14.2	14.5	15.4	14.4
21–30	6.3	9.0	11.3	9.6	7.8	11.7	9.3
>30	25.7	23.5	26.7	19.4	23.6	19.8	23.1
Russian							
4–10	25.9	34.4	25.0	29.5			28.7
11–20	14.9	7.0	10.6	4.8			9.3
21–30	12.2	8.6	12.6	6.5			10.0
>30	15.1	23.2	21.3	28.8			22.1

utterances account for the highest proportion of utterances, frames with over 30 utterances are also accounting for over 20% of the data in the case of Russian and German and around 13% for English.

3.1.2. Core frames

To determine how consistent these frames were across different mothers within a language, we also looked at core frames in each language—frames which half or more of the mothers each use at least four times. The core frames of course accounted for less of the input data, but the amount accounted for was nonetheless very high. From the right-hand group on Fig. 1 we can see that the core frames accounted for a mean of 56% in Russian, 63% in German, and 64% in English of the mothers' utterances in all three languages. Interestingly, as we can see from Table 2, the actual number of core frames was much lower than for frames overall. Here it can be seen that each Russian mother used an average of 45 core frames. However, there were only 64 core frames in Russian overall (because most of the mothers used most of the same core frames). Since the core frames accounted for 56% of all the Russian mothers' utterances (a total of 2,745 out of 4,806 utterances), this means that one core frame accounts on average for 43 Russian utterances. This figure ends up being very similar for English; one core frame accounts on average for 46 English utterances.[4] It was, however, quite a bit higher for German for the following reason. Each German mother used an average of 58 core frames. However, there were only 79 core frames in German overall. Since the core frames accounted for 63% of all the German mothers' utterances (a total of 4,877 out of 7,585 utterances), this means that one core frame accounts on average for 62 German utterances.

Thus, this core frame analysis shows that in addition to the high level of lexical restrictiveness shown by each mother, there is quite some overlap in the lexical frames the mothers speaking the same language use. Moreover, each core frame has considerable predictability because they each account on average for between 43 and 62 utterances, depending on the language.

3.1.3. Differences between languages

Despite the cross-linguistic similarity in lexical repetitiveness, there may still be differences between the language corpora because Russian and German allow much more word-order variation than does English. Therefore, we examined whether the number of frames per mother and the proportion of utterances that they account for differed significantly between the three languages. This is important, since if children learning German and Russian are hearing less sentence-initial, linearly based, lexical repetitiveness, this might affect the speed or process of distributional analysis.

English has on average more frames per mother than German, which, in turn, has more than Russian.[5] A one-way ANOVA shows that the difference in the mean number of frames is significant ($F(2,13) = 34.685$, $p < .001$). Pairwise comparisons indicate that all three languages differ significantly from each other ($p < .05$). In addition, the percentage of utterances accounted for by these frames is not the same across the three languages ($F\ (2,33) = 79.137$, $p < .001$). English frames account for more of the data than do German

frames, and both English and German frames account for more of the data than do Russian frames ($p < .001$).

Likewise, the number of core frames also differed significantly across the three languages ($F(2,33) = 60.654$, $p < .001$). Pairwise comparisons again found that English mothers on average use most core frames, Russian mothers the least, and German mothers are between the two ($p < .005$). The proportion of data accounted for by Russian, English, and German core frames again differs significantly ($F(2,13) = 10.163$, $p = .002$). This time pairwise comparisons show no difference between English and German ($p = .529$), but again both English and German core frames account for more utterances than Russian core frames ($p > .005$). Thus, Russian has fewer core frames accounting for less of the data. English shows the highest number of core frames and of utterances falling into these frames. German lies somewhere between Russian and English. However, for all three languages, a high proportion of utterances in these CDS corpora can be accounted for by highly lexically restricted, linearly based sentence-initial frames.

3.2. Analysis 2: One-, two-, and three-word frames

The frames included in analysis 1 varied in length from one word (***He —***) to three words (e.g., ***That's a —***.). We also examined whether the language samples differ on this factor. The reader is reminded that, as a result of our coding system, the relative proportion of different length frames is related. If a corpus has many three-word frames, it will automatically have fewer one- and two-word frames: since each utterance is only coded once, utterances captured by a three-word frame will not be available for coding into either one- or two-word frames.

A repeated measures ANOVA with the length of frames (one-, two- and three-word frames) as the within subject factor and language the between subject factor found a significant interaction between language and length of frames ($F(4,26) = 59.1$, $p < .001$). For the ANOVA, we used log-transformed data because the error variances were not homogeneous as revealed by visual inspection. To find out about the exact differences between one-word, two-word, and three-word frames we carried out two further analyses.

First, we tested whether there is a significant difference in the number of types of frames (one-, two- and three-word) across the three languages, that is, whether one language has for instance more one-word frames than the others. Then, second, we examined whether, within a language, there are preferences for frames of a particular length, that is, whether, for instance, Russian has significantly more one-word frames than two- or three-word frames.

The numbers of frames of different length did indeed vary across languages. Fig. 2 shows the number of frames used by each mother for each language separately.[6] Looking across languages, there are significant differences in the number of frames for all three types (one-word-frames: $F(2,13) = 3.733$, $p = .052$; two-word-frames: $F(2,13) = 38.364$, $p < .001$; three-word-frames: $F(2,13) = 59{,}330$, $p < .001$). Pairwise comparisons show that Russian has significantly more one-word frames ($p < .05$) and significantly fewer two-word frames than either German or English ($p \leq .001$) with virtually no three-word frames. Thus, Russian has significantly shorter frames than either German or English. German and English show

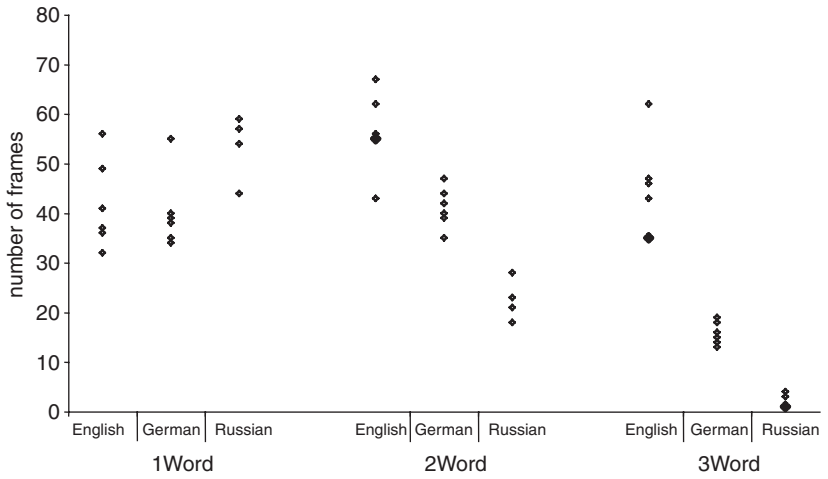

Fig. 2. Number of one-, two-, and three-word frames of individual mothers.

no difference in the number of one-word frames ($p = .724$), but English has significantly more two- and three-word frames than German ($p \leq .001$). This indicates that English has significantly longer frames than either German or Russian. In summary, English has the longest frames, Russian the shortest, and German in between.

As a second step we test whether there are differences within languages, that is, whether a particular language is significantly more likely to have one-word than three-word frames. The picture here basically ties in with the previous analysis. In all three languages there is a significant difference in the number of one-, two-, and three-word frames as shown by a repeated-measures analysis of variance ($F(2,13) = 24.970$, $p < .001$). However, pairwise comparisons reveal that the three languages differ as to which particular frame lengths tend to occur more frequently. In English there is no significant difference between one-word and three-word frames but a significant difference between one- and two-word frames ($p = .011$) and two- and three-word frames ($p = .016$). This is because the most frequent frame length in English was the two-word frame. In German there is no difference between one- and two-word frames, but there are clearly fewer three-word frames ($p \leq .001$). In Russian, there is a significant difference between all three types of frames. This is because the most frequent frame length in Russian was the one-word frame, followed by two-word frames, and hardly any three-word frames ($p < .01$).

Fig. 3 shows the mean proportion of input utterances accounted for by one- two-, and three-word frames, respectively, for each language separately. This again shows a very similar pattern of results. The proportion of utterances accounted for by one-, two-, and three-word frames is significantly different between all three languages: Russian one-word frames account for more utterances than do German one-word frames ($p = .001$), which, in turn, account for more utterances than English one-word frames ($p \leq .001$). The situation is reversed for two- and three-word frames with English accounting for significantly more utterances in both cases than German ($p \leq .001$), which again is significantly different from

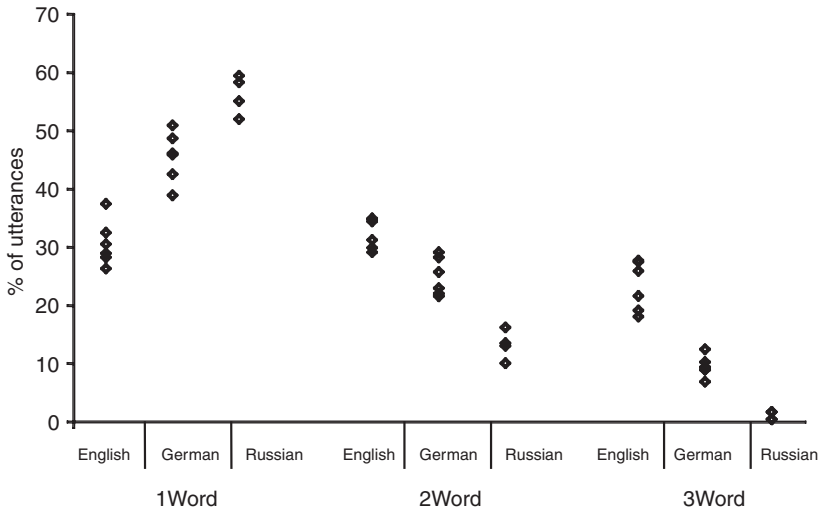

Fig. 3. Percentage of utterances of individual mothers accounted for by one-, two-, and three-word frames.

the proportion of utterances in Russian accounted for by two- and three-word frames ($p \leq .001$).

To sum up, lexical consistency at the beginning of utterances in Russian CDS is mainly carried by the first word, with Russian two-word frames accounting for only 13% of utterances. In German, it is carried either by the first or the first two words, whereas in English it is carried fairly equally between the first one, two, or three words.

3.3. Analysis 3: What's in the core frames?

To give a little more detail of what these lexically specific utterances look like, we will take a look at the core frames in each language. Table 4 summarizes the overall number of core frames and shows the percentage of core frames that start with different words.

The three languages vary strongly in how repetitive they are in the first word used in core frames. In Russian 76% of all core frames start with a different word; in German, 52%; and in English only just over a third (36%). English, thus, in addition to having significantly

Table 4
Overall number of core frames and percentage of core frames starting with different words

	Overall Number of Frames	Percentage of Core Frames Starting With Different Words
English	122	36
German	79	52
Russian	63	76

Table 5
Category of the first word of core frames in percentages

	Pronouns	Wh-words	Verbs	Adverbs	Demonstratives	Prepositions	Determiners	Negators	Modal Particles	Other
English	27.9	23.0	26.2	1.6	12.3	2.4	1.6	1.6	1.0	2.4
German	8.9	16.4	29.1	8.9	21.5	2.5	10.1	1.3	–	1.3
Russian	14.3	17.5	11.1	11.1	19.0	7.9	–	4.8	6.3	7.9

more core frames, is also more repetitive with the beginnings of these core frames than is Russian or German. For instance, for English, eight core frames begin with *I (I can, I don't know, I don't think, I'll, I'm, I think, I've, I*) and two begin with *how* (*how many*, *how*).

Table 5 shows the category that the first word of each frame belongs to (as percentages of the core frames for that language).

For English, this means that 27.9% of core frames begin with pronouns and 23% with *wh*-words. The full list of core frames can be found in the Appendix. Table 5 shows that, for all three languages, quite a high proportion of core frames start with wh-words, verbs, or demonstratives. However, the table also gives a relatively clear picture of how the typological differences between the languages interact with the linearly based, lexical specificity at the beginnings of utterances. The higher frequency of core frames starting with pronouns in English than in either Russian or German is because the relative rigidity of English SVO word order results in having a pronoun in first position as subject more often than in the other two languages. Russian has also proportionally more modal particles, prepositions, and negators at the start of core frames than do the other two languages. If we look inside the verb category (Appendix), we can see that, while a large proportion of German and English verbs that start frames are auxiliaries rather than main verbs, this is not the case for Russian, which does not have auxiliaries. The differences in the proportion of verbs clearly results from this absence of auxiliaries and the absence of a copula in Russian. There is an interesting sense, therefore, in which these Russian frames contain more lexical substance than do the English and German ones. It is also interesting that there is a greater variety of negators in the initial words of Russian core frames, whereas in German and English, negation often occurs further through the utterance or, in the case of English, is contracted onto the auxiliary.

4. General discussion

The focus of the current study was on the language that children are actually hearing at the point in time at which they are in the early stages of multiword speech. In particular we were interested in extending previous findings on English child-directed speech to languages with more morphology and potentially much freer word order. Our main question was whether these languages nonetheless show the lexically based word-order patterning found in English and, if so, the extent to which this is modulated by typological differences. To this end we compared English with German, which has a number of word-order variants,

and Russian, which potentially has free word order. We analyzed corpora of child-directed speech in these languages in terms of sentence-initial, lexically based patterns. We found the highest degree of sentence-initial repetitiveness in English, followed by German and then Russian. A mean of between 143 (English) to 78 (Russian) frames per mother accounts for over 75% of the utterances in all three languages. For core frames (those which were frames for at least half the mothers) these figures are 88–45 frames accounting for over 58% of utterances. This order follows the degree of word-order restrictiveness shown by these languages and differences in inflectional morphology between them. Nonetheless, even in Russian, the sentence initial lexical ''frames'' we found accounted for over 75% of the mother's utterances. This indicates that even in some languages where word order is potentially free, the input children receive still contains a great degree of predictability at the beginnings of utterances.

However, there were clearly differences between the three languages in the number of frames, the amount of input data accounted for by frames, and in frame length. English tended to have more and longer frames and these accounted for more of the input data. That is, English mothers provide the most sentence-initial, lexically restricted input and Russian the least. One reason for this is clearly the rigid word order of English compounded with the paucity of verbal and nominal morphology. Our criteria required, first, a strict adherence to sentence-initial linear order and, second, word forms were counted as separate words if they were morphologically diverse (e.g., ''play'' vs. ''plays''). In addition, a perusal of the Appendix reveals that the longer English frames tend to contain either a copula or an auxiliary. English requires an auxiliary in all questions and negative sentences. Since there are only a limited set of auxiliaries, this leads to a greater likelihood of longer frames. German does not require an auxiliary in such sentences. However, since modal auxiliaries, such as *kann* ''can'' and *hat* ''have,'' are fairly common and because German requires the finite verb (or modal) to come in verb-second position in main clauses, this leads to a reasonably large number of two-word frames. The situation is very different for Russian, which has no overt copula in the present tense and does not invert the subject and verb in questions. Consequently English has longer frames than German and German has longer frames than Russian.

Despite these typological differences between the three languages, there was a considerable degree of similarity both in terms of the amount of input data accounted for the initial one to three words and in terms of the predictability of the first word. Even in Russian, the core frame analysis shows that the first word of a sentence tends to be one of only 45 words for more than half the sentences which individual children are hearing. Moreover, since these are core frames, which are by definition those used by 50% of mothers within a language, there is marked consistency between children in the particular sentence-initial words they hear.

These cross-linguistic similarities are partly due to the following facts. First, mothers in all three languages tended to use a fairly large proportion of imperatives (between 11% and 21% of utterances) and these imperatives tended to start with one of a very restricted set of verbs, consistent with previous findings by Broen (1972) for English. Second, we can see from Table 5 that all three languages have relatively high numbers of frames beginning with demonstratives. Finally, from Table 5 and the Appendix we can see that across the three

languages, wh-words occur frequently at the beginnings of utterances and since there is a limited set of these words, this will produce relatively high amounts of utterance-initial repetitiveness.

Many of these similarities across the three languages are interesting in that they are not what would be predicted by a purely syntactic description. For instance, note that while, for English, the position of wh-words, imperative verbs, and demonstratives at the beginning of the utterances is required by English syntax, this is not the case for Russian, for which much more word-order variability is grammatically possible. This of course raises the question of why these similarities occur. Part of the explanation is likely to simply be that descriptive grammars of a language do not define word-order patterns which speakers actually tend to use. Another part of the explanation may concern the pragmatics of talking to small children. The high proportion of utterance-initial demonstratives probably reflects the typically ostensive basis of much adult speech to children of this age within joint attention episodes: pointing things out, naming objects, and talking about the immediate environment. Likewise, the very restricted number of utterance-initial imperative verbs probably relates to the kind of things mothers typically tell their children to do. Thus, one question for future research is thus whether the same degree of utterance-initial restrictiveness is also found in adult-to-adult speech. We suspect that the answer may depend in part on the particular genre and context of speech chosen. Adult-to-adult conversations while collaborating on cooking, repairing a car, or putting up a tent, for example, may show much more lexical restrictiveness than a telephone conversation about last week's trip to Paris.

The nature of adult-to-adult speech is, however, not as crucial for theories of language acquisition as is the nature of child-directed speech. It is clear that in Russian, German, and English, CDS is lexically restricted and patterned in a way that could in principle help account in part for empirical findings of initial lexically based performance in child grammatical acquisition (particularly of English e.g., Braine, 1976; Dąbrowska & Lieven, 2005; Peters, 1983). Indeed, Cameron-Faulkner et al. (2003) did find a significant correlation between the frequency of various copula frames in the English input and the emergence of the same frames in their children's speech. Further evidence for the role of utterance-initial strings comes from the Rowland (2007) study mentioned in the Introduction.

Frequency is obviously not the only factor involved in learning: some highly frequent words are learned relatively late. Phonological salience and semantic relevance to the child must also play important roles. Most crucial for this study is the role of diversity in relation to frequency. Clearly, the child needs to hear something frequently enough to be able to learn it, but if there is no diversity then words and constructions will remain fully lexically specific rather than providing a basis for abstraction. In support of this, Naigles and Hoff-Ginsberg (1998) showed that the absolute frequency of verbs in the input and the number of different constructions that verbs appeared in were both related to the order of acquisition of children's first 25 verbs. A lexically specific string, once it has been learned, may well become partially schematic because a number of similar utterances show diversity in the same position, thus allowing the development of a more abstract slot (Goldberg, 2006).

Future research is needed to pin down the relationship between children's abstraction of the more schematic aspects of the language they are learning and the relative frequency of particular words, cues, and constructions. Such research will need to include languages for which repetition in morphological units might be more important than the repetition of whole words, as for instance in polysynthetic languages. Further, the role of frequency of languages with ''free'' word order needs to be assessed in more detail. For instance, if in Russian there are differences in the frequencies with which mothers use different word orders for the same construction, for example, wh-questions, is this correlated with the speed with their children show evidence of more abstract wh-syntax? Another line of research should explore whether differences between mothers in the frequency and diversity of particular strings relate to their children's development not only of these strings but also of the abstractions that we are suggesting could arise from them.

To sum up: We have found high levels of lexical restrictiveness at the beginnings of CDS utterances in all three languages. But we have also found significant differences between the languages. The question arises as to how to assess the relative importance of these similarities and differences for children's language learning. The goal of this paper was simply to characterize the input from the three typologically different languages; a definitive answer to this question awaits empirical testing, both experimental and modeling, but we make the following suggestions.

4.1. Learning word categories

In terms of learning grammatical categories of words, we suggest that the degree of repetitiveness at the beginnings of utterances would indeed be facilitative. If we just look at the first words of core frames, from Table 4 we know that 17.5% of the Russian core frames start with a wh-word and, from the Appendix, that these consist of 11 types. The figures for English and German are roughly similar. Since children already know something about questioning and answering at this age, it seems likely that these facts would help them start to form a category of wh-words. Thus, already knowing, for instance, that utterances with a particular intonation require answers, and learning that a subgroup of these start with a small group of often phonologically related words, might help in identifying wh-words as a semantically related group and this in turn could contribute to the process of internally analyzing the frequent wh-aux-subject strings that Rowland (2007) has shown protect the child from inversion errors. Once this is achieved a more abstract representation of the syntax of wh-questions will be possible. Similar considerations apply to Russian demonstratives and imperative verbs with which 19% and 11% of core frames, respectively, begin.

A second way in which categories could become identified is through slots in frequent frames. For example, these could provide a basis for the development of both the syntax and the semantics of the noun phrase. Lieven (2006) showed that while the productivity of children's novel multiword utterances at 2;0 could largely be characterized in terms of filling low-scope frames with words for objects or people, these slots then developed more complex NPs. Initially the children placed bare nouns into these semantically identified referent slots. They then moved to using either *a/the* before the noun in the slots and only after this

did other determiners and longer sequences involving adjectives appear. Distributional learning could also lead to the development of a more semantically abstract NP category (one that goes beyond words for objects and people, for instance) since it will cluster words that appear in the slots of frequent frames in the input (e.g., Redington et al., 1998). A number of the German and Russian core frames starting with demonstrative pronouns and, in the case of German, determiners, may well afford the same development, given their frequency and lexical repetitiveness.

The central tasks facing distributional learning approaches to the learning of grammatical categories are (1) to specify the nature of the distributional information being registered, (2) how the varying types of information building up might interact with each other, and (3) how this changes with development. An excellent survey of this problem is provided by Morgan and Demuth's (1996) introduction to their edited volume *Signal to Syntax*. The volume is largely concerned with how the developing characteristics of infant speech perception might facilitate word segmentation and, later, the identification of putative categories of words in a language, for instance, function words versus content words or nouns versus verbs. Morgan and Demuth point out that no single measure is going to provide a fully valid cue to, for instance, word class membership but that, depending on the particular characteristics of the language, sets of measures may interact to provide valid predictors. Christiansen, Allen, and Seidenberg's (1998) ''multiple cue integration'' model is an example. In a model aimed at finding word boundaries, they discovered that information provided individually to the model about phonemes, utterance boundaries, and relative lexical stress did not lead to good identification but that together they were much more successful.

Children learning languages with richer inflectional morphology are building up distributional paradigms around this morphology, and this is also not an instantaneous process. Dąbrowska (2001), for instance, shows that Polish-speaking children, while often producing correct morphology from early on, take a considerable period of time to fill in the whole paradigm and that even some adults are not at ceiling on low frequency, nonphonologically transparent parts of the paradigm. Thus, at the same time as children are learning linear relationships between words, they will also be learning the distributional relationships of the morphological paradigms of their language through processes that may have much in common. The point is that distributional relationships will be building up not just around frequent frames but around all the typologically relevant aspects of a language that have surface perceptual cues available to the language learner and that these distributional patterns will interact with each other. To assess the precise ways in which different degrees of repetitiveness in initial strings interact with other features of the language being learned, for instance, inflectional morphology and variants in word order, is going to require both experimental studies with children and modeling studies in which these variables can be systematically manipulated.

4.2. Learning constructions

In terms of learning constructions, it is quite likely that this high level of repetitiveness provides children with some basic constructional types from which they can start to vary the

word order for semantic/pragmatic purposes. However, again the question is whether the language-specific differences in the degree of repetition have any impact on language learning strategies. As we have seen, Russian has not only fewer but also shorter frames than the other two languages. Does this mean that Russian children's input makes it more difficult to learn the language? The answer to this lies in the grammatical differences mentioned in the Introduction and earlier. If a grammatical feature is not present in a language, there is no occasion to repeat it as is the case in Russian for articles, the copula, and auxiliaries. A Russian child does not need to learn that a presentational construction has three words before the slot: *There is a X*. S/he needs to learn only *Eto X*. Thus, the input provides exactly the type of repetition that is useful for that particular language. While the copula construction exists in a variety of forms across all three languages, we would expect Russian children to learn the highly frequent, though shorter, copula construction from the levels of frequency we have found here. When and how they learn its word-order variability awaits a study of word-order variation in Russian-speaking children from the earliest multiword combinations. There are two possibilities: first that the construction is learned first with one word order and then subsequently ''frees up''; second, that since variability is a typological feature of Russian and if this is reflected in the language that they hear, then Russian children would develop sensitivity to this from the outset (Slobin, 2001). An interesting topic for future research would be to model different levels of word order variability and construction frequency to see when variability emerges in learning.

A similar case where language specific influences become apparent is that of core frames beginning with determiners. As expected there are none in Russian, in English there are exactly two beginning with *the* and *a*, but in German we find 17 core frames. German children have to learn that case is marked on the determiner and thus they have to learn a variety of determiner forms; therefore, it is useful if they are repeated in the input. The most striking discrepancies in the number and length of frames are found in frames beginning with a pronoun, wh-word, and verb. For frames beginning with pronouns neither Russian (seven core frames) nor German (nine core frames) has a large number, whereas English has 28 different core frames, which are often also two- to three-words long and include an auxiliary or a copula. For wh-constructions, English again has 28 core frames, whereas Russian has only 11 and German, 13. The high number of core frames in English wh-construction is most likely due to the strict word order in this construction. After the wh-word the verb has to follow immediately, in German this is almost always the case as well with auxiliaries and copula exhibiting a larger variety than English. Russian has no copula in the present tense and no auxiliaries; thus, the number of frames is reduced and the length of frames is again predominantly one-word as expected. Thus, one could even argue that the level and type of repetitiveness found in the three languages is specifically tailored to the needs of the children learning that language. In all these cases the child is learning a mapping between a particular function (e.g., pointing something out or naming it; asking a wh-question) and a particular constructional form consisting of the initial string followed by a variable. We do not have the direct evidence to say whether the typological differences between the languages would make this an easier or more difficult task for learning a particular construction in a particular language, but our guess is that the systematicity that we have revealed in what

children hear is enough to get them started on assembling an inventory of basic construction types.

That said, we are certainly not claiming that the only way that children initially break into syntax is by using linear, lexically based patterns. This may be a strategy of choice for languages like English and German, in which particular word-order variants frequently occur. But other forms of learning are also obviously taking place and an additional or alternative strategy for children learning languages like Russian might be to learn slot-and-frame patterns based around particular, frequently occurring morphemes. In some sense, morpheme-based frames and lexical, string-based frames might essentially provide the same level of structure in that they would both involve ''markers'' for particular grammatical categories. Both may also provide a similar level of assistance for learning argument structure preferences.

Finally, we should note that the current study used a very conservative set of criteria for deriving frames, which may have actually underestimated the degree of lexical patterning in the input. Many sentence-initial patterns, in German in particular, are not captured by our strict adherence to linear order. That is, many sentence-initial, lexically restricted patterns in German differ slightly in linear order, but these differences are related. Thus, German has frequent sentence-initial patterns such as *jetzt ist der…* (lit: ''now is he…'') and *der ist jetzt…* (lit: ''he is now …''), which are basically the same speech act for the same event, contain the same three lexical items, and have *ist* in invariant verb-second position. Likewise, English child-directed speech has previously been found to contain many highly frequent distributed lexical frames of the type X_Y (e.g., *you_it*), which consist of two constant lexical items separated by only one word. Mintz (2003) found that by selecting the most frequent 45 frames of this type from English CDS in CHILDES corpora, he was able to grammatically categorize the intervening words in approximately half of the corpora at an accuracy rate of between 91% and 98%. An interesting question for future research would be apply various different criteria for lexically restricted frames to child-directed speech in languages like German and Russian, which do not have the rigid word order of English.

4.3. Beyond lexically based schemas

Obviously children move beyond lexically based constructions. The current study and that of Cameron-Faulkner et al. (2003) do not really touch on the issue of how they might progress from initial lexically based, syntactic representations to converge on adult-like syntax. Below we sketch out some possibilities. One is suggested by Abbot-Smith and Tomasello (2006). They argue that as the child cumulatively learns more slot-and-frame schemas such as *What is he ACTION? What does she ACTION Where is Mummy ACTION?* and *Where do they ACTION?* the semantic and distributional commonalities between these lexical items and what they refer to will gradually be reinforced. This should mean that the child eventually represents such sentence-types in terms of *WH-word AUXILIARY PERSON ACTION*. Their view is close to that of Langacker (2000:7), who claims that the more abstract schema which emerges from the process of exemplar learning is ''immanent'' in the sum of the similarities of the individual learned exemplars and is not stored separately from them. The nature of the exemplars, the frequency with which new items need to be

categorized, and the semantic or analogical distance between the new instances and the previously learned exemplars will determine exactly how abstract this ''summed similarity'' (''schema,'' for shorthand) is.

Nonetheless, some generativists have argued that the child's language input cannot in principle be sufficient to allow a gradual abstraction process to converge on an adult-like syntax, particularly for certain rare constructions (e.g., Crain, 1991). Such theorists might, for example, look at the fact that English wh-questions in the input are predominately single-clause questions and wonder how children ever learn to produce and understand complex wh-questions such as *Who did Sandy persuade to see Max?* or indeed passives with by-phrases (such as *the fence was painted by the man*).

Part of the solution is that children DO actually hear these rarer constructions, as evidenced by studies which have examined densely sampled data (e.g., Abbot-Smith & Behrens, 2006). Moreover, there is some evidence that children may, in addition, be assisted in learning these rarer and more complex constructions from their prior acquisition of related and more frequent constructions (e.g., Abbot-Smith & Behrens, 2006). Morris, Cottrell, and Elman (2000) illustrated how this might work in a simulation with a connectionist simple recurrent network, which was trained to assign semantic categories, such as agent or experiencer to words in a series of different constructions involving action (e.g., ''kiss'') and experience verbs (e.g., ''see''). The generalization test involved two systematic gaps in the data presented to the network, both involving experience verbs. The network generalized to the first construction (the questioning of embedded subject sentences like *Who did Sandy persuade to see Max*?) because it was part of an interlocking group of training constructions which ''conspired'' to compensate for the gap. That is, to perform correctly on the embedded subject question test, the network had to assign the semantic role of experiencer to the question word ''who,'' although the network had previously been trained to assign the semantic role of agent to the question word in this construction. In the training, however, the semantic roles agent and experiencer shared a distribution over a number of active voice constructions which are related to embedded subject questions; namely, simple transitives, questions of simple transitives, and embedded transitives, intransitives, and questions of these.

5. Conclusions

We have demonstrated a considerable degree of lexical repetitiveness at the beginnings of utterances addressed to children in languages with less rigid word order and more inflectional morphology than English. Clearly much more research is required to work out exactly what can be extracted by the learner as a function of both the particular level of repetitiveness and the differences in what children hear as a function of typological differences in the language that they are learning. It should be clear, though, that even using our conservative frame criteria, it is possible to account for between 58% and 75% of the input data using a very restricted set of ''frames'' even in a so-called ''free word order'' language like Russian. Thus, it would appear likely that the input in very many languages is more predictable than might be thought from a purely grammatical description.

Notes

1. Utterances are given in italics and frames in italics and bold.
2. Contracted negation in English was counted as one word (e.g., *don't, doesn't, isn't*). Since, with the exceptions noted in the text, we counted words rather than morphemes, we refer throughout the paper to one-, two- and three-word frames.
3. Cameron-Faulkner, Liever, and Tomasello (2003) conducted similar analyses but within each construction, summing the results to assess the number of frames and core frames in their data and the proportion of utterances that they accounted for. The difficulty with this is that there may have been frames that were counted more than once, or, alternatively, not detected at all.
4. The mean number of core frames per mother shows that each English mother used an average of 88 core frames. Overall there were 122 core frames in English. The core frames accounted for 65% of all the English mothers' utterances (a total of 5,638 out of 8,632 utterances).
5. Cameron-Faulkner, Liever, and Tomasello (2003) report a different overall number of frames and core frames for the 12 mothers in their sample. Their overall number of frames is 156. Of course, this may be due to the greater number of mothers but it is also because, as mentioned above, in their study, frames were counted within constructions and added up. Thus, there might have been several frames that were counted more than once, or, alternatively, not detected at all. The same reasoning goes for core frames.
6. Again, this does not equate to the total number of frames for a particular language, as mothers overlapped in the particular frames they used.

Acknowledgments

We are extremely grateful to Gisela Szagun for placing her data on the CHILDES database and to the families who took part in her study and, of course, to the Russian and English families of the Stoll and Manchester corpora. We also would like to thank Thea Cameron-Faulkner for help with coding. Heartfelt thanks go to the research assistants who worked so hard on this project: particularly Wiebke Binder and Kristin Wolter, but Frank Binder, Suse Grassman, Susann Feik, Tatjana Welikanowa, Anya Zimmermann, and Anita Strauss-Naumann also put in a great deal of hard work. We are grateful to Daniel Stahl and Roger Mundry for help with the statistics, to Franklin Chang who helped us with aspects of the data analysis, and to Michael Tomasello, Colin Bannard, and Julian Pine as well as to the Action Editor and three anonymous reviewers for their comments. Some parts of these data have been presented previously at the 2004 meeting of the Deutsche Gesellschaft für Sprachwissenschaft, the 2005 meeting of the International Association for the Study of Child Language, the 2005 Boston University Child Language Conference, and the 2006 Second Biennial meeting of the Russian Cognitive Science Society.

References

Abbot-Smith, K., & Behrens, H. (2006). How known constructions influence the acquisition of other constructions: The German passive and future constructions. *Cognitive Science*, *30* (6), 995–1026.

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Ambridge, B., Rowland, C., Theakston, A., & Tomasello, M. (2006). Comparing different accounts of inversion errors in children's non-subject wh-questions: ''what experimental data can tell us?'' *Journal of Child Language*, *30* (3), 519–557.

Bannard, C., & Matthews, D. E. (2008). Stored word sequences in language learning: The effect of familiarity of children's repetition of four-word sequences. *Psychological Science*, *19* (3), 241–248.

Braine, M. D. S. (1976). Children's first word combinations. With commentary by Melissa Bowerman. *Monographs of the Society for Research in Child Development*, *41*, 1–104.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44.

Broen, P. A. (1972). The verbal environment of the language-learning child. *Monographs of the American Speech and Hearing Association*, *17*, 1–103.

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32* (1), 76–95.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27* (6), 843–873.

Carey, S. (1978). The child as word-learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, *8* (6), 500–508.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26* (5), 609–651.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically different languages. *Cognitive Systems Research*, *9* (3), 198–213.

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, *37* (6), 739–748.

Chomsky, N. (1959). A review of verbal behavior, by B. F. Skinner. *Language*, *35*, 26–58.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2* (10), 406–416.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*, 597–611.

Dąbrowska, E. (2001). Learning a morphological system without a default: The Polish genitive. *Journal of Child Language*, *28* (3), 545–574.

Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16* (3), 437–474.

DeVilliers, J. G. (1985). Learning how to use verbs—lexical coding and the influence of the input. *Journal of Child Language*, *12* (3), 587–595.

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). Comprehension of case marking and word order cues by German children. *Child Development*, *79*, 1152–1167.

Fernald, A., & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science*, *9* (3), F33–F40.

Goldberg, A. (2006). *Constructions at work: The nature of generalisation in language*. Oxford, England: Oxford University Press.

Gómez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70* (2), 109–135.

Hauser, M. D., Weiss, D., & Marcus, G. (2002). Rule learning by cotton-top tamarins. *Cognition*, *86* (1), B15–B22.

Hoff, E. (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74* (5), 1368–1378.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, D. (2006). Functional elements in infant speech processing: The role of determiners in the syntactic categorisation of lexical elements. *Infancy*, *5*, 341–353.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45* (3), 337–374.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Langacker, R. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmerer (Eds.), *Usage-based models of language* (pp. 1–63). Stanford: SLI Publications.

Lewis, J. D., & Elman, J. L. (2001). *A connectionist investigation of linguistic arguments from poverty of the stimulus: Learning the unlearnable*. Paper presented at the Proceedings of the twenty-third annual conference of the cognitive science society, Mahwah, NJ.

Lieven, E. (2006). Producing multiword utterances. In B. Kelly & E. Clark (Eds.), *Constructions in acquisition* (pp. 83–110). Stanford, CA: CSLI Publications.

Lieven, E. V. M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30* (2), 333–370.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24* (1), 187–219.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30* (5), 678–686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90* (1), 91–117.

Monaghan, P., Christiansen, M., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*, 259–305.

Morgan, J. L., & Demuth, K. (1996). Signal to syntax: an overview. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 1–24). Mahwah, NJ: Erlbaum.

Morris, W., Cottrell, G., & Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In S. Wermter & R. Sun (Eds.), *Hybrid neural systems* (pp. 175–193). Berlin: Springer Verlag.

Naigles, L., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25* (1), 95–120.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 109–149). Cambridge, England: Cambridge University Press.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264–275.

Peters, A. M. (1983). *The units of language acquisition*. New York: Cambridge University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: Bradford Books.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.

Rojina, N. (2004). The acquisition of wh-questions in Russian. *Nordlyd*, *32* (1), 68–87.

Rowland, C. (2007). Explaining errors in children's questions. *Cognition*, *104*, 106–134.

Rowland, C. F., & Pine, J. M. (2000). Subject-auxiliary inversion errors and wh-question acquisition: What children do know? *Journal of Child Language*, *27*, 157–181.

Saffran, J. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Seidl, A., & Johnson, E. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9* (6), 565–573.

Slobin, D. I. (2001). Form-function relations: How do children find out what they are? In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 406–449). Cambridge, England: Cambridge University Press.

Stoll, S. (n.d.). Unpublished Russian corpus.

Szagun, G. (2004). Learning by ear: On the acquisition of case and gender marking by German-speaking children with normal hearing and with cochlea implants. *Journal of Child Language*, *31* (1), 1–30.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28* (1), 127–152.

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds ''listen through'' ambiguous adjectives in fluent speech. *Cognition*, *100*, 389–433.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Wells, G. (1981). *Learning through interaction: The study of language development*. New York: Cambridge University Press.

Wilson, S. (2003). Lexically specific constructions in the acquisition of inflection in English. *Journal of Child Language*, *30* (1), 75–115.

Appendix 1: Core frames across the corpus categorized by the first word
Russian

Pronouns	Wh-words	Verbs	Adverbs	Demonstratives	Prepositions	Negators	Modal particles	Other
ja 'I'	*chego* 'what.GEN.'	*davaj* 'give.IMP, come on'	*potom* 'then'	*tam* 'there'	*na* 'on'	ne 'not'	*nu* 'well'	*a* 'and'
ja tebe 'I you.-DAT'	*chto* 'what'	*podozhdi* 'wait.IMP'	*sejchas'* 'now'	*tut* 'here'	*u* 'at'	*ne nado* 'not should'	*nu chto* 'well what.NOM'	*no* 'but'
my 'we'	*chto takoe* 'what (is) this'	*pojdem* 'let's go'	*sejchas' ja'now I'*	*von* 'over there'	*v* 'in'	*nikak* 'by no means'	*nu davaj* 'well, come on,/give.IMP'	*mama* 'mom'
on 'he'	*chto ty* 'what you'	*sadis'* 'sit.IMP'	*teper'* 'now'	*zdes'* 'here'	*za'*for,behind'		*nu-ka* 'well'	*papa* 'daddy'
ona 'she'	*gde* 'where'	*skazhi* 'say.IMP'	*tol'ko* 'only'	*vot* 'here'	*s* 'with'			*Vanja*
oni 'they'	*kak* 'how'	*smotri* 'look.IMP'	*tozhe* 'also'	*vot tak* 'here so'				
ty 'you'	*kak ty* 'how you'	*vidish'* '(you) see'	*eshche* 'still'	*vot eto* 'here- this is'				
vsjo 'everything	*kakoj* 'which'			*eto* 'this is'				
vse 'all'	*kto* 'who'			*eto chto* 'this what'				
	kto eto 'who this'			*eto kto* 'this who'				
	kuda 'where to'			*eto ne* 'this not'				
				tak 'so'				

German

Pronouns	Wh-words	Verbs	Adverbs	Demonstratives	Prepositions	Determiners	Negators	Other
du hast 'you have'	*was denn* 'what particle'	*geht* 'goes'	*dann* 'then'	*da* 'there'	*in* 'in'	*das 'this / that / it/'*	*nicht* 'not'	*wenn* 'when'
du kannst 'you can'	*was hast du 'what have you'*	*guck mal 'look.IMP particle'*	*jetzt ist* 'now is'	*hier* 'here'	*mit* 'with'	*den 'the.ACC'*		
du 'you'	*was ist da 'what is there'*	hast du *'have you'*	*jetzt 'now'*	*da ist der 'there is the.MASC'*		*der hat 'the. MASC has'*		
ich 'I'	*was ist das 'what is this'*	*hast 'have'*	*ganz* 'very'	*da ist die 'there is the.FEM'*		*der ist 'the. MASC is'*		
ich glaube 'I think/ believe'	*was ist denn 'what is particle'*	*ist da 'is there'*	*genau* 'exact / exactly'	*da ist 'there is'*		*der 'the.MASC'*		
ich habe 'I have'	*was ist 'what is'*	*ist das 'is it / this / that'*	*noch* 'still'	*da kann 'there can'*		*die 'the.FEM'*		
wir 'we'	*was soll 'what shall'*	*ist 'is'*	*vielleicht* 'perhaps'	*da sind 'there are'*		*ein 'a / an.MASC/ NEUT'*		
	was 'what'	*kann 'can'*		*hier ist 'here is'*		*eine 'a / an.FEM'*		
	wer 'who'	*kannst du 'can you'*		*hier sind 'here are'*				
	wie 'how'	*kannst* 'can'		*das geht* 'this /that / it works'				
	wo 'ist der where is the-NOM'	*Komm mal her 'come.IMP particle here'*		*das ist der* 'this / that / it is the.MASC'				
	wo ist 'where is'	*komm 'come.IMP'*		*das ist die* 'this / that / it is the.FEM'				
	wo 'where'	*lass 'let.IMP'*		*das ist ein* 'this / that / it is a'				
		mach mal 'do.IMP particle'		*das ist ja* 'this / that / it is particle'				
		mach 'do.IMP'		*das ist* 'this / that / it is'				
		möchtest du 'want you'		*das sind 'these/those are'*				
		musst du 'have to you'		*so* 'so'				
		musst *'have to'*						
		soll ich 'shall I'						
		warte mal 'wait particle'						
		willst du 'want you'						
		willst'want'						
		wollen wir 'want we'						

English

Pronouns	Wh words	Verbs	Adverbs	Demonstratives	Prepositions	Determiners	Negators	Others
he-'has	how many	are you go-ing	now	here	in the	a	don't	because
he-'is	how	are you	just	here-'is	in	the	not	so
He	what about	can you remember		there-'is the	on			daddy
I can	what are you	can you see		there-'is				mummy
I do-'nt know	what are	can you		there we are				
I do-'nt think	what did you	can		there				
I-'ll	what do you	come on		that-'is a				
I-'m	what does	come		that-'is it				
I Think	what else	did you		that-'is not				
I -'ve	what happen-ed	did		that-'is right				
I	what-'has happen-ed	do you know		that-'is the				
she-'is	what-'has	do you want		that-'is				
She	what is it	do you		that one				
they're	what-'is that	do		that				
They	what-'is the	does		this				
we-'ll	what-'is this	has						
We	what-'is	have you got						
you can	what-'re you	have you						
you can-'nt	what	have						
you-'ll	where did	is he						
you-'re go-ing	where-'has	is it						
you-'re not	where-'is	is that						
you-'re	where-'is the	is						
you-'ve	where	let-'us see						
you-'ve got	which one	let-'us						
you want	who-'is	let						
You	who	look						
Your	why	put the						
it does-'nt		put						
it-'is a		shall I						
it-'is not		shall we						
it-'is		was						
it was								
It								