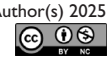
Special Issue: Multi-Word Units in Multilingual Learning: Chunks, Phrasemes and Formulaic Language in Learning and Teaching Contexts

The International Journal of Bilingualism

Recycling constructional patterns: The role of chunks in early bilingual acquisition

International Journal of Bilingualism
1–21



DOI: 10.1177/13670069251346103
journals.sagepub.com/home/ijb


Nikolas Koch
Ludwig-Maximilians-Universität München, Germany

Antje Endesfelder Quick
Leipzig University, Germany

Stefan Hartmann
Heinrich Heine University Düsseldorf, Germany

Abstract

Aims and Objectives: Over the last decades, usage-based research into child language acquisition has shown that multi-word units play a key role in child language acquisition. This paper sets out to explore the role of multi-word units in the bilingual speech of two German-English bilingual children, focusing on their code-mixing and starting from the hypothesis that code-mixed utterances can be accounted for with the help of constructional patterns that are in turn abstracted away from recurrent multi-word units.

Approach: We used the Chunk-Based Learner (CBL), a computational model developed by Stewart McCauley and Morten Christiansen.

Data and Analysis: We applied the CBL to longitudinal data from two children growing up bilingually with the same language pair (German-English) but notably different input situations, to detect recurrent multi-word units in the data.

Conclusions: The study shows that both monolingual and code-mixed utterances are built from the same constructional patterns, highlighting the crucial role of chunks in language acquisition. Although code-mixed utterances appear highly creative, they largely rely on fixed, formulaic patterns similar to those found in monolingual speech. This finding supports the usage-based approach, suggesting that chunk-based learning is fundamental across different language acquisition contexts.

Originality: Using a strictly bottom-up approach allows us to identify recurrent chunks that are used as "building blocks" of early child language, including, importantly, code-mixed utterances.

Significance: Our study adds to previous research emphasizing the role of multi-word units in early bilingual acquisition and contributes to ongoing efforts to pinpoint the way in which constructional patterns emerge from multi-word units in more detail.

Corresponding author:
Nikolas Koch, Institut für Deutsche Philologie, Ludwig-Maximilians-Universität München, Schellingstraße 3 / RG, 80799 München, Germany.
Email: Nikolas.Koch@lmu.de

Keywords
Chunks, bilingual language acquisition, usage-based approach, code-mixing, language modeling

Introduction

In language acquisition, children must identify the basic components of their language(s), such as phonemes, morphemes, and words, as well as how these elements combine to form larger structures, for example, inflected words and sentences. The development of children's language use is often described as progressing from simple syllables to complex multi-word phrases. While this perspective highlights the combinatorial nature of language learning, it arguably overlooks the important role that larger patterns play in helping children understand language(s) and its rules. Peters (1983) was among the first to propose that children utilize larger phrases during language acquisition, emphasizing the importance of *gestalt* processes. She pointed out that there is a distinction between the linguistic units analyzed by researchers and those that children use when they begin to speak[1] (Arnon & Christiansen, 2014). By focusing solely on "words," we may overlook the multi-word sequences that children extract and employ early in their speech. Since children do not hear adult speech neatly divided into phonemes, morphemes, and words, they must parse the continuous flow of speech to identify the relevant linguistic units. This process involves breaking down larger segments of speech into smaller, manageable units. Wong Fillmore (1976, 1979) explored similar ideas in the context of children's second language acquisition and observed that children learning a new language often use whole phrases or sentence fragments without fully analyzing them at first. These so-called chunks serve as a kind of "entry point" into the language, enabling children to express themselves early on, even if they do not yet understand the underlying grammar. For example, they might repeat and use phrases like "Can I have?" or "I don't know" without knowing how these expressions are constructed. Chunks play a vital role because they allow children to engage in social interactions right away, which is crucial for learning a language. As children gain experience, they gradually start to break down these chunks, analyze their components, and use them in creative ways. This process of chunk decomposition helps children to develop a deeper understanding of the grammar and structure of the target language. Thus, chunks act as both a practical tool for immediate communication and as a foundation for more advanced language skills—in other words, they act as building blocks for more complex and productive structures.

The idea that larger chunks are important in language acquisition has been further explored in usage-based theories of language learning, according to which children learn grammar by abstracting and generalizing from stored multi-word sequences (Goldberg, 2019; Tomasello, 2003). Multi-word chunks, which span across individual words, provide specific lexical phrases that children can use in their early speech. These chunks help them uncover grammatical relationships and patterns of co-occurrence between words. Two key processes contribute to this: undersegmentation, where a multi-word phrase is initially learned as a single unit and later segmented into smaller parts, and chunking, where frequent usage leads to the merging of words into multi-word units (Arnon & Christiansen, 2014). Both processes suggest that children rely on these multi-word chunks during their language learning journey. This perspective is increasingly supported by psycholinguistic studies that highlight the importance of fixed multi-word combinations in language comprehension and production for both, children (e.g., Arnon & Clark, 2011; Bannard & Matthews, 2008) and adults (e.g., Arnon & Snider, 2010; Jolsvai et al., 2013). Computational modeling further strengthens this view, illustrating that knowledge of multi-word sequences can explain children's real-time comprehension and production processes (e.g., McCauley & Christiansen, 2011, 2014, 2017) and contributes to the development of abstract grammatical knowledge (e.g., Solan et al., 2005). While substantial empirical evidence supports the role of chunks in children's

language acquisition, both in monolingual and second language contexts, questions remain about their role in bilingual language acquisition. Specifically, to what extent do multi-word chunks contribute to the formation of mixed utterances, where a child uses both languages in one utterance, which are often regarded as highly creative linguistic phenomena? This question also connects to earlier work by Backus (2003), who demonstrated that such mixed utterances often consist of multi-morphemic units situated at the interface between the lexicon and grammar. Analyzing these elements as lexical chunks suggests that code-switching is not merely a creative act but also builds on routinized patterns grounded in language use.

In this paper, we investigate the role of chunks in both monolingual and mixed utterances of bilingual children from a usage-based perspective, using a computational model. We present an exploratory study employing the Chunk-Based Learner (CBL) model (McCauley & Christiansen, 2017, 2019a) to analyze bilingual language acquisition data. Using longitudinal data from two German-English bilingual children, we analyze their code-mixing utterances, monolingual speech, and the input they receive from their caregivers. In doing so, we follow up on previous research reported in Koch et al. (2022b), where we investigated the bilingual speech of one child using the CBL model. In the present paper, we extend this approach to a second child, which allows for a more detailed comparison in terms of individual differences, and we offer a more detailed discussion of the implications of our results for assessing the role of multi-word units in early bilingual language acquisition.

The CBL approach helps us determine whether the children's utterances reflect specific patterns found in their input. From a usage-based perspective, we assume that children's mixed utterances are constructed from patterns that are also present in their input and their monolingual utterances and that multi-word chunks play a central role in this. We begin with a brief overview of the usage-based approach, highlighting its implications for bilingual language acquisition, particularly in relation to mixed utterances. Next, we describe the empirical study, including the data and the CBL model employed. The findings are then presented and discussed, focusing on the role of chunks in mixed and monolingual utterances and pointing out some individual differences between the two children studied, while also taking a critical look at the methodology. Finally, we consider the broader implications of these results for understanding the role of chunks in bilingual language acquisition.

Bilingual acquisition from a usage-based perspective

The question of how children acquire language(s) is one of the most intensely debated topics in linguistics and cognitive science (see Rowland, 2014). Consequently, a variety of theories and methods have been developed to address this issue. One influential approach in this field is the usage-based approach (e.g., Tomasello, 2003; Tomasello & Lieven, 2008), which builds upon earlier lexically oriented theories of grammatical development (e.g., Braine, 1976) and aligns with linguistic approaches that blur the line between grammar and lexicon (e.g., Langacker, 1987). Tomasello (2009, p. 69) summarizes the core principle of the usage-based approach with two aphorisms: "meaning is use" and "structure emerges from use." In the context of language acquisition, this means that children learn all aspects of their language(s)—including grammatical structure, meaning, and the pragmatic dimensions of language use (e.g., Bruner, 1983, p. 18)—from the input they receive.

The usage-based approach posits that children's language experience serves as the foundation for their ability to abstract linguistic patterns, using their cross-domain cognitive skills such as pattern finding and intention reading (Tomasello, 2009).[2] In particular, the ability to recognize patterns is considered essential for language acquisition (see Tomasello, 2003, p. 34). Children are capable of filtering and categorizing various regularities from their input at an early age. Consequently, the "rules" of a language are continuously acquired through a process of

generalization from actual language use. The communicative contexts prevalent at the beginning of language acquisition support this process, as child-directed speech is often highly repetitive and contains many formulaic expressions (Cameron-Faulkner et al., 2003; Lester et al., 2022; Stoll et al., 2009; Szagun, 2019, pp. 206–238). In addition, already recognized patterns help children further perceive and categorize their linguistic environment (Romberg & Saffran, 2010). Thus, children receive rich linguistic input that enables them to discover regularities and gradually build their grammars. However, it is important to note that input can vary significantly from child to child: children have different language experiences, contexts, and conversation partners.

According to the usage-based approach, children learn language gradually, starting with individual words and fixed word combinations, referred to as multi-word chunks (Tomasello, 2006, p. 261). From the very beginning of language acquisition, children use such larger linguistic units. They combine or segment them to discover the units of their language(s) and the regularities of their combinations. As shown by Lieven et al. (1992), up to 50% of the first 400 multi-word utterances produced by 2-year-olds are "frozen," that is, their components are not used flexibly but appear in limited, fixed combinations (see also Lieven et al., 1997; Pine & Lieven, 1993). Two- and three-year-olds are quicker and more accurate in producing frequently occurring chunks and are influenced by chunk frequency when making syntactic generalizations (Arnon & Christiansen, 2014). As such, children make use of fixed multi-word chunks, allowing them to simultaneously uncover co-occurrence patterns between words and grammatical relationships.

In recent decades, first steps have been taken to apply the ideas and methods from usage-based language acquisition research to multilingual contexts (e.g., Quick & Backus, 2022; Quick et al., 2021; Vihman, 1999; see Backus, 2020 for an overview). While the usage-based approach does not make an a-priori distinction between monolingual and multilingual acquisition, certain linguistic phenomena occur exclusively in multilingual language acquisition and use, potentially challenging usage-based theory. One such phenomenon is code-mixed utterances like (1).

(1) *look at the ampel, it's kaputt.* 'look at the traffic light, it's broken' (Fion, 03;10.22)[3]

A variety of approaches and models have explored the reasons behind language mixing[4] especially in adults. In multilingual societies, it is common for individuals to switch between their languages in various ways (Gardner-Chloros, 2009). Early research (e.g., Poplack, 1980) primarily focused on the sociolinguistic and pragmatic reasons why people—mainly adults—combine multiple languages in their speech, as well as the question of what structures and rules underlie mixed utterances (for an overview of recent research approaches, see Bullock & Toribio, 2009; Gardner-Chloros, 2009; MacSwan, 2020). The primary goal of these studies has been to uncover explanations for these constraints within the fundamental syntactic structures of language and the typological features of individual languages (Poplack, 1980; see also Belazi et al., 1994; Di Sciullo et al., 1986; MacSwan, 1999; Myers-Scotton, 1993). These two research perspectives—one predominantly sociolinguistic and the other more grammar-oriented—have largely been pursued independently. From a usage-based perspective, however, these aspects are inseparably linked. This approach focuses on the constructional patterns available to speakers, how these patterns are acquired, and how they are used in specific (social) contexts during language interaction and acquisition.

Language is always tied to social contexts. The acquisition of words and more complex linguistic constructions is not only shaped by various situations and usage contexts but also by different languages, each with distinct collocational patterns and frequency profiles. As Schmid (2020, p. 343) puts it, "cotextual, situational, interpersonal, and social aspects of usage are entrenched as parts of routinized patterns of associations." Thus, linguistic knowledge emerges from language use in specific contexts. The close connection between language(s) and social, cultural, and

interactional factors, which is central to usage-based approaches (see also Croft, 2009, for an early programmatic work in cognitive linguistics), is particularly evident in bilingual language acquisition. However, bilingual language acquisition often occurs under very different conditions and circumstances. Some children grow up in bilingual families, others in societies where multilingualism is the norm, while some are raised in monolingual families living in an L2 society. From a usage-based perspective, it is important to consider how these differences affect the acquisition process, whether linguistic knowledge changes in response to various input situations, and to what extent children's output reflects their input contexts. In bilingual acquisition, examining the interaction between linguistic input and children's output should provide insights not only into the development of each individual language but also into how linguistic patterns appear in code-mixing utterances.

Based on these previously discussed theoretical implications regarding the role of chunks in language acquisition, as well as the assumptions of a usage-based language acquisition theory, the present study aims to investigate the significance of chunks for bilingual language acquisition, focusing on two research questions:

1. Which similarities and differences can be observed in monolingual and code-mixed utterances of bilingual children with regard to the use and complexity of chunks?
2. How does the role of chunks change over the course of language development, and what impact does the linguistic input received by the children have on this process?

Methods

Participants and data

The study is based on longitudinal corpora from two bilingual children growing up with German and English. Both children are experiencing simultaneous bilingual first language acquisition, with their parents following the OPOL (one-parent-one-language) strategy, meaning that each parent communicates with the child in a different language—one parent speaks German while the other speaks English. Both children lived in middle-class households in Germany and were recorded in their home environment during everyday interactions, such as playtime or dinner. The first child, "Fion," is the second child of a German-speaking mother and an English-speaking father. Although the parents generally adhered to the OPOL strategy when speaking to Fion, they did not commit to a family language and sometimes alternated between both languages when all family members were present. In addition, the data include utterances from Fion's older brother, who also grew up bilingually and occasionally used code-mixing when speaking with Fion or their parents. From the age of 18 months, Fion attended a German-speaking daycare 5 days a week. Data collection took place from ages 02;03 to 03;11, averaging about 2 hours per week, resulting in a total corpus of approximately 205 hours.[5] Fion's input situation was initially predominantly German, which was also evident in his own language use. As he was more proficient in German, his parents decided to increase the English input to support its development shortly after his third birthday. This resulted in a noticeable shift in the distribution of the two languages in his output data: while he predominantly spoke German in the earlier recordings, a significant portion of his later utterances in the corpus is in English. The data indicate a gradual return to a higher proportion of German in the following months (see the Supplementary Material available at https://doi.org/10.17605/OSF.IO/ZN4KF for a longitudinal overview of the language use of both children). The second child, a girl named "Silvie," has a German-speaking father and an English-speaking mother, who was Silvie's only source of English. Her father's English skills were quite limited, so the parents communicated in German. Silvie's mother was at home with her

Table 1. Corpus overview.

Child	# Child's utterances (total)	# Child's code-mixing utterances	# Parents' utterances
Fion	47,812	3,492	180,292
Silvie	37,995	4,279	140,387

during the first year and most of the second, during which time Silvie was primarily exposed to English throughout the day. However, from the age of 18 months, Silvie's input situation shifted as she began attending a German kindergarten for 45 hours per week, where she was predominantly exposed to German from that point onward. The corpus includes recordings from ages 02;04 to 03;10, with Silvie being recorded on average 2.5 hours per week, resulting in a total of 135 hours. All recordings were transcribed in SONIC CHAT format (MacWhinney, 2000). The data have not yet been tagged for part-of-speech, but they have been manually coded for language at the utterance level, indicating whether each utterance is in English, German, or code-mixed. The code-mixed utterances were further tagged manually on a word-by-word basis to specify the language of each individual word. In their monolingual utterances, both children predominantly spoke German. Table 1 gives an overview of the corpus.

Chunk-Based Learner

Usage-based theory posits that language users' grammatical knowledge is grounded in concrete linguistic events. Linguistic structures are acquired and modified through their use in specific contexts. Methodologically, this leads to an empirical approach to studying grammatical structures. This approach has yielded significant insights, particularly in research on monolingual language acquisition (for overviews, see Behrens, 2021 and Tomasello, 2009). One of the methods employed in this context is the CBL (McCauley & Christiansen, 2014, 2017, 2019a, 2019b). In the present study, we apply this model to multilingual language acquisition. Our focus is on investigating the role of chunks in both monolingual and code-mixing utterances of the children and their relationship to the children's input during the bilingual acquisition process. By analyzing two longitudinal corpora, the study also enables the examination of inter-individual variation in the use of code-mixing. This approach provides valuable insights into the interaction of construction patterns across different languages. As previously mentioned, the CBL model has been used to analyze similarities and differences between first and second language acquisition. Since a usage-based approach posits that bilingual acquisition is not fundamentally different from monolingual language learning, as in both cases the same cognitive mechanisms are employed to acquire a repertoire of linguistic constructions, the CBL model should significantly contribute to understanding bilingual acquisition, particularly in relation to children's code-mixing utterances.

The CBL model is based on the incremental recognition of chunks, or related multi-word units, using a relatively simple metric: backward transitional probabilities (BTPs). The BTP value indicates how likely it is that the previous word precedes the current word. This value is calculated solely based on the linguistic units processed up to that point. The model tracks both frequency information for words and word pairs, as well as the BTPs between these pairs (McCauley et al., 2015). In other words, the model "learns" linguistic units by being fed input word by word, recognizing related structures through a bottom-up process. The choice of this metric for chunk identification is grounded in a series of studies from experimental psychology and developmental biology,

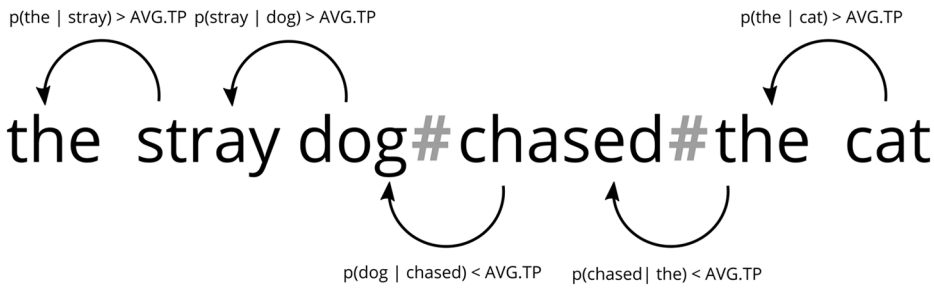


Figure 1. Simplified illustration of the Chunk-Based Learner algorithm, adapted from McCauley and Christiansen (2019a, p. 10). The "#" signs indicate where the algorithm identifies boundaries between chunks. At these points, the BTP between the lexemes p(lexeme1|lexeme2) falls below the current average transitional probability (avg.tp).

which have shown that even infants as young as 8 months old respond sensitively to BTPs in word segmentation tasks (Pelucchi et al., 2009; Saffran et al., 2008).

The basic idea of the model can be illustrated with the following example:[6] when the model encounters the phrase *the stray dog chased the cat*, it captures the frequencies of the individual words as well as the bigrams such as *the stray*, *stray dog*, *dog chased*, *chased the*, and *the cat*. It also calculates the BTPs, which represent the likelihood that, for instance, *dog* directly follows *stray*, *chased* directly follows *the*, and so on. The model's algorithm then computes a running average of the BTPs for all previously identified word pairs. If the BTP of the current word pair exceeds the current average BTP value, this bigram is classified as a potential chunk or part of a chunk. Conversely, if the BTP value falls below the average, the model's algorithm sets a boundary between the word pair. Therefore, in the example provided, if the BTP values for *the stray* and *stray dog* are above the average, the model recognizes *the stray dog* as a cohesive chunk. This would be added to the so-called "Chunkatory," which contains the "learned" chunks of the model (McCauley & Christiansen, 2019a). If we assume that the BTP value for "dog chased" is below the average BTP, the CBL model will set a chunk boundary in this case (illustrated in Figure 1, where the chunk boundary is marked with <#>). There is no a-priori limit on the size of chunks. The key factor is the calculation of the BTP values: if they exceed the average, a cohesive structure is defined. Thus, anything situated between two boundaries—whether those are the boundaries of utterances or those set by the model when the BTP of a word pair falls below the running average BTP—is considered a chunk.

The procedure described so far is part of the "understanding side" of the model. This aspect was used in the present study to identify chunks in both the monolingual and code-mixing utterances of the children, as well as in their respective input. The second component of the model, the "production side," is not included in this study and will therefore not be discussed further (see McCauley & Christiansen, 2019a for details). We used McCauley's CBL script, available at https://github.com/StewartMcCauley/CBL/ (accessed on 11 November 2024), to obtain the chunking results.

Results

This section presents the results of the data analysis, combining the quantitative CBL approach with qualitative assessments. The focus of our analysis is on the distribution of chunks in the code-mixing utterances and monolingual utterances and on the longitudinal development of the role of chunks. This also reveals interesting individual differences between the children. It is important to

stress that the CBL model does not recognize languages; rather, it calculates chunks and word boundaries based on the statistical metrics described earlier. Consequently, when processing the input received by the children as well as the children's output, the model does not differentiate between German, English, and code-mixing utterances to define cohesive chunks or word boundaries.

Distribution of chunks across monolingual and code-mixing utterances. First, we will quantitatively assess the differences between the chunks identified in the children's code-mixing utterances and their monolingual utterances. Second, we will look at the results qualitatively by taking a closer look at the most frequent chunks.

In the first step, we compared the average number of chunks per utterance and the average number of words per chunk across the different utterance types. This allows us to check whether the transitional probabilities in the code-mixed utterances differ considerably from those in the monolingual utterances. Figure 2 displays the average number of chunks per utterance and the average number of words per chunk for both children, showing that the two types of utterances exhibit clear differences. As Figure 2 shows, the results for both children (Fion in the upper, Silvie in the lower panel) are very similar. The algorithm identified a larger number of chunks in code-mixed utterances than in monolingual utterances. However, chunks in monolingual utterances were on average longer, containing more words, than chunks in code-mixed utterances. To some extent, this is expect, as code-mixed utterances are more likely to contain very rare word combinations, leading to very low transitional probabilities, which entails that the algorithm will set a chunk boundary. Still the results provide a clear indication that there are considerable differences between monolingual and code-mixed utterances that can be detected by an algorithm that is blind to the languages the individual words belong to.

Looking more closely at the code-mixed utterances, we see that the position of chunk boundaries determined by the CBL algorithm often coincides with the position of code switches (see Figure 3). For Fion's code-mixed utterances, this occurs about one third of the time ($n = 1{,}120$), that is, all code switches in the respective utterance coincide with chunks. This can be seen in the following utterances, for example, where <#> again stands for the chunk boundary assumed by the model:

(2) a. zeig # ice-cream. "show ice-cream" (Fion, 02;03.16)
b. ein kleinen # shark. "a little shark" (Fion, 02;03.16)
c. nein # a ice hockey player. "no a ice hockey player" (Fion, 02;03.16)

In 1960 cases, there is at least a partial overlap between the position of chunk boundaries and of language switches, that is, some but not all chunk boundaries in the utterance coincide with language boundaries:

(3) a. jetzt gleich # this zumachen. "right now this close" (Fion, 02;10.08)
b. I have # that werkzeug. "I have that tool" (Fion, 03;10.22)

In example (3a), the utterance begins with the German chunk "jetzt gleich," followed by a switch to English with "this." But Fion then returns to German with the verb "zumachen." Only in 426 code-mixed utterances, there is no overlap. In those utterances, we often find recurrent code-mixed sequences such as *time out machen* "do time out" or *(ein) anderer frog* "another frog." In Silvie's data, the proportion of full overlaps is smaller (979 out of 4,286), but in the vast majority of cases (2,810), we find a partial overlap between chunk and language boundaries. The fact that the proportion of complete overlaps is lower for Silvie could be due to the fact that Silvie's

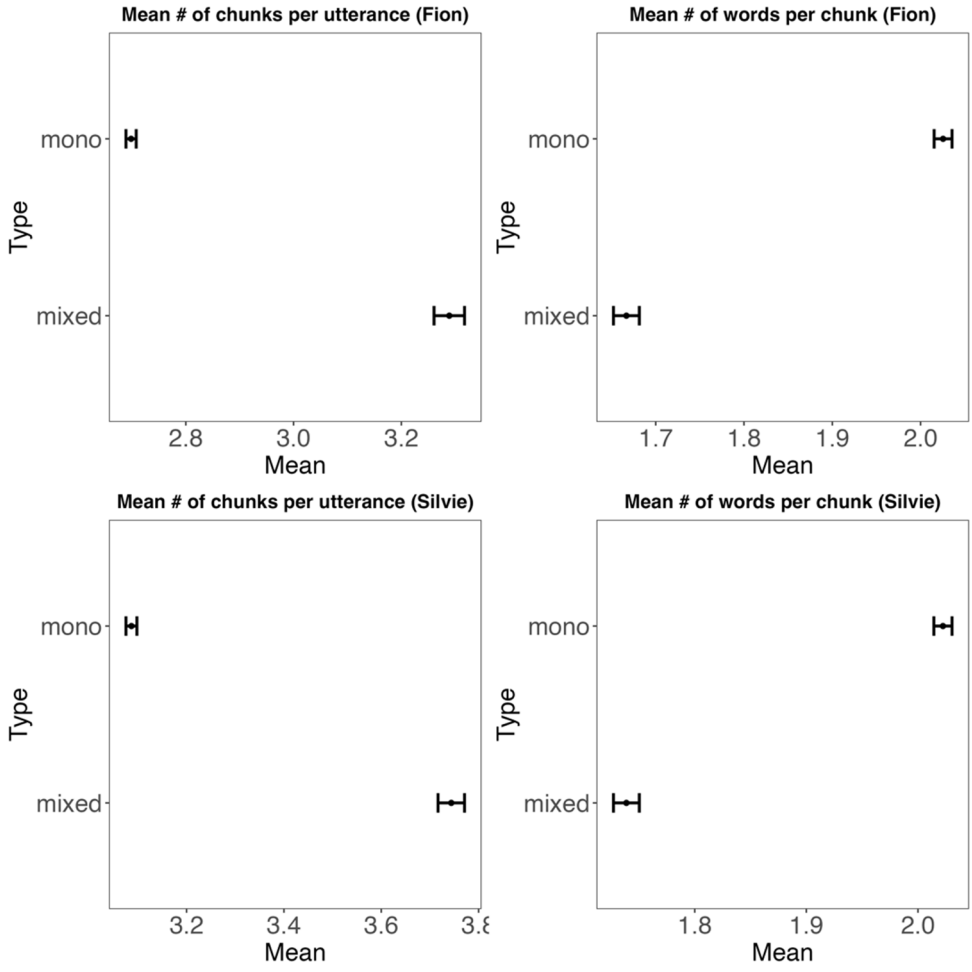


Figure 2. Mean number of chunks per utterance and mean number of words per chunk in the code-mixed vs. monolingual data for Fion (upper part) and Silvie (lower part). The error bars show the standard error.

utterances are more complex overall, as shown by the mean length of utterance (MLU) of the two children (see Figure 4). This will be discussed in the following section.

We will now proceed to the qualitative analysis of the data, focusing on the ten most frequent multi-word chunks consisting of two or more words, or of at least three words (following the CHAT conventions that were used in transcribing the data, contractions like I'm are counted as two words), identified by the CBL in the datasets of Fion and Silvie (see Table 2). The left part of Table 2 contains the top 10 chunks for the code-mixing utterances, the right one for the monolingual ones. It is noteworthy that several of the identified chunks appear prominently in both the top ten lists for code-mixed and monolingual utterances. In Fion's case, five of the ten most frequent chunks comprising two or more words are present in both types of utterances (highlighted in light gray in the table). For chunks consisting of three or more words, three out of ten also appear in both categories (shaded dark gray). A similar pattern is observed for Silvie: seven out of ten chunks with two or more words are used in both utterance types, while two out of ten with three or more words

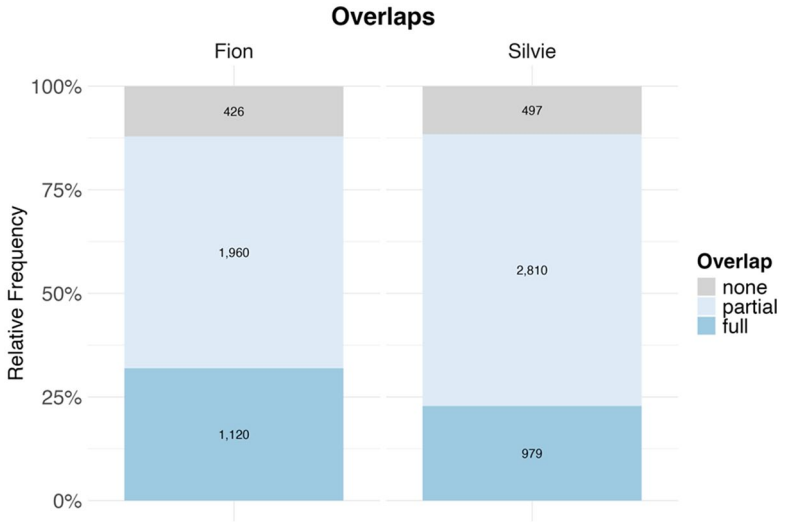


Figure 3. Proportion of overlaps between chunk boundaries and language switches in Fion's and Silvie's code-mixed utterances.

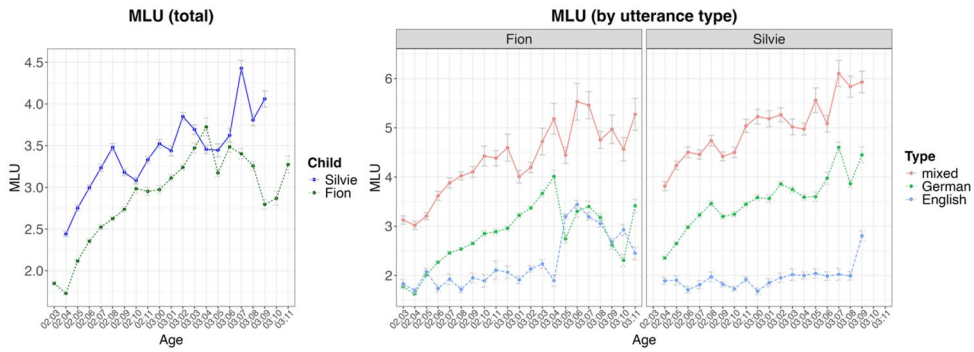


Figure 4. Mean Length of Utterance (MLU) for Fion and Silvie. The MLU was calculated based on words, meaning the graph displays the average number of words per utterance at each age. The left graph shows the MLU across all utterance types, while the right graph breaks down the MLU by type (monolingual German and English, as well as code-mixed utterances).

overlap as well. This overlap highlights the interconnectedness of language use across different contexts and suggests that certain chunks may play a significant role in the children's overall linguistic development (see Section 4 for further discussion).

Longitudinal development and individual differences. The previous results regarding the chunk boundaries are based on the complete dataset, comprised of data collected over the course of several years; however, it may be insightful to examine the longitudinal developments more closely. In this subsection, we will therefore analyze how the number of chunks per utterance evolves over time. To be able to classify the results in the overall linguistic development of the children, we first looked at the development of the MLU. MLU has often been used as a measure of syntactic complexity since Brown (1973). Throughout most of the study period, Silvie's MLU is higher than Fion's, with an interesting decline occurring around the time when his predominant language shifts

Table 2. Top 10 chunks with 2 or more words/with 3 or more words identified by CBL in Fion's and Silvie's data.

Fion

CM 2 + words		CM 3 + words		Mono 2 + words		Mono 3 + words	
Chunk	Freq.	Chunk	Freq.	Chunk	Freq.	Chunk	Freq.
ich will	63	ich will nicht	14	guck mal	723	i can t	71
ein sword	36	choo choo train	5	ich will	436	ich kann nicht	66
this istis	36	i want to	5	ich bin	428	i want to	49
i m	33	that s a	5	i m	265	ich weiss nicht	48
that s	33	eine sexy lady	4	das ist	244	do nt know	47
ich bin	28	i can t	4	ich habe	196	i do nt know	40
das istis	27	noch mehr orange juice	4	look at	194	was ist das	39
und dann	23	und hierhere isist	4	noch mehr	188	ich will nicht	36
guck mal	22	a builder handy	3	du bist	140	it s not	33
das ist	21	cup o tea	3	was denn	138	in the garden	28

Silvie

CM 2 + words		CM 3 + words		Mono 2 + words		Mono 3 + words	
Chunk	Freq.	Chunk	Freq.	Chunk	Freq.	Chunk	Freq.
ich moechte	103	a little bit	10	ich moechte	483	noch ein bisschen	37
ich bin	88	in the oven	10	ich haben	342	eins zwei drei	32
das istis	87	ein eisice cube	8	das ist	333	eins zwei drei vier	17
ich haben	87	das ist mein	7	ich will	319	das geht nicht	15
das ist	68	in the box	6	das hier	314	noch ein stueck	13
ich habe	48	noch ein bisschen	6	guck mal	272	one two three	13
du bist	39	das ist kein	4	ich bin	263	kannst du mir	12
ich will	39	kannst du mir	4	ich kann	253	nein nein nein	12
und dann	36	on the beach	4	ich habe	183	one two three four	12
ich mache	33	the combine harvester	4	und dann	182	a b c	10

from German to English (see Figure 4). As shown in Figure 4, this pattern is especially evident in the code-mixed utterances, which have the highest MLU for both children. This is not surprising, given that the monolingual utterances contain a significant number of single-word utterances (Fion: 17,042 out of a total of 47,761 utterances; Silvie: 10,716 out of a total of 37,995 utterances). In contrast, code-mixing utterances typically consist of more than one word, with single-word utterances occurring only in rare cases, such as mixed compounds like "firewehr" (Fion, 02;04.22).

If we now look at the development of the chunks in the data for the two types of utterance, we see the following: As the children's utterances, whether code-mixed or not, become longer on average over time, and as the model's "chunkatory" keeps growing as it continues to process input data, its likelihood to detect already encountered chunks even in the code-mixed data becomes higher.

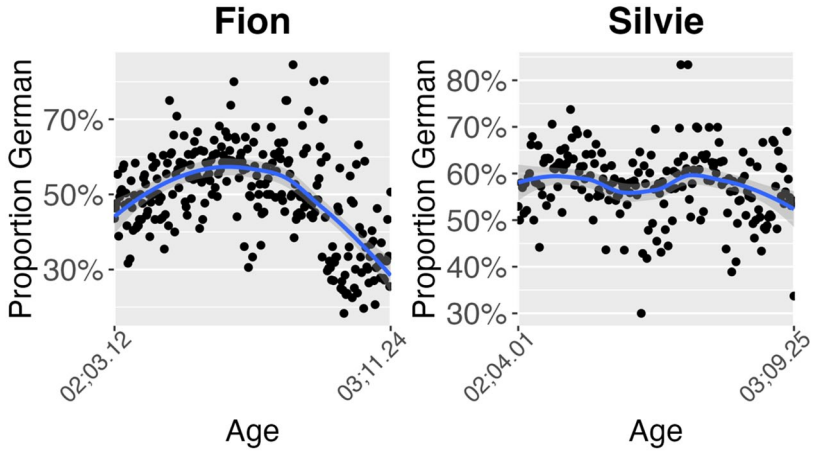


Figure 5. Proportion of German elements in Fion's and Silvie's code-mixed data (assessed using the word-by-word language tags as described in the Methods section).

This is exactly what we found (see Figure 5). The normalized number of chunks per utterance becomes more similar across different utterance types over time. Note that the ups and downs in the data can partly be accounted for by data sparsity. As both children tend to use relatively few monolingual English utterances (except for Fion toward the end of the investigation period), these few utterances have a tremendous effect on the proportions displayed in Figure 5, causing considerable ups and downs. Still, the data show a few general tendencies: For both children, the normalized number of chunks in monolingual German utterances remains relatively stable. Unsurprisingly, the code-mixed data contain more chunks than the monolingual ones, as we have already seen in Section 3.3.1. The number of chunks in Fion's code-mixed utterances is subject to more variation than in Silvie's utterances. In Fion's case, the mean number of chunks per utterance goes back quite drastically and consistently until the age of 02;11, followed by an increase in the average number of chunks per utterance. This can partly be attributed to Fion's language shift to English explained above, which also has ramifications on his code-mixed utterances: As Figure 5 shows, the proportion of German elements in his code-mixed data decreases. This suggests that in his later code-mixed utterances, Fion tends to insert individual German words in his otherwise English utterances, and a cursory qualitative look at the data lends support to this idea, as utterances like I'm going *in my Urlaub* (03;10.22) or *they are want to essen* (03;11.23) abound in his later code-mixed data. As the model encounters more novel transitions not attested in previous stages of the data, it sets more chunk boundaries, leading to a higher number of chunks. In Silvie's case, the number of chunks in her code-mixed utterances slightly decreases over time, indicating that she is re-using some patterns that the algorithm has encountered before and categorized as chunks more frequently.

Discussion

The role of chunks in children's monolingual and code-mixed utterances

A comparison of the mean number of chunks per utterance and the mean number of words per chunk across the different utterance types revealed the same pattern for both children: the chunks identified in code-mixing utterances contain, on average, fewer words than those in monolingual utterances. In other words, the number of chunks per utterance is significantly higher in code-mixing utterances than in the monolingual data. However, the number of words per chunk is likely

to be higher in the monolingual data than in the code-mixed data. One reason for this is the method used: The code-mixing utterances are likely to contain many word combinations that rarely or never appear in the OPOL input the children received, and hence, can be expected to show a low transitional probability. Importantly, this observation also highlights that code-mixing is not a phenomenon triggered solely by the input, that is, generated by caregivers, but an inherent attribute of bilingualism. Even in an OPOL setting, bilingual children produce code-mixed utterances that go beyond their input, combining elements from both languages in one utterance. This points to the active role children play in constructing their bilingual grammar and suggests that code-mixing reflects cognitive flexibility rather than mere imitation. It thus underscores the need to view code-mixing as a natural and productive part of bilingual development. As a result, despite the fact that many high-frequency chunks attested in monolingual data also occur in the code-mixed data, the model encounters many "surprising" transitions, which are likely to yield BTP values below the average BTP value, leading the model to set a chunk boundary. Therefore, it could be expected that the model sets more chunk boundaries in code-mixed utterances than in monolingual ones. However, we believe it would be too narrow to interpret this solely as a methodological artifact. At the same time, as previously mentioned, this result also reflects the children's OPOL input situation. The results show that chunks are generally present in both monolingual and code-mixed utterances, and the differences in number and complexity can be well explained by the input the children receive. These findings directly address the research question of what differences and similarities can be observed in monolingual and code-mixed utterances of bilingual children regarding the use and complexity of chunks. The higher chunk frequency in code-mixed utterances, alongside the reduced complexity of individual chunks, highlights a distinct structural difference in how chunks are employed across the two types of utterances. An in-depth analysis of the chunks in code-mixed utterances has also revealed that chunk boundaries frequently align with language switches. This also offers a new perspective concerning code-mixing because it indicates that language mixtures are largely composed of fixed formulaic phrases and that individual words or specific syntactic features are not the primary factors driving code-mixing (Quick et al., 2021; Quick & Hartmann, 2021). The chunks children use in their language also feed into their code-mixed utterances. Code-mixing can therefore also be seen as a recycling process in that frequent chunks are also used to construct code-mixed utterances (Dąbrowska, 2014). This is also what we see in our data: the ten most frequent chunks appear to a large extent in both types of utterances (see Table 2).

This also suggests that code-mixed utterances can be identified as typical instances of item-based constructions that play a key role in language acquisition according to usage-based approaches (Tomasello, 1992, 2003). In addition, the list of top 10 monolingual chunks in the data reflects each child's input situation. Whereas Fion initially had more German input and later experienced a shift to more English input, Silvie's input is predominantly German throughout the recordings. Silvie's most frequent chunks mirror this input situation in that she mainly uses German chunks which are also "recycled" in her code-mixing. By contrast, Fion's input shift also shows in his production of chunks, both in the monolingual as well as in the code-mixed utterances. As such, in both children we can see that they extract linguistic knowledge, such as chunks, from the input they receive and use these chunks in both utterance types.

Examining the longitudinal development of the data reveals that the difference between utterance types—monolingual versus code-mixed—in terms of chunks decreases over time. As all children's utterances become longer on average (see Figure 4) and more chunks are detected by the chunkatory, the likelihood that the chunkatory will recognize chunks already identified in code-mixed utterances also increases (see Figure 6). This applies equally to both children. However, comparing the two different input situations of the children reveals highly interesting results, which align with a usage-based approach to language acquisition. In Silvie's case, the number of chunks

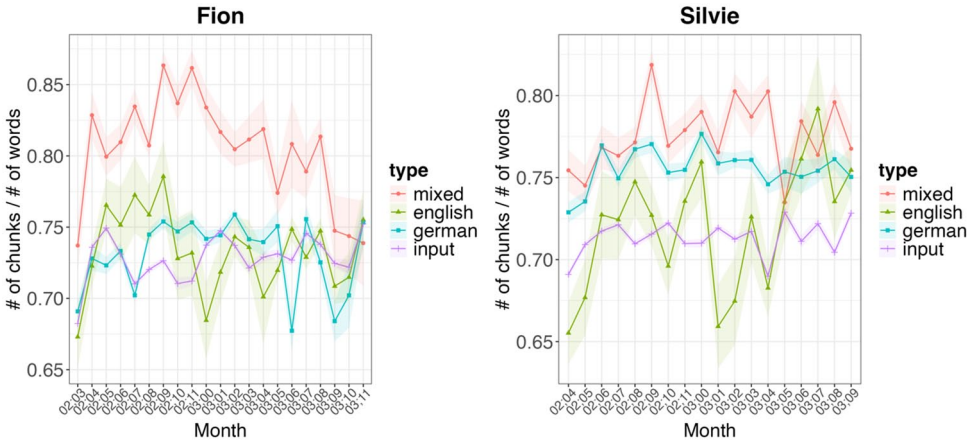


Figure 6. Normalized number of chunks per utterance over time. The ribbons show the standard error of the mean.

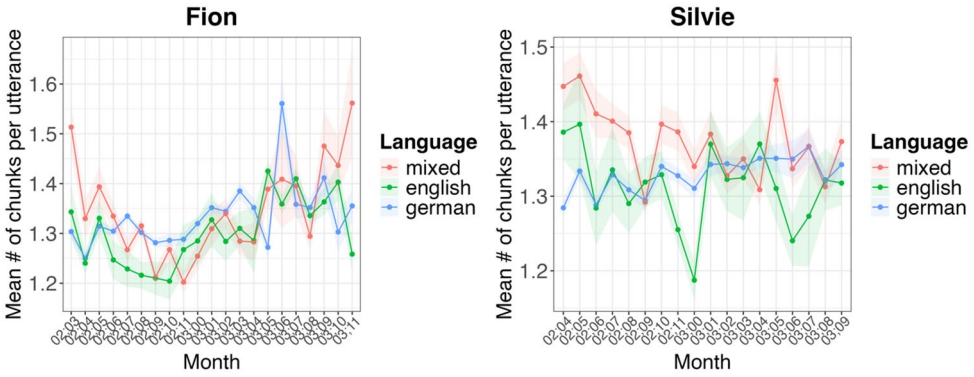


Figure 7. Average number of chunks per utterance.

in her monolingual German utterances and code-mixed utterances remains relatively stable throughout her development. The notable fluctuations in her English utterances can be explained by the very limited number of monolingual English utterances in the data, which strongly affects the proportions observed. Regarding chunk complexity over time, her monolingual German utterances also show a high degree of stability, while the complexity of her code-mixed utterances decreases slightly. In Fion's case, we observe that the number of chunks in his code-mixed utterances decreases (see Figure 6), while the average number of words per chunk increases significantly from around the age of 3;3 (see Figure 7). At the same time, the number of words in his English chunks also shows a slight increase from this point onward. This shift can be partly explained by Fion's changing input situation. Until about the age of three, his main input was in German. However, after his third birthday, there was a shift toward more English, due to an extended stay in his father's home country and more frequent visits from his English-speaking grandparents, who neither spoke nor understood German. It is noteworthy that this increased exposure to English also influenced his code-mixed utterances, as shown in Figure 5. This underscores

the role of input as a key factor in first language acquisition, significantly shaping the linguistic elements children use to construct their utterances. Notably, this seems to apply equally to both monolingual and code-mixed utterances in terms of chunk usage. This observation aligns with the usage-based theory of language acquisition, which posits that children are initially conservative learners, relying heavily on fixed multi-word sequences to form their utterances, and only gradually develop more complex linguistic patterns through abstraction processes. These processes lead to productive schemas, allowing children to use language in increasingly productive and creative ways. Using the CBL model, these fixed patterns were incrementally extracted from the data in a bottom-up process. However, the question arises as to how cognitively plausible such an approach is and, consequently, how meaningful results are. This will be examined in more detail in the following section.

A critical evaluation of the method

As we have seen, the model yields interesting results, even though the results of the CBL algorithm cannot be taken at face value of course: Although the algorithm aims at modeling language acquisition in a cognitive plausible way, some of the results we obtained can be explained by the inner workings of the model without necessarily being very informative about the process of acquisition per se. For instance, the shift of Fion's dominant language from German to English has a considerable impact on the results because the model is of course trained on the (predominantly German) input data that the corpus contains—the model cannot know about the rich English input that Fion received at around that time outside of the recording sessions. This limitation applies not only to the CBL model but must also be considered in general when interpreting corpus studies, which can only capture a portion of the linguistic experience that a child receives during the study period. In other cases, results that can be easily explained by the inner workings of the model can at the same time be argued to reflect learning in a plausible way. For instance, the observation that we find a higher number of chunks in code-mixed utterances can be accounted for by the fact that the transitional probabilities between words from different languages are very low because such word combinations hardly occur in the input. On the other hand, however, this also reflects the children's OPOL input situation, and a potential difference between monolingual and code-mixed utterances: while monolingual utterances often contain larger phrases that are "recycled" as entire units, code-mixed utterances combine different patterns in new and creative ways.

Unlike other quantitative methods, such as the traceback method (Dąbrowska & Lieven, 2005; Koch, 2019; Koch et al., 2022a, 2022b), the CBL does not use fixed frequency thresholds. This means that word combinations that occur dozens of times are not treated the same as combinations that occur only twice. The CBL model has the significant advantage of incrementally extracting chunks by considering the relatively simple metric of BTPs. The CBL model thus offers the advantage of defining chunks in a more cognitively plausible way through the incremental identification of linguistically related units. However, it should be noted that linguistic patterns are not acquired solely based on frequencies. In addition to frequency, salience also plays a crucial role in the formation of chunks and in language acquisition more generally (Koch, 2019; Schmid, 2007), and should therefore be included in further investigations.

A further limitation of the model in its current implementation is its focus solely on chunks. As such, it cannot account for linguistic abstractions which play a key role in usage-based approaches, such as frame-and-slot patterns which are considered a gateway to more complexity and growing productivity (e.g., Lieven et al., 2009). In addition, the CBL focuses solely on the formal and distributional aspects of constructions, therefore neglecting semantic dimensions. This limitation also applies to other methods of pattern identification, such as the traceback method (Dąbrowska &

Lieven, 2005; Koch, 2019; Koch et al., 2022a, 2022b) and the dynamic network model (Ibbotson et al., 2019). Consequently, an open question remains regarding how semantic, as well as phonological and prosodic, aspects can be more systematically integrated into the characterization of patterns in early language acquisition.

Finally, it is an open question to what extent models like CBL or traceback are equally informative for typologically distinct languages that differ in, for example, the rigidity of their word order or the way grammatical relations are encoded. Constituent order conventions could also have ramifications for the usefulness of applying backwards transitional probabilities: It could be hypothesized that in a language in which determiners precede nouns, for example, *the boy*, *boy* is a much better predictor for *the* than vice versa, which could help explain why BTPs have been shown to outperform forward transitional probabilities in a number of studies (Pelucchi et al., 2009; Saffran et al., 2008). However, it seems plausible that the opposite may hold in a language in which the determiner follows the noun, such as Korean. As such, a further open question concerns the cognitive plausibility of a BTP-based approach across different languages. Future studies should therefore take a closer look at different language pairs, and also at different input situations. For instance, it would be quite interesting to investigate the effect of input data from parents who frequently code-mix themselves. In general, a closer and more systematic investigation of the relationship between input and output would be another important desideratum for future studies.

All in all, the results of the CBL, explorative as they may be, provide important insights into the role of chunks in early (bilingual) language acquisition. They lend further support to the idea that both utterance types, monolingual utterances and code-mixed ones, are constructed around chunks and lexically fixed patterns that are based on experience.

Conclusion

The acquisition of multilingualism has long been an important topic in language acquisition research, and much of the research implicitly shares many of the central assumptions of the usage-based approach. However, researchers have only recently begun to examine multilingualism from an explicitly usage-based, cognitive-linguistic perspective (for an overview, see Backus, 2020). The usage-based approach posits that cognitive processes such as pattern recognition are central to the language acquisition process. Accordingly, one of the primary goals of this approach is to perform a kind of *reverse engineering* of the processes at work during language acquisition: using quantitative methods such as the CBL employed here, the "building blocks" upon which children acquire language are identified. It is assumed that these "building blocks" do not fundamentally differ between monolingual and multilingual language acquisition. The comparison of the role of chunks in monolingual and code-mixing utterances has shown that code-mixed utterances are made up to a large extent of the same constructional patterns that can also be found in monolingual utterances. This is in line with the usage-based approach to language acquisition. Especially a closer look at the most frequent chunks attested in both monolingual and code-mixed utterances reveal many commonalities between the two utterance types. In general, the results are very much in line with the idea that speakers, and language learners in particular, continuously "recycle" utterances (Dąbrowska, 2014), or, more generally, constructional patterns. Chunks in particular seem to play an important role here. This study provides further empirical evidence that chunks and their combinations are a crucial component of language acquisition, both in monolingual and code-mixed utterances. At first glance, code-mixed utterances may seem like a challenge for a usage-based approach, as they appear to be examples of highly productive and creative language use rather than formulaic speech. However, upon closer examination, it becomes clear that fixed linguistic patterns also play an important role in code-mixed utterances. Just as early monolingual

language acquisition is largely formulaic and characterized by simple schemas that serve as gateways to greater complexity, code-mixed utterances can also be understood as combinations of well-established chunks from different languages and/or the productive use of formulaic patterns with one or more open slots.

Much of the research on first language acquisition still focuses on the combinatorial aspects of language: the transition from words to a (presumably) abstract syntax. However, chunk-based learning, in which smaller chunks are combined into larger ones or larger chunks are segmented into smaller ones, plays an equally important role in the learning process by creating new linguistic units and discovering the relationships between them. Acknowledging the significance of chunk-based language acquisition raises several open questions and also opens up new opportunities for examining language acquisition processes comprehensively, including both monolingual and bilingual first language acquisition, as well as second and foreign language acquisition in adults.

Taken together, the findings of this study allow us to address the two central research questions outlined above. The first research question was: *Which similarities and differences can be observed in monolingual and code-mixed utterances of bilingual children with regard to the use and complexity of chunks?* This study has shown that monolingual and code-mixed utterances of bilingual children are structurally similar in that both rely heavily on recurring chunks. While code-mixed utterances contain more chunk boundaries and tend to be composed of shorter chunks, many of the most frequent chunks are shared across both utterance types. This suggests that code-mixing is shaped by the same item-based patterns that underlie monolingual speech. The second research question was: *How does the role of chunks change over the course of language development, and what impact does the linguistic input received by the children have on this process?* The analysis has demonstrated that chunk use is highly dynamic over time and closely tied to the children's input. Changes in the linguistic environment, such as shifts in dominant input language, are reflected in both the frequency and complexity of chunks in the children's speech. This underscores the importance of input in the development of chunk-based knowledge and supports usage-based assumptions that language acquisition is driven by frequency, salience, and gradual abstraction from concrete linguistic experiences.

Acknowledgements

We are grateful to the parents and children who took part in this study and to the anonymous reviewers who significantly improved the manuscript. We would like to extend our thanks to our project team for preparing and annotating the data, especially Nina Julich-Warpakowski and our student assistants Maximilian Adolphi, Annika Klotz, Asude Kölün, and Luca Müller.

Author contributions

Conceptualization, investigation, methodology, formal analysis, funding acquisition, validation, writing—review & editing: N.K., A.E.Q., S.H. Data curation: A.E.Q. Visualization: S.H. Writing—original draft: N.K.

Data availability

The script that was used to analyze and visualize the data as well as a small excerpt of the corpus data are available at https://doi.org/10.17605/OSF.IO/ZN4KF. For privacy reasons, the full datasets cannot be made available at this point. However, we are currently working on anonymizing the corpus data. Once anonymized, they will be published on the CHILDES database.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported on here was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), project number 504095269. In addition, our research was conducted in the context of the scientific network "Language contact phenomena in multilingual language acquisition" (LaCoLA), also funded by the DFG, project number 496468900.

Ethical considerations

The data investigated here were collected at the Max Planck Institute for Evolutionary Anthropology, Leipzig, from 2005 to 2007. All ethics requirements applicable at that time were taken into account. All parents gave informed consent for the recordings, and all published data have been carefully pseudonymized.

ORCID iDs

Nikolas Koch https://orcid.org/0000-0001-6917-9318
Antje Endesfelder Quick https://orcid.org/0000-0002-9240-1068
Stefan Hartmann https://orcid.org/0000-0002-1186-7182

Supplemental material

Supplemental material for this article is available online.

Notes

1. The research reported on here focuses exclusively on spoken languages, which is why we use monomodal terms in the present paper, although it can be expected that most of the processes discussed here also play a role in the acquisition of signed languages.
2. Note that while generative approaches still regard language as an autonomous cognitive module and assume a genetic basis for its acquisition, even some generative approaches now emphasize the role of linguistic experience (see Yang, 2016).
3. Chapter 4 provides a detailed description of the corpora cited in this and the following examples. As is common in studies on first language acquisition, we refer to a child's age in the format "Years;Months. Days," that is, *03;10.22* indicates that Fion was 3 years, 10 months, and 22 days old when he made that utterance.
4. The terms language mixing, code-mixing, code-switching, and so on are sometimes used inconsistently. In this paper, we use mixing to refer to the use of units from more than one language in the same utterance.
5. For comparison, Tomasello and Stahl (2004, p. 105) suggest that a child is awake for about 10 hours a day, during which they produce and comprehend language structures. Therefore, recording 2 hours per week accounts for roughly 2.9% of the child's awake time. However, this is a very rough estimate, and there are likely significant individual differences in children's awake times.
6. The example and its description are adopted from Koch et al. (2022).

References

Arnon, I., & Christiansen, M. H. (2014). Chunk-based language acquisition. In P. Brook, & V. Kempe (Eds.), *Encyclopedia of language development* (pp. 88–90). Sage.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth—Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language*, *62*, 67–82.

Backus, A. (2003). Units in code-switching: Evidence for multimorphemic elements in the lexicon. *Linguistics*, *41*(1), 83–132. https://doi.org/10.1515/ling.2003.005

Backus, A. (2020). Usage-based approaches. In E. Adamou & Y. Matras (Eds.), *The Routledge handbook of language contact* (pp. 110–126). Routledge.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Behrens, H. (2021). Constructivist approaches to first language acquisition. *Journal of Child Language*, *48*(5), 959–983.

Belazi, H. M., Rubin, E. J., & Toribio, A. J. (1994). Code switching and X-bar theory: The functional head constraint. *Linguistic Inquiry*, *25*, 221–237.

Braine, M. D. S. (1976). *Children's first word combinations* (Monographs of the Society for Research in Child Development, Serial No. 164). University of Chicago Press.

Brown, R. (1973). *A first language: The early stages*. George Allen & Unwin.

Bruner, J. S. (1983). *Child's talk: Learning to use language*. Oxford University Press.

Bullock, B. E., & Toribio, A. J. (2009). Themes in the study of code-switching. In B. E. Bullock & A. J. Toribio (Eds.), *Cambridge handbook of linguistic code-switching* (pp. 1–17). Cambridge University Press.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child-directed speech. *Cognitive Science*, *27*(6), 843–873. https://doi.org/10.1207/s15516709cog2706_2

Croft, W. (2009). Toward a social cognitive linguistics. In V. Evans, & S. Pourcel (Eds.), *New directions in cognitive linguistics* [Human Cognitive Processing] (Vol. 24; pp. 395–420). John Benjamins.

Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, *25*, 615–653.

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

Di Sciullo, A.-M., Muysken, P., & Singh, R. (1986). Government and code-mixing. *Journal of Linguistics*, *22*, 1–24.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.

Goldberg, A. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Ibbotson, P., Salnikov, V., & Walker, R. (2019). A dynamic network analysis of emergent grammar. *First Language*, *39*(6), 652–680. https://doi.org/10.1177/0142723719869562

Jolsvai, H., McCauley, S., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 692–697. https://escholarship.org/uc/item/5cv7b5xs

Koch, N. (2019). *Schemata im Erstspracherwerb: Eine Traceback-Studie für das Deutsche* (Linguistik, Impulse & Tendenzen 80). De Gruyter. https://doi.org/10.1515/9783110623857

Koch, N., Hartmann, S., & Endesfelder Quick, A. (2022a). The traceback method and the early constructicon: Theoretical and methodological considerations. *Corpus linguistics and linguistic theory*, *18*, 477–504. https://doi.org/10.1515/cllt-2020-0045

Koch, N., Hartmann, S., & Endesfelder Quick, A. (2022b). Traceback and chunk-based learning: Comparing usage-based computational approaches to code-switching. *Languages*, *7*(4), 271. https://doi.org/10.3390/languages7040271

Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford University Press.

Lester, N. A., Moran, S., Küntay, A. C., Allen, S. E. M., Pfeiler, B., & Stoll, S. (2022). Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. *Cognition*, *221*, 104986. https://doi.org/10.1016/j.cognition.2021.104986

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

Lieven, E., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*(2), 287–310.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481–507. https://doi.org/10.1515/COGL.2009.022

MacSwan, J. (1999). *A minimalist approach to intrasentential code switching*. Routledge.
MacSwan, J. (2020). Theoretical approaches to the grammar of code-switching. In E. Adamou & Y. Matras (Eds.), *The Routledge handbook of language contact* (pp. 88–109). Routledge.
MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.
McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1619–1624). Cognitive Science Society.
McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*, 419–436.
McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*(3), 637–652. https://doi.org/10.1111/tops.12258
McCauley, S. M., & Christiansen, M. H. (2019a). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51. https://doi.org/10.1037/rev0000126
McCauley, S. M., & Christiansen, M. H. (2019b). Modeling children's early linguistic productivity through the automatic discovery and use of lexically-based frames. In A. K. Goel, C. M. Seifert & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 782–788). Montreal, QB: Cognitive Science Society.
McCauley, S. M., Monaghan, P., & Christiansen, M. H. (2015). Language emergence in development: A computational perspective. In B. MacWhinney & W. O'Grady (Eds.), *Handbook of language emergence* (pp. 415–436). Blackwell.
Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press.
Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.
Peters, A. M. (1983). *The units of language acquisition*. Cambridge University Press.
Pine, J. M., & Lieven, E. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language*, *20*, 551–571.
Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español. *Linguistics*, *18*, 581–618.
Quick, A., & Backus, A. (2022). A usage-based approach to pattern finding: The traceback method meets code-mixing. *Languages*, *7*(2), 135. https://doi.org/10.3390/languages7020135
Quick, A., Endesfelder, A., Backus, A., & Lieven, E. (2021). Entrenchment effects in code-mixing: Individual differences in German-English bilingual children. *Cognitive Linguistics*, *32*(2), 319–348. https://doi.org/10.1515/cog-2020-0036
Quick, A. E., & Hartmann, S. (2021). The building blocks of child bilingual code-mixing: A cross-corpus traceback approach. *Frontiers in Psychology*, *12*, Article 682838. https://doi.org/10.3389/fpsyg.2021.682838
Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wires Cognitive Science*, *1*, 906–914.
Rowland, C. (2014). *Understanding child language acquisition*. Routledge.
Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, *107*(2), 479–500. https://doi.org/10.1016/j.cognition.2007.10.010
Schmid, H.-J. (2007). Entrenchment, salience and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 117–138). Oxford University Press.
Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford University Press. https://doi.org/10.1093/oso/9780198814771.001.0001
Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*(33), 11629–11634. https://doi.org/10.1073/pnas.0409746102
Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child-directed speech. *Cognitive Science*, *33*(1), 75–103. https://doi.org/10.1111/j.1551-6709.2008.01004.x
Szagun, G. (2019). *Sprachentwicklung beim Kind* (7th ed.). Beltz.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

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

Tomasello, M. (2006). Acquiring linguistic constructions. In W. Damon, R. M. Lerner, D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 255–298). Wiley.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 69–87). Cambridge University Press.

Tomasello, M., & Lieven, E. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. J. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 168–196). Routledge.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, *31*(1), 101–121.

Vihman, M. M. (1999). The transition to grammar in a bilingual child: Positional patterns, model learning, and relational words. *International Journal of Bilingualism*, *3*(2–3), 267–301. https://doi.org/10.1177/13670069990030020801

Wong Fillmore L. (1976). *The second time around: Cognitive and social strategies in second language acquisition* [Unpublished PhD dissertation, Stanford University]. ProQuest LLC.

Wong Fillmore, L. (1979). Individual differences in second language acquisition. In C. J. Fillmore, D. Kempler, & S.-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 203–228). Academic Press.

Yang, C. D. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.

Author biographies

Nikolas Koch is a Senior Researcher in German Linguistics at the Institute for German Philology at LMU Munich. His research focuses on the acquisition of German in multilingual contexts from a usage-based perspective.

Antje Endesfelder Quick is a Senior Researcher of English Linguistics at Leipzig University. In her research, she is primarily interested in multilingual first language acquisition and language contact phenomena from a usage-based perspective.

Stefan Hartmann is a professor of German linguistics at Heinrich Heine University in Düsseldorf, Germany. His research focuses on all aspects of language dynamics: acquisition, variation, change, and evolution.