Learning the surface structure of *wh*-questions in English and French with a non-parametric Bayesian model

An Nguyen, Colin Wilson
Johns Hopkins University, Department of Cognitive Science
an.nguyen@jhu.edu, colin.wilson@jhu.edu

Introduction. There is considerable variation in the surface structure of *wh*-questions (e.g. Soare, 2007). Across languages, questions can differ with respect to overt displacement of *wh*-phrases (i.e., *ex-situ* vs. *in-situ* questions), inversion of subjects and auxiliaries or verbs, presence of a Q-marker, and prosodic properties such as pitch accent. These differences present a problem for language acquisition: How does a child learn the grammatical surface structure of *wh*-questions in her native language(s), given that within a language there can be typically multiple *wh*-question types, each with its own characteristic surface properties?

In this paper, we propose a model for learning *the number of wh-question types* in a given language and *the properties of each type*. The idea pursued here is that a child learns the number of *wh*-question types in her language and their surface structures by *clustering* observed utterances. Because the child cannot know a priori how many types are present, our model specifically employs *non-parametric clustering*. This allows the number of types to be learned from the input data rather than specified beforehand. We focus here on overt properties such as displacement, morphological marking, and prosody.

The learning problem. In English, there are at least three types of *wh*-question available to children in child-directed speech (CDS), namely fronted information-seeking questions, in-situ probe questions, and in-situ echo questions (Nguyen and Legendre, 2020). Fronted questions and probe questions are both information-seeking and have similar prosody: the *wh*-word typically has flat/falling fundamental frequency (F0) and receives no stress accent (Reinhardt, 2019). Echo questions, on the other hand, are typically used to ask for repetition of a previous utterance rather than to request new information. Echo questions prosody is distinctive: the *wh*-word typically has high rising pitch and receives heavy stress (cued in part by longer duration) (Artstein, 2012; Cheng and Rooryck, 2002).

Children learning English must determine the number of *wh*-question variants present in the language and the characteristics of each type. They must learn that while fronted questions and probe questions have different morphosyntactic properties, they are both information-seeking and can be used in similar contexts. They must also learn that depending on the prosody, an in-situ question can call for a new-information response (probe questions) or a repetition/clarification (echo questions).

The model. We adopt a non-parametric approach because, unlike classical parametric finite mixture models, it does not force the learner to commit to the existence of a particular number of clusters (question types) in advance of analyzing the input data. For purposes of implementation, we place an upper bound of $K = 10$ on the number of *wh*-question types that the model can learn, and examine only overt morphosyntactic and prosodic properties. The morphosyntactic properties are discrete variables that can take on two values (1 = presence and 0 = absence). They include the position of the *wh*-word, the inversion status of the auxiliary, and the presence of a Q-marker. The prosodic properties consist of two continuous variables: the duration and F0 contour of the *wh*-word. The non-parametric model proposed here is technically a truncated Dirichlet Process Mixture Model (e.g. Gershman and Blei, 2012), as specified below.

The probability that the ith question utterance, represented as three binary morphopsyntactic variables and two continuous prosodic variables, belongs to question type k is given by Bayes' Rule:

$$p(k|\mathbf{y}_i) = \frac{p(\mathbf{y}_i|k)\ w_k}{\sum_{k=1}^{K} p(\mathbf{y}_i|k')\ w_{k'}}$$

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 403-405.
Held on-line February 14-19, 2021

Cluster probabilities

$$\begin{aligned}
\alpha &\sim Gamma(1,1) \\
v_\ell \mid \alpha &\sim Beta(1,\alpha) \qquad \text{for } \ell = 1,\ldots,K-1 \\
\alpha &\sim Gamma(1,1) \\
w_1 &= v_1 \\
w_k &= v_k \prod_{\ell=2}^{k-1} 1 - v_\ell \qquad \text{for } k = 2,\ldots,K-1 \\
w_K &= \prod_{\ell=1}^{K-1} 1 - v_\ell
\end{aligned}$$

Parameters of each cluster

$$\begin{aligned}
p_{kj} &\sim Beta(1,1) \qquad \text{for } j = 1,\ldots,J \\
\mu_{k\ell} &\sim Normal(M_\ell, S_\ell) \qquad \text{for } \ell = 1,\ldots,L \\
\log \sigma_{k\ell} &\sim Normal(4,2)
\end{aligned}$$

Distribution of observations

$$\begin{aligned}
p(\mathbf{y}_i|\mathbf{w},\mathbf{p},\mu,\sigma) &\qquad \text{for } n = 1,\ldots,N \\
= \sum_{k=1}^{K} w_k \prod_{j=1}^{J} &Bernoulli(y_{ij}|p_{kj}) \\
&\cdot \prod_{\ell=1}^{L} Normal(y_{i\ell}|\mu_{k\ell},\sigma_{k\ell})
\end{aligned}$$

where each $p(\mathbf{y}_i|k)$ is a product of three $Bernoulli$ probabilities and two $Normal$ densities. The mixture weights w_K are given by a stick-breaking process (Sethuraman, 1994).

Data and results. The model was trained on 2000 simulated instances of the 3 questions types. 88 English *wh*-questions were collected from four CHILDES audio corpora, HSLLD (Dickinson and Tabors, 2001), Snow (MacWhinney and Snow, 1990), Van Houten (Van Houten, 1986), and Weist (Weist and Zevenbergen, 2008), for testing. The frequency distributions of question types in the simulated data matched distribution in CDS. Inference proceeded by MCMC sampling for 5000 iterations with the initial 2500 samples discarded as burn-in. Trace plots indicated that all parameters settled on stable values within the burn-in period, therefore without loss of detail we present only average values over the remaining 2500 samples. The sampling run shown in Table 1b (see appendix A) converged on three clusters, ordered in descending probability, that closely approximate the actual *wh*-question types in the training data. (The other clusters inferred by the model had a total probability of 0.05 and are ignored here as noise.) The model accurately classified 97.7% of the simulated question utterances on which it was trained, and 86.0% of the natural CDS test utterances. The main confusion for test utterances was misclassification of echo questions as probe questions (i.e., as the somewhat more frequent type of in-situ question).

Discussion. We have described a non-parametric Bayesian model that infers the number of *wh*-variants within a language and classifies question utterances accordingly. The model does fairly well in terms of identifying the number of clusters. The current model addresses the acquisition problem posed by differences in the surface properties of *wh*-questions across and within languages. While much work remains to be done to expand the range of variation that the model can accommodate, to integrate the model with other approaches to the acquisition of *wh*-questions (e.g., Pearl and Sprouse, 2013) and other aspects of syntax, and to synergistically combine the learning of surface properties with that of semantics and pragmatics, the present results show the promise of applying non-parametric Bayesian methods to cross-linguistic and especially language-internal syntactic variation.

References

Ron Artstein. 2012. A focus semantics for echo questions. In Ágnes Bende-Farkas and Arndt Riester, editors, *Workshop on Information Structure in Context*, pages 98–107. IMS, University of Stuttgart.

Lisa Cheng and Johan Rooryck. 2002. Licensing wh-in situ. *Syntax*, 3:1–19.

David Dickinson and Patton Tabors. 2001. *Beginning literacy with language: Young children learning at home and school*. Paul Brookes Publishing, Baltimore, MD.

Samuel J Gershman and David M Blei. 2012. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Brian MacWhinney and Catherine Snow. 1990. The child language data exchange system: An update. *Journal of Child Language*, 17:457–472.

An Nguyen and Géraldine Legendre. 2020. Covert movement in English probing wh-questions. *Proceedings of Linguistic Society of America 2020 Annual Meeting*, 5(1):180–186.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax

and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Janina Reinhardt. 2019. *Regularity and variation in French direct interrogatives: The morpho-syntax and intonation of question forms in reality TV shows, audio books and teaching materials*. Ph.D. thesis, University of Konstanz.

Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Gabriela Soare. 2007. A cross-linguistic typology of question formation and the antisymmetry hypothesis. *Generative Grammar in Geneva*, 5:107–131.

Lori Van Houten. 1986. Role of maternal input in the acquisition process: The communicative strategies of adolescent and older mothers with their language learning children. Paper presented at the 11th Annual Boston University Conference on Language Development.

Richard Weist and Andrea Zevenbergen. 2008. Autobiographical memory and past time reference. *Language Learning and Development*, 4(4):291–308.

A Appendices

Table 1. English question types based on CDS (a) and inferred by the non-parametric Bayesian model (b)

(a)

Type	**Proportion**	**WhFront**	**Inver**	**InitialQ**	**WhDur (sd)**	**WhΔF0 (sd)**
Fronted	.84	1	1	0	150 (49)	-6 (36)
Probe	.09	0	0	0	208 (85)	-21 (61)
Echo	.07	0	0	0	254 (60)	108 (64)

(b)

Cluster	w	$p_{WhFront}$	p_{Inver}	$p_{InitialQ}$	μ_{WhDur} (σ)	$\mu_{Wh\Delta F0}$ (σ)
$1 \approx$ Fronted	0.79	1.0	1.0	0	119 (48)	4 (35)
$2 \approx$ Probe	0.11	0	0	0	203 (78)	-8 (68)
$3 \approx$ Echo	0.05	0.01	0.01	0.01	270 (49)	120 (58)