Arnon, I. & Christiansen, M.H. (2014). Chunk-based language acquisition. In P. Brooks & V. Kempe (Eds.), Encyclopedia of Language Development (pp. 88-90). Thousand Oaks, CA: Sage Publications.

Chunk-Based Language Acquisition

In learning to talk, children have to discover the linguistic units of their language (sounds, morphemes, words) and the ways these units can be combined to create larger patterns (inflected words, sentences). Children's progression is often characterized as a move from smaller building blocks to larger combinations: from syllables to words to multi-word combinations. This characterization captures the combinatorial aspect of language learning but does not address an equally important process: the use of larger chunks to discover the units of language and the regularities governing their combination. The idea that children use larger chunks in learning language was first formulated by Ann Peters in her seminal work on the units of language acquisition. She emphasized the role of gestalt processes in language learning, and highlighted the difference between the units linguists use to analyze language and the ones children employ when learning to talk. By looking for 'words' we miss out on the larger multi-word sequences children extract and use in early production. The use of such units reflects the fact that infants don't hear adult speech neatly segmented into phonemes, morphemes, and words. To discover these units, children first need to break into the speech stream and identify the relevant linguistic units, a process that necessarily involves decomposing larger chunks – stretches of unsegmented speech – into smaller linguistic units. Similar whole-to-part processes play a role in children's discovery of the sounds and morphemes of their language: children can learn about phonological contrasts by comparing whole words (whole-word phonology) and use inflected words to learn about the inflectional system in their language.

The insight that larger chunks play a role in language acquisition was further developed in usage-based approaches to language learning – where children learn about grammar by abstracting and generalizing over stored utterances. Multi-word chunks (crossing lexical word boundaries) provide children with lexically specific chunks to be used in early production and allow them to discover grammatical relations and co-occurrence patterns that hold between words. Such building blocks can be formed through *undersegmentation* – where a multi-word sequence is first acquired as a chunk and only later properly segmented, or through *chunking* – where patterns of usage cause words to fuse together into one multi-word unit. Both processes make the prediction that that children make use of multi-word chunks in the learning process.

Empirical evidence for children's use of multi-word chunks

There is growing evidence that children's early building blocks include multi-word chunks and that they are sensitive to the properties of multi-word combinations. Children produce 'frozen' multi-word utterances at a stage where most of their other productions consist of single words. Many later productions are still not fully productive. As Elena Lieven and her colleagues demonstrate, up to 50% of the first 400 multi-word utterances produced by two-year-olds can be classified as 'frozen': their components are not used productively but instead appear only in restricted combinations. Children's later productions are also impacted by their knowledge of larger chunks. Two- and three-year olds are faster and more accurate to produce higher frequency chunks (*a drink of milk* compared to *a drink of tea*), and are impacted by chunk frequency in making syntactic generalizations. Children's morphological accuracy is also affected by the larger contexts in which words appear: four-year-olds are more accurate at producing irregular plurals (e.g., *teeth*) inside higher frequency chunks (*brush your --- teeth*). Similar patterns are seen in computational simulations. In a model that uses data-oriented parsing to parse a corpus of child speech many of the units identified in the early stages of language production (up to age three) were multi-word ones. In another model – which uses backwards transitional probabilities to identify units of language use in a corpus of child speech - children's language is better captured when the lexicon contains multi-word chunks in addition to single words, reflecting the 'chunked' nature of children's early language.

Children's use of multi-word chunks is also reflected in their error patterns. Four-year-olds have difficulty changing a first-person prompt such as *I think* or *I believe* into a third-person one (e.g., *he thinks*) for verbs (e.g., mental-state verbs) that predominantly appear with a first-person subject. Children's me-for-I errors (pronoun case errors such as *me do it* where the accusative-marked pronoun is used instead of a nominative one) can be related to the proportion of preverbal uses (e.g., *let me do it*) in their input. Children are less likely to make inversion errors in questions for strings that appeared inverted frequently in the input. Over a range of constructions, children's correct and incorrect productions show sensitivity to multi-word units.

Using multi-word chunks as building blocks for language use

The reliance on multi-word information is not limited to child learners. Adult speakers are also sensitive to the distributional properties of multi-word chunks and draw on such information in

production and comprehension. Adults have better memory for higher frequency sequences and show reduced processing cost for object relative clauses with more frequent subject-noun combinations. Adults are faster to recognize higher frequency phrases compared to lower frequency ones even when all part frequencies are controlled for (e.g., *don't have to worry* vs. *don't have to wait*), suggesting that they represent frequency information about the entire complex form. Similar patterns are found in production: speakers produce higher frequency sequences more quickly; they are more likely to use contractions in higher frequency sequences, and show shorter phonetic duration for the same phonetic material when it appears inside higher frequency chunks (e.g., *don't have to worry* vs. *don't have to wait*). Importantly, the sensitivity to multi-word frequency is not limited to idiomatic phrases or highly frequent collocations, but instead is found for compositional sequences along the frequency continuum. Taken together, these findings show that multi-word chunks continue to be an important part of native knowledge of language.

The potential differential role of multi-word chunks in first and second language learning

One of the long-standing questions in language learning is why children seem to be better language learners than adults, despite being worse at a range of other cognitive tasks. Unlike children learning a first language, adults rarely reach native-like proficiency in a second language. However, contrary to what might be expected from a critical or sensitive period perspective, adults are not always worse than children when it comes to learning language. While adult learners clearly experience problems in many linguistic domains, they do not find all aspects of the novel language equally hard. Older learners, for instance, are generally faster and more efficient in the early stages of learning, and seem to master certain domains (e.g., vocabulary) better than children. More importantly, while some facets of language are learned with relative ease (e.g., vocabulary, word order, yes/no questions), other aspects—such as grammatical gender, article use, and classifiers—continue to pose difficulty even for highly proficient speakers. The currently unresolved challenge is to explain what gives rise to the specific pattern of difficulty for adult language learners.

Part of the answer to this challenge may be that adults are less likely to learn from multi-word chunks, and that this affects the way they learn certain grammatical relations. Whereas children are learning segmentation, meaning and structure at the same time, adults—because of their prior knowledge and different learning situation—will learn from input that is largely segmented into words for which the semantics is already known. This tendency will affect how grammatical relations are learned by changing the information conveyed by the various linguistic elements. To give an example, when presented with the sequence 'la-pelota' [the-ball] in Spanish, an adult learner who already knows what a ball is, can focus on the noun-label at the expense of learning the pairing of the article and the noun. A child learning their first language is more likely to associate the entire article-noun sequence with the meaning of ball, thereby strengthening the link between the article and the noun.

Thus, the suggestion is that children and adults differ in their sensitivity to multi-word chunks and that this can affect learning outcomes. Unlike children, the language of second language learners is often characterized as non-formulaic, and is fraught with non-native idioms and collocations. The use of collocations and formulaic expressions by second language learners is more flexible than that of native-speakers, and even advanced learners produce fewer formulaic sequences than do native-speakers in both spoken and written language. Indeed, when adults do acquire chunked units, they seem to use them differently in learning. Adult learners in immersion settings clearly learn some fixed expressions early on, such as greetings or requests for information. But they do not seem to use them in the same way as children, to further grammatical development. The idea that using multi-word building blocks can enhance learning is supported by a recent study that manipulated the linguistic units participants were exposed to early on. In this study, adults showed better learning of grammatical gender in an artificial language when they were exposed first to sentences (multi-word building blocks) and only then to individual words (single-word building blocks). They learned the association between the article and the noun better when the whole sequence was first associated with meaning, suggesting that there is an effect of early building blocks on learning outcomes.

Summary

Much of the work on language learning focuses on the combinatorial aspects of language: the move from smaller units to larger and more complex ones. Chunk-based learning – where larger chunks are used to discover linguistic units and the relations between them – plays an equally important role in the learning process. While understudied, such processes are found across a range of linguistic domains (phonology, morphology, syntax). Acknowledging the importance of chunk-based language acquisition raises a new set of questions and challenges, and adds new ways of accounting for how language is learned by children and adults. One major challenge lies in identifying the building blocks of language:

how can we discover what children are using as building blocks? One way of addressing this question is by running computational models on child and child-directed speech to identify the most likely building blocks found in children's speech. A second challenge lies in demonstrating the role of larger chunks in learning. Multiple studies document the existence of multi-word chunks in children's speech, yet there is less work showing how they impact the learning process: how do children learn to segment and analyze chunks and how does that affect learning grammar? Here also, the combination of computational and experimental work may prove promising: computational models can be used to identify chunks and generate predictions about their role in learning, which can be tested experimentally. While many questions remain open, the study of chunk-based processes allows us to enrich our understanding of how children acquire language.

Inbal Arnon, Hebrew University & Morten H. Christiansen, Cornell University

See also: Item-based/Exemplar-based learning; Grammatical Gender (Acquisition of); Syntactic development: Construction grammar perspective.

Further readings

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

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: how order-of- acquisition affects what gets learned, *Cognition*, 122, 292-305.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, 19, 241-248.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20, 481-507.

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

Pine, J.M., & Lieven, E.V.M. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language, 20*, 551-551.

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). Austin, TX: Cognitive Science Society

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.