Cognition 175 (2018) 131–140

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

A universal cue for grammatical categories in the input to children: Frequent frames

Steven Moran[a,*], Damián E. Blasi[a,b], Robert Schikowski[a], Aylin C. Küntay[c], Barbara Pfeiler[d], Shanley Allen[e], Sabine Stoll[a]

[a] Department of Comparative Linguistics, University of Zurich, Switzerland
[b] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Germany
[c] Department of Psychology, Koç University, Turkey
[d] Universidad Nacional Autónoma de México, Mexico
[e] Department of Cognitive Science and Psychology, University of Kaiserslautern, Germany

ARTICLE INFO

Keywords:
Statistical learning
Nonadjacent dependency
Cross-linguistic language acquisition
Frequent frames
Input patterns
Child-directed speech

ABSTRACT

How does a child map words to grammatical categories when words are not overtly marked either lexically or prosodically? Recent language acquisition theories have proposed that distributional information encoded in sequences of words or morphemes might play a central role in forming grammatical classes. To test this proposal, we analyze child-directed speech from seven typologically diverse languages to simulate maximum variation in the structures of the world's languages. We ask whether the input to children contains cues for assigning syntactic categories in frequent frames, which are frequently occurring nonadjacent sequences of words or morphemes. In accord with aggregated results from previous studies on individual languages, we find that frequent word frames do not provide a robust distributional pattern for accurately predicting grammatical categories. However, our results show that frames are extremely accurate cues cross-linguistically at the morpheme level. We theorize that the nonadjacent dependency pattern captured by frequent frames is a universal anchor point for learners on the morphological level to detect and categorize grammatical categories. Whether frames also play a role on higher linguistic levels such as words is determined by grammatical features of the individual language.

1. Introduction

Humans learn language through exposure to surrounding speech. Speech is rich with distributional regularities encoded in adjacent and nonadjacent sequences, which reflect grammar constraints. Experimental studies suggest that infants are sensitive to dependencies between sequences and they can use general mechanisms of statistical learning to process and acquire language (for a review see Sandoval & Gómez, 2013). Infants can, for example, segment the speech stream into words given only dependencies between adjacent syllables (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996). But to attain linguistic proficiency, children must also learn to generalize the behavior of words into grammatical categories, so that they can be used productively in syntax.

The mechanisms that children use to assign and remember grammatical category membership are not well understood. How does a child learn to map words to classes when words are not overtly marked, cross-linguistically, neither lexically nor prosodically? Language-specific phonological cues such as stress or segment length (Cassidy & Kelly, 2001) have been shown to facilitate word category assignment (Monaghan, Christiansen, & Chater, 2007). However, not all languages have phonological cues that accurately predict grammatical categories. So how are they learned? One other promising candidate are structural cues such as neighboring words or discontinuous dependencies, that are indicative of grammatical category.

Words belonging to the same category typically behave similarly in similar morphological and syntactic contexts (Bloomfield, 1933; Harris, 1951). Members of the same class, such as 'noun' or 'verb', can be substituted for one another without changing the grammaticality of an utterance. Presumably, these distributional patterns provide input regarding grammatical function to the learner. Maratsos and Chalkley (1980) propose that adjacent sequences in word cooccurrence distributions are a cue for word categorization. Cartwright and Brent (1997) and Redington, Crater, and Finch (1998) use bigram frequencies from natural language and computer simulations to demonstrate categorical learning effects. And Mintz, Newport, and Bever (2002) show that distributional structures in adjacent dependencies (bigram cooccurrences) successfully categorize nouns and verbs in child-directed speech in four English corpora.

* Corresponding author.
E-mail address: steven.moran@uzh.ch (S. Moran).

https://doi.org/10.1016/j.cognition.2018.02.005
Received 17 February 2017; Received in revised form 18 January 2018; Accepted 6 February 2018

Table 1
Results from previous studies.

Language (corpus)	Utterances	Mean accuracy		Mean completeness	
		Words	Morphemes	Words	Morphemes
English (Mintz, 2003)	103,191	0.91		0.12	
Chinese (Xiao et al., 2006)	22,137	0.70			
Dutch (Erkelens, 2009)	49,635	0.71			
French (Chemla et al., 2009)	2006	1.0		0.33	
Spanish (Weisleder & Waxman, 2010)	37,588	0.75			
Turkish (Wang et al., 2011)	37,765	0.47	0.91	0.10	0.06
German (Wang et al., 2011)	5685	0.86	0.88	0.07	0.05
German (Stumper et al., 2011)	30,601	0.77			

In addition to adjacent dependencies, nonadjacent dependencies in natural language exist and they can encode grammatical structures. An example is morphosyntactic agreement, e.g. he **is** sleep**ing**. There is ample evidence that infants make use of nonadjacent dependencies in categorizing elements presented between two repetitive surrounding elements (Gómez, 2002; Gómez & Maye, 2005; Höhle, Weissenborn, Kiefer, Schulz, & Schmitz, 2004; Mintz, 2006; Nazzi, Floccia, Moquet, & Butler, 2009; Nazzi, Barrière, Goyet, Kresh, & Legendre, 2011; Onnis, Monaghan, Christiansen, & Chater, 2004; Santelmann & Jusczyk, 1998; Van Kampen et al., 2008). Artificial language learning experiments also show that learners are sensitive to nonadjacent dependencies (Wang & Mintz, 2016). The simplest nonadjacent dependency is the so-called *frame*, a sequence of three elements, like A_B_C, in which A and C predict information about B. In our example of morphosyntactic agreement, only verbs can appear between the auxiliary verb *is* and the progressive suffix *-ing*. Therefore this nonadjacent dependency, or frame, signals the grammatical class of the intervening element.

Mintz (2003: 91) defines the frame as, "two jointly occurring words with one word intervening", and shows that words A and C in frequently occurring frames accurately categorize the grammatical category of word B in English. Across longitudinal corpora of child-directed speech in parent-child dyads, the results are robust as evaluated by measures of accuracy and completeness. In technical terms, accuracy is equivalent to *precision* in Information Retrieval, i.e. true positives/true positives + false positives (aka a Type I statistical error). And completeness is analogous to *recall*, i.e. true positives/true positives + false negatives (aka a Type II statistical error). In plain speak, accuracy measures how precise is the set of elements selected from a sample. For example, you want to select apples from a bag of apples and pears, but you cannot see in the bag. Out of the pieces of fruit you pick from the bag, how many are apples? Completeness measures how many apples you selected from all the apples present in the bag.

Since Mintz (2003), studies of frequent word frames in languages other than English have had mixed results, summarized in Table 1. French and Spanish frames are a robust cue for word categorization, especially for nouns and verbs (Chemla, Mintz, Bernal, & Christophe, 2009; Weisleder & Waxman, 2010). Frames in Dutch, German and Turkish, however, are not accurate (Erkelens, 2008; Stumper, Bannard, Lieven, & Tomasello, 2011; Wang et al., 2011). Erkelens (2008) found that on all levels of analysis, English frames were more predictive than Dutch frames. Weisleder and Waxman (2010: 1098) conclude in their comparison of Spanish and English frames that "the clarity of the distributional information available in frequent frames varies across languages, and within languages it varies across different distributional environments and grammatical form classes". Additionally, different studies using the same methods on different datasets of the same language obtain different results (see the results for German in Table 1). Wang et al. (2011) study word frames in a small corpus of German and found a high degree of accuracy for frames. Stumper et al. (2011), by contrast analyze a much larger corpus of German child directed speech and they find less robust accuracy for word frames.

Hence, it has become a matter of debate whether the nonadjacent dependency captured by the frame is a universally available pattern to children that might aid in categorization. Most studies analyze frames at the word level, i.e. word1_word3. To account for the differences in morphological and grammatical features in typologically different languages, Wang et al. (2011) propose analyzing frames in languages with richer morphology on the morpheme level. They find both Turkish and German frequent morpheme frames are accurate predictors of the target morpheme's grammatical category (morpheme2). This suggests that the morphological complexity of a language might be relevant for the level of granularity of the units where frames are to be found. Whether this finding translates to other languages, however, is an unresolved issue so far and therefore the focus of this paper.

In this paper, we test whether frequent frames are a universally salient nonadjacent distributional pattern at the word and morpheme levels in child-surrounding speech.[1] In Section 2, we describe our data sample, which includes longitudinal corpora from seven typologically diverse languages. Because the corpora differ in size and the languages differ in their morphological complexity, we operationalize a relative frequency measure to make the data comparable. In Section 3, we evaluate frequent frames in child-surrounding speech in each corpus by accuracy and completeness scores and compare the results to previous findings. These measures, however, do not lend themselves to investigating frequent frames cross-linguistically at the level of specific parts of speech. We therefore propose two novel measurements, called global accuracy and global completeness, and test whether certain parts of speech are more accurately captured in frequent frames across languages. In Section 4, we discuss our results and reasons why frequently occurring nonadjacent dependencies captured by frames are indeed a universally available cue for children.

2. Materials and methods

2.1. The corpora

Linguistic diversity poses many challenges for cognitive science (Evans & Levinson, 2009). In language acquisition studies, it is not practical or even possible to test for statistical patterns across all languages. Instead, we simulate linguistic diversity by examining languages which differ maximally in their grammatical structure. To develop a typologically-diverse set of languages, Stoll and Bickel (2013) applied a fuzzy clustering algorithm used by Rousseeuw and Leonard (1990) that takes as input thousands of languages and their typological feature values (e.g. grammatical case, inflection categories, degree of synthesis, inflectional compactness) as encoded in two broad coverage typological databases: the World Atlas of Linguistic Structures (WALS; Haspelmath, Dryer, Gil, & Comrie, 2008) and AUTOTYP (Bickel et al., 2017). The algorithm outputs five clusters of maximally diverse

[1] Data and source code are available at: https://github.com/acqdiv/frequent-frames.