The Learnability of Syntactic Islands

Introduction Movement can be found in all languages, but not all languages follow the same set of constraints on movement [1, 2]. We present a formulation of Wh-movement that enables the learnability of island constraints and report results on the distributional learning of strong as well as selective/weak islands from language-specific input data.

Proposal We represent movement as a chain of maximal projections through which the moved element must cross. In a simple structure such as (1a), the moved element originates from its base position and traverses through VP, then TP, and finally lands in the matrix CP (C_MP), or VP→TP→C_MP.

(1) a. [$_{C_MP}$ Who did [$_{TP}$ you [$_{VP}$ see _]]]?

b. [$_{C_MP}$ Who [$_{TP}$ did John [$_{VP}$ think [$_{C_EP}$ (that) [$_{TP}$ Bill [$_{VP}$ saw _]]]]]]?

In general, a movement chain is written as $X_1P \rightarrow X_2P \rightarrow ... \rightarrow X_nP$. The chain is grammatical iff each successive step sequence $X_iP \rightarrow X_{i+1}P$ is grammatical; that is, extraction out of X_iP to $X_{i+1}P$ is allowed. The key question, then, is how children determine the grammaticality of a local sequence $X_iP \rightarrow X_{i+1}P$. Because X_iP is a syntactic category (e.g., Tense, transitive verb, preposition) that may include an open list of members, the child must learn whether $X_i \rightarrow X_{i+1}P$ holds for all/none/some members of X_i based on finite input data. Here we turn to the Tolerance Principle [3]. In particular, if the learner has N members of a category X_i in their vocabulary, $X_iP \rightarrow X_{i+1}P$ holds if at least $(N - N/\ln N)$ of the N members are attested with extraction. If the attested items falls below the Tolerance Principle threshold, the learner will not generalize but simply memorize the specific heads for X_iP that do allow extraction. Because children's Wh-questions obey island constraints from a very early age [4, 5], learning must take place on a relatively small and high frequency vocabulary [6]. As a proxy for children's early vocabulary and experience, our corpus analysis focuses on the most frequent words within syntactic categories and their distributional properties.

Under this view of learning, **strong islands**, from which no movement can ever take place, are trivially accounted for. There would be no data to support, say, the extraction out of the subject or the CP headed by *if/whether* (i.e. subject and *if/whether* island): the the chain NP→TP and TP→C_MP, where M is *if/whether*, are ungrammatical. Similarly, the absence of P-stranding in Spanish, and the absence of long-distance movement in German, are entirely the result of learning. In the more than 460,000 German child-directed utterances in the Leo corpus [7], we could only find one example of long-distance movement: at the level of negligibility. In 87724 sentences of child-directed Spanish, we find 12063 utterances that start with a preposition, but only ever find sentence final prepositions in 22 single-word questions where the wh-element has been ellided. We find a singular example of a proper yes-no question ending in a preposition, as negligible as long-distance movement in German. More interesting cases concern non-islands, from which extraction can take place freely, and selective or weak islands, which allow the extraction of some but not all elements. We turn to a study of these in English.

Non-islands In a 2.7-million-utterance corpus of child directed English [7], we identified 188,000 Wh-questions, of which just over 1,300 involve long-distance extraction out of an embedded CP. Focusing on the matrix questions, we found that of the 100 most frequent verbs, 95 are attested with subject extraction (*who/what*) to support TP→C_MP and 97 are attested with adjunct (*where/why/how/when*) extraction to support VP→TP. And 86 of the most frequent 100 transitive verbs are attested with object extraction (*who/what*) to support VP→TP.

The Markovian formulation of chains facilitates generalization. As noted above, long-distance extractions such as (1b), with the corresponding chain VP→TP→C_EP→VP→TP→C_MP, are very rare. However, the Markovian child only needs to learn TP→C_E and C_EP→VP where C_E stands for the embedded C. This is easy because T and C_E have very few members: ±past for T, and null/*that* for declarative C_E, respectively, which are all attested even in the very modest sample of 1,300 sentences. The other local sequences, which are established from the abundant matrix questions, are "reused" for long-distance movement.

Selective islands Movement constraints may not apply uniformly. For example, English has about 200 CP-taking verbs but relatively few — the so-called bridge verbs [8] — allow movement out of the embedded CP then across them. As a result, most speakers reject (2), taken from [8, p85].

(2) *Who did John complain that he had to do _ this evening? *What did John quip that Mary wore _?

Similarly, as it has been long noted [9], an NP may extract out of certain PPs (3a) but not others (3b):

(3) a. Who did you see a picture of_? What did you receive a request for_?
b. *Who did you admire a book by_? *What color do you buy the car in_?

Under the present account, selective islands are structures X_P for which an insufficient number of X heads in a learner's vocabulary is attested with extraction: the child will only memorize these and do not generalize beyond. The child-directed English corpus contains 24 CP-taking matrix verbs but only 10 are used as bridge verbs, falling below the Tolerance principle threshold (17 needed for 24). Thus bridge verbs may be nothing more than an idiosyncratic subset of CP-embedding verbs [8].

We now turn to the data concerning PP islands (3). We begin by noting that the extraction out of PPs that are attached to a VP is completely free. All top 20 prepositions are attested as stranded inside the PP, which may be either adjunct ("Who did you go with?") or argument ("Who did you give the book to"). And of the top 100 verbs attested with PPs as adjuncts or complements, 92 are attested in p-stranded Wh-question. Thus PP→VP→TP→CP can be generalized as grammatical by the Tolerance Principle. However, when a PP is attached to an NP as in (3), extraction out of the PP is much more restricted. We manually examined all Wh-questions that end with an NP followed by a stranded preposition to ensure that the PP is attached to the NP rather than some higher VP. We found that only 3 stranded prepositions (*about*, *of*, and *from*) in such positions, far below the Tolerance threshold for generalization: the top 20 requires 14. Therefore, the sequence PP→NP cannot be generalized for all prepositions but will be limited to those attested in the input data.

Discussion It is useful to compare our approach with the Pearl-Sprouse learning model of islands [10] which, like ours, decomposes movement into local sequences. The absence of extraction across strong islands facilitates acquisition under this model as easily as ours. However, the Pearl-Sprouse model is probabilistic; the goodness of local sequences is measured by their *token* frequencies. This leads to serious problems for non-islands, among others. If some sequences are attested very frequently, other sequences become unlikely even though they are grammatical. For example, because the complementizer in long-distance movement is almost always null, the probability of crossing the complementizer *that* is almost zero, leading the model to reject such examples [11]. Our model, by contrast, is non-probabilistic and forms generalization over types. The embedded complementizer has two options: null and *that*, and both are attested, and their token frequency disparity plays no role in generalization.

Given the range of cross-linguistic differences, and the level of idiosyncrasy within a single language, a learning theory must be a significant part of any theory of islands especially under the Minimalist conception of UG [12]. We regard as our approach as a null hypothesis: islandhood is established entirely empirically. As long as movement does not violate true universals of language (e.g., hierarchy and structure dependence), everything can be an island but nothing must be one. The possibility of selective islands, and rote memorization of specific items that allow extraction (e.g., bridge verbs), suggest a limit on the orderliness of island constraints. At the same time, language-internal theoretical explanations would be strengthened and better motivated if its coverage of the data surpasses a theory strictly based on distributional learning.

References [1] Rizzi 1982. Foris. [2] Engdahl 1986. Reidel. [3] Yang 2016. MIT Press. [4] Otsu 1981. MIT dissertation. [5] De Villiers et al. 1990. In *Language processing and language acquisition*. [6] Fenson et al. 1994. SRCD Monograph. [7] MacWhinney 2000. Erlbaum. [8] Chomsky 1977. In *Formal syntax* [9] Chomsky 1973. In *A Festschrift for Morris Halle*. [10] Pearl & Sprouse. *Lg Acq*. [11] Phillips 2013. *Lg Cog Proc*. [12] Berwick & Chomsky 2016. MIT Press.