Root Infinitives and the Acquisition of Morphological Marking

Sarah Payne (sarah.payne@stonybrook.edu)
Department of Linguistics & Institute for Advanced Computational Science, Stony Brook University

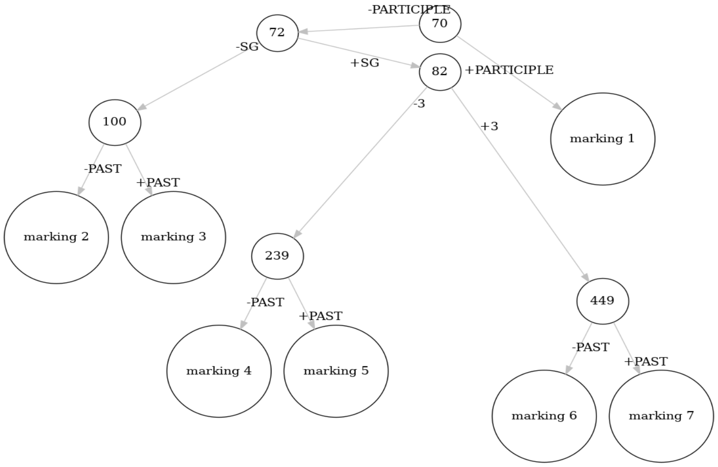
During the **Root Infinitive (RI)** stage of acquisition, children use non-finite verb forms in matrix sentences that require a finite form. Cross-linguistic work has shown that the length of the RI stage is related to properties of the verbal paradigm: its "richness" [1] or how much evidence it provides for a tense-marking grammar [2]. This suggests that RIs are closely intertwined with morphological acquisition, but exactly how these phenomena relate remains an open question. We propose that RIs are a consequence of the acquisition of morphological marking: children must learn which morphosyntactic features are marked in their language from the input [3], and RIs emerge before the child learns that their language marks tense. We present a model that learns which features are marked from developmentally plausible vocabularies, and show that this model (a) matches well with developmental findings on morphological acquisition and (b) correctly predicts cross-linguistic differences in RI as a consequence of an initial lack of sufficient quantitative evidence for tense marking.

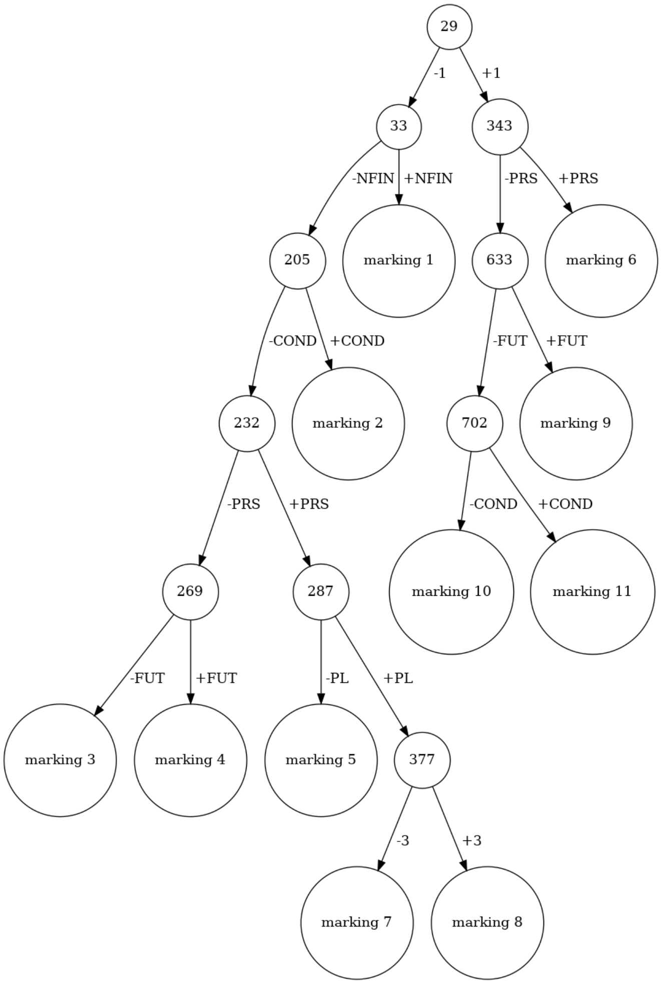
Proposal: Our model employs the **Principle of Contrast** (PC, [4]), the hypothesis that distinct forms generally indicate distinct meanings, and the **Tolerance/Sufficiency Principle** (TSP, [5]), a measure of how much positive evidence is needed to generalize a linguistic rule. We propose that the PC and the child's early segmentation of inflectional morphology [6] allow them to make use of **collisions**: a single stem appearing in multiple inflected forms (e.g. *walk-walking*). Once the child learns what features distinguish individual collisions (e.g. ±PARTICIPLE distinguishes *walk-walking*), we propose that they use the TSP to determine whether there is sufficient evidence for marking across the language (e.g. if enough other verbs in their lexicon exhibit a ±PARTICIPLE collision). If there is sufficient evidence, the child will subdivide their input into +PARTICIPLE and -PARTICIPLE forms, and apply the same learning procedure recursively to each resulting set. Thus, after learning that ±PARTICIPLE is marked in English, the child can go on to learn that ±3,SG is marked for -PARTICIPLE forms, and so on.

Results: To simulate morphological learning, we extract the most frequent English, Spanish, and French verbs from CHILDES as a proxy for early vocabulary. We use the UniMorph annotated morphological database to annotate these verbs with morphosyntactic features and the Lexique French lexical database to phonologically transcribe the French verbs, since some syncretisms in the French paradigm are not reflected in the orthography. The input is provided incrementally to the learner as triples of stem, inflected form, and features (e.g. {*walk, walked*, 3-SG-PAST}), and the model produces tree-like learning traces. Each node in the trace indicates the vocabulary size when the marking of a given feature set is acquired, and the numbers on the leaf nodes indicate order of acquisition. English, French, and Spanish learning traces are shown in Figs 1, 2, and 3 respectively.

Our model matches well with developmental findings on morphological acquisition. Regarding order of acquisition, it correctly predicts the participle and third singular preceding the past tense in English [7,8] and subject agreement mostly preceding tense, aspect, and mood in Spanish [9] and French [10]. Regarding vocabulary size, learning is complete by just 449 inflected forms (188 unique stems) in English, 904 inflected forms (286 stems) in Spanish, and 702 inflected stems (232 stems) in French, matching well with child vocabulary sizes at ages 2-3 in these languages [11].

Our model also correctly predicts cross-linguistic differences in the RI stage. Previous work on RIs considers languages in which agreement typically emerges before tense [1,2,7,8,9,10], and the recursive nature of our model means that it will first subdivide by agreement and then learn tense marking within each agreement node for such languages. Because the TSP tolerates relatively more exceptions for smaller sets [6], we can expect tense to emerge more quickly when there is greater agreement subdivision, thus predicting a shorter RI stage in languages with richer subject agreement. Indeed, our model learns tense marking across the English paradigm by 449 inflected forms (188 stems, Fig 1), French by 343 inflected forms (124 stems, Fig 2), and Spanish by 237 inflected forms (103 stems, Fig 3). The RI phase is indeed shortest in Spanish and longest in English, with French in the middle, as predicted by our model [1, 2]. Our model thus shows that the well-studied RI stage can be accounted for as a consequence of the acquisition of morphological marking: it is the stage before children accumulate sufficient evidence for tense marking.





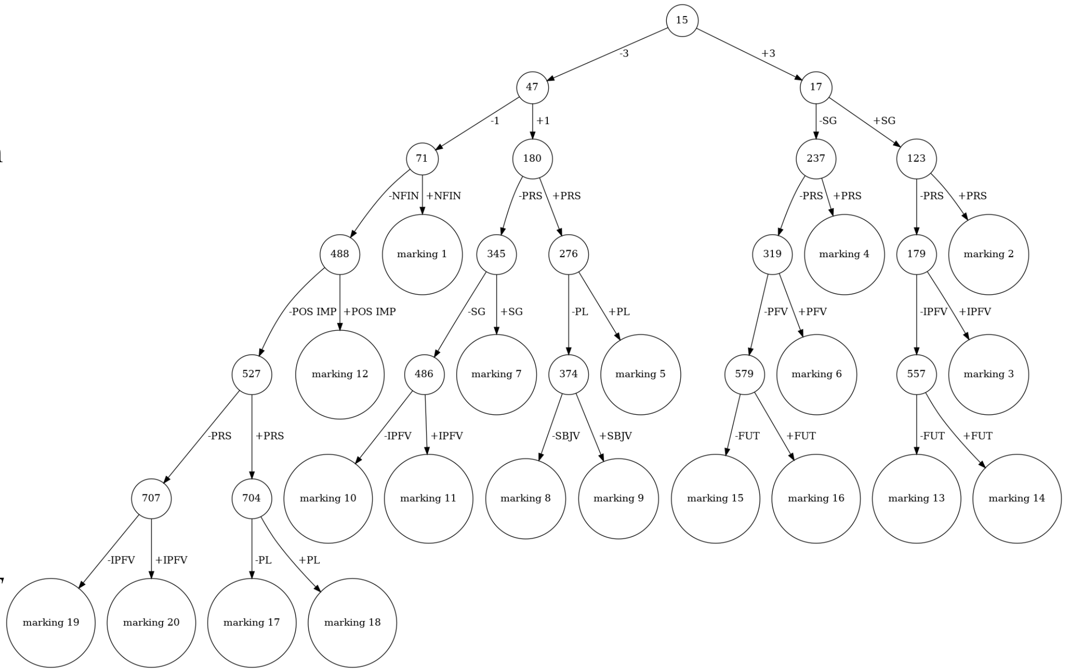


Fig 1 (above), **Fig 2** (top right), & **Fig 3** (bottom right): The learning traces of our model on the English, French, and Spanish CHILDES data, respectively.

References: **[1]** Philips 1996. *Proceedings of the 20th BUCLD.* **[2]** Legate & Yang 2007. *Language Acquisition.* **[3]** Marantz 2013. *Language and Cognitive Processes.* **[4]** Clark & MacWhinney 1987. *Mechanisms of Language Acquisition.* **[5]** Yang 2016. *MIT Press.* **[6]** Kim & Sundara 2021. *Developmental Science.* **[7]** Brown 1973. *Harvard Press.* **[8]** Berko 1958. *Word.* **[9]** Montrul 2004. *John Benjamins Publishing.* **[10]** Prevost 2009. *John Benjamins Publishing.* **[11]** Bornstein et al. 2004. *Child Development.*