Functional category induction with theory-neutral cognitive biases

Mila Marcheva[1] Theresa Biberauer[2,3,4] Weiwei Sun[1]
[1]Department of Computer Science & Technology, University of Cambridge
[2]Department of Theoretical and Applied Linguistics, University of Cambridge
[3]General Linguistics Department, Stellenbosch University [4]Linguistics Department, University of the Western Cape

Abstract

This paper probes the influence of a particular kind of domain-general cognitive bias in first language acquisition with the aid of computational models. We introduce a novel task: inducing functional categories from morphemically tokenised sentences, and supply a manually annotated dataset of English child-directed speech (CDS). We operationalise a widely assumed type of cognitive bias, "less-is-more", as three computational principles—ordering input, gradually increasing model complexity, and priming the learner—and develop a theory-neutral experimental setup to evaluate their impact on functional category induction. Our experiments with CDS demonstrate that models incorporating reflexes of "less-is-more" outperform the purely statistical baseline. As part of our exploration of ordering effects, we employ the morpheme acquisition order proposed by Brown (1973) and, for the first time in the literature, present statistical evidence that Brown-compliant orders outperform non-Brown-compliant ones.

Keywords: Brown's order of acquisition; domain-general cognitive biases; functional categories; online learning

Introduction

A core syntactic aspect of first language acquisition (FLA) is acquiring functional categories (FCs) as well as their surface forms (Dye, Kedar, & Lust, 2018). FCs are the linguistic components that encode grammatical content, in contrast to lexical categories, which encode semantic content. This paper investigates the influence of a particular kind of domain-general cognitive bias on FC acquisition with the aid of computational models. We situate our discussion in relation to the novel task of functional category induction (FCI), supplied with a manually annotated dataset.

We operationalise a widely assumed type of cognitive bias, "less-is-more" (Newport, 1990; Biberauer, 2019b), as three computational principles: **ordering** the training input, gradually **unlocking** the available number of categories, and **priming** the learner with small amounts of information. We design and implement an experimental framework for FCI based on Online Expectation Maximisation (Liang & Klein, 2009) and Hidden Markov Models, and explore how the three principles affect FCI. Finally, we present a case study centring on the attested order in which functional morphemes are acquired in English, first documented in Brown (1973).

The three principles greatly benefit the learning outcome: the augmented model achieves 59.9% FCI accuracy, compared to 32.1% for the baseline. Furthermore, only the augmented model achieves linguistically plausible performance and generalises beyond the primed information. Finally, for the first time in the literature, we computationally demonstrate the superiority of Brown's order: Brown-compliant orders achieve statistically significantly higher log-likelihood than non-Brown-compliant ones.

Background

Functional categories as the core of syntax This paper is concerned with functional items because they shed light on the grammatical organisation of languages as per the *Borer-Chomsky Conjecture* (Baker, 2008). For our proposed model, we build on the "lexical-before-functional" pattern established by acquisitionists (Gentner & Boroditsky, 2001; Dye et al., 2018). Specifically, we assume that initial knowledge of nouns and verbs is already in place before the model is confronted with the functional items studied by Brown (1973). According to the *edge significance* hypothesis, (unanalysed) functional items are key during the lexical stage by virtue of their being high-frequency edge-marking elements, which facilitate confirmation of the categorial status of the content items they occupy predictable peripheral positions in relation to (Biberauer, 2019a).

Our focus on functional items and the *edge significance* hypothesis motivates morphemic tokenisation. In English, which is a largely isolating language, most functional items manifest as free morphemes (independent words). However, some bound functional morphemes exist as affixes. Existing computational operationalisations overlook bound morphemes, treating forms like *runs* as atomic tokens. This then leads to *runs* and *run* being stored as separate vocabulary entries. If the bound morpheme *-s* were tokenised, there would instead be a single entry for *run* and one for *-s*.

Cognitive biases in language acquisition Some linguists consider humans' general cognitive faculty as solely responsible for FLA and others (generativists) postulate an additional language faculty (Ambridge & Lieven, 2011; Guasti, 2017). Today domain-general cognitive biases are, however, widely recognised as playing a role in shaping the structure of linguistic systems (Chomsky, 2005; Saffran & Thiessen, 2007). This makes them very adaptable for computational experimentation as their use precludes the need to subscribe to a specific linguistic framework.

Recently, the focus has been on "less-is-more"-type biases (Biberauer, 2019a; Culbertson, Compostella, & Kirby, 2024). "Less-is more" (Newport, 1990) poses learning in *stages* of increasing *complexity* as a key factor in the success of FLA and is thought to be prevalent in different components of acquisition, including functional category acquisition (Bosch & Biberauer, 2024). Another insight from "less-is-more" is that learning and development co-occur: the input from the environment stays static, but the learner is dynamic (Elman, 1993; Lidz & Gagliardi, 2015). Thus, "less-is-more" ties learning in stages to the distinction between *input* and *intake*, which

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Stage	MLU	Functional morphemes
I	< 2	none: the child-produced utterances consist of bare (non-inflected) nouns and verbs
II	2-2.5	(1) present progressive *-ing*, (2) preposition *in*, (3) preposition *on*, (4) regular plural *-s*
III	2.5-3	(5) past irregular (e.g. *ran*), (6) possessive *'s*, (7) uncontractible copula (e.g. *is*)
IV	3-3.75	(8) articles *the*, *a*, (9) past regular *-ed*, (10) third person regular *-s*
V	3.75-4.5	(11) third person irregular (e.g. *has*), (12) uncontractible auxiliary (e.g. *is*), (13) contractible copula (e.g. *'m*), (14) contractible auxiliary (e.g. *'m*)

Table 1: Brown's acquisition order (Brown, 1973). Morphs are identified in *blue* and the functional morphemes are in black. Mean Length of Utterance (MLU) is a language-development measurement, which measures the length of child speech utterances at a given time and counts functional morphemes as separate units from lexical morphemes.

is often overlooked in computational work. The *input* entails all the utterances a child hears, whereas the *intake* is the subset of the input that the child is able to attend to (Wijnen, 2000). The "less-is-more" interpretation of *intake* is that this constitutes the utterances that allow acquirers to maximise the use of their currently accessible resources—what they already know about the grammar they are acquiring—in order to make sense of the overall input (Biberauer, 2018).

We assume that the *intake* will become successively more complex as the acquirer's grammatical knowledge increases as per Evers and van Kampen (2008); Lidz and Gagliardi (2015); Dye et al. (2018); Biberauer (2018, 2019a). This leads to the idea of ordering training data with increasing complexity to simulate intake. Aspects of the input that are accessible to the acquirer serve a **priming** function, facilitating further acquisition, while inaccessible aspects do not contribute.

Brown's order Brown (1973) conducted a longitudinal study of three North-American-English-acquiring children, and observed that the emergence of the first *functional morphemes* follows the staged order shown in Table 1. Note the terminological distinction between functional morpheme (the abstract general type of the morpheme) and morph (which is the manifestation of the abstract functional morpheme) (Haspelmath, 2020); for example: the indefinite article is a functional morpheme and in English it is expressed by two morphs *a* and *an*. Mean Length of Utterance (MLU) defines the stages, and functional items only begin to be systematically produced when MLU exceeds 2. The emergence in stages is confirmed by more recent works (van Kampen, 2004; Dye et al., 2018; Bosch & Biberauer, 2024).

All functional morphemes in a given stage need to be acquired before the functional morphemes from the next stage can be acquired, but the order within the stage can differ across children. For example, all morphemes from stage II need to be acquired, regardless of the order, before the child can acquire possessive *'s*, which is from stage III. We denote the set of orders resulting from within-MLU-stage shuffling as *Brown-compliant*. Brown (1973, p. 317) presents the mean order of acquisition of the 14 functional morphemes, which is equivalent to the order (1)-(14) in Table 1; this is what is commonly referred to as *Brown's order*.

Hidden Markov Model (HMM) An HMM is a chain-like structure representing a Markov process of hidden states and observations. The essential components of an HMM are: transition T_{ij}, emission E_{iw}, and start S_i probabilities. We use an HMM to approximate the problem of FCI, where tokens (morphs) are the observations and their functional categories are the hidden states. For detailed textbook overview of HMMs refer to Jurafsky and Martin (2009).

Online Expectation Maximisation (EM) Expectation Maximisation (Dempster, Laird, & Rubin, 1977) is a probabilistic framework which can estimate the parameters of a generative model, such as an HMM, based only on observed data. Online EM (Liang & Klein, 2009) processes batches leading to more frequent updates and decreased computational load than standard EM. The training objective (see eq. 1) is to maximise the log-likelihood (*LL*) of the sequence of observables $\mathbf{x}$, which is determined by the sequence of corresponding hidden states $\mathbf{z}$. Online EM with batch size one emulates child processing during FLA: a linear stream of utterances, arriving one by one. Another aspect that makes Online EM cognitively plausible is that it performs soft assignment: a token can be emitted from several clusters, which is desirable as a token may belong to more than one category (e.g. *run* can be both a noun and a verb). Finally, statistical modelling, of which Online EM is an example, is well suited for testing language acquisition hypotheses, as exhibited by past work (Yang, 2004; Pearl & Goldwater, 2016). Empirical evidence suggests that statistical learning is part of humans' domain-general skillset (Saffran & Kirkham, 2018).

$$LL \equiv \log p(\mathbf{x}) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \tag{1}$$

Functional category induction (FCI)

We simulate a new computational task—**functional category induction (FCI)**. The input for FCI is morphemically tokenised utterances and the output is categorially labelled utterances. The task is illustrated in Figure 1.

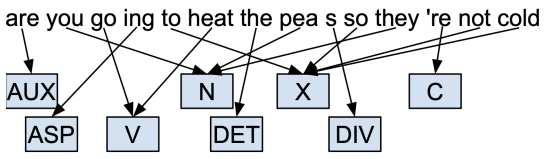

Figure 1: A sentence tokenised and annotated in accordance with the task of functional category induction (FCI).

FCI is similar to the existing task of POS induction: tokens need to be grouped in (labelled) categories based on their syntactic affiliation. The putative target categories for FCI are listed in Table 2: as the primary goal of this project is scientific exploration, we consider a customised subset of categories as opposed to one of the complete POS sets (Marcus, Marcinkiewicz, & Santorini, 1993; Petrov, Das, & McDonald, 2012). At MLU stage V (see Table 1), we estimate there to be 10 functional category clusters as displayed in Table 2. Justification for the groupings of functional morphemes into clusters and detailed annotation guidelines are included in the dataset release[1].

Id	Label	Short description
0	N	nouns
1	V	verbs, irregular past
2	X	hold-all
3	ASP	progressive aspect
4	P	prepositions
5	DIV	plural noun morpheme
6	DET	possessive clitic, articles
7	T	regular past, third person present
8	AUX	auxiliary verbs
9	COP	copular verbs

Table 2: FCI category labels and descriptions.

Data

We curated a corpus of North American English child-directed speech (CDS) from several CHILDES derived corpora (MacWhinney, 2014; Pearl & Sprouse, 2013; Brent & Siskind, 2001; Phillips & Pearl, 2015). The training dataset consists of 295K sentences and 1,645M tokens, where we count both functional and lexical items akin to the method used for measuring MLU. For the purposes of cognitive plausibility, the dataset size reflects the quantity of data children have available during FLA until roughly age five, as stipulated in previous works[2]. We manually annotate a dataset of 1,4K utterances for testing.

Morphemic tokenisation of the data is performed using the `en_core_web_lg spaCy` model (Montani et al., 2020). Additional rule-based tokenisation separates the functional morphemes included in Brown's order. Punctuation tokens are excluded due to their debatable role for language acquisition. After the automatic tokenisation, a subset of the CDS is manually verified on 2,9K sentences (1% of the whole corpus) from the training data to show $< 1\%$ error rate. The annotated corpus consists of 100 randomly selected sentences for each of the 14 Brown morphemes. The sentences are annotated with the labels from Table 2.

[1] `https://huggingface.co/datasets/milamarcheva/morphemically_tokenised_english_cds`

[2] The lower bound is 736K words between ages three and five as per (Pearl & Sprouse, 2021, Table 1) and the upper bound 50M tokens until age five as per (Warstadt & Bowman, 2022, Figure 1). It is challenging to provide a precise estimate due to high individual variation between acquirers (Hart & Risley, 1995).

Experimental design

We set out to explore how "less-is-more"-type cognitive biases can be incorporated in a statistical learning framework and what their effect would be on the learning process and the final learned categories. The experimental framework allows us to probe the effect of three factors (order, gradual unlocking, and priming) which we derived from our consideration of the role of cognitive biases in FLA. We expect that each of these factors will benefit the learning routine and outcome independently, but hypothesise that the best outcome will be achieved when using all three simultaneously. In that idealised case, we simulate *intake*: first, the learner is focused on only one type of functional morpheme at a time; second, the input for the duration of that phase comprises only examples that contain an instance of the functional morpheme(s) currently being learned.

Computational operationalisation

Initialisation:
`primeCluster`$(0,$ BLC nouns$)$
`primeCluster`$(1,$ Swadesh verbs$)$
`lockCluster`(3) Lock clusters 3–10
select utterances without Brown morphemes
estimate HMM parameters with Online EM
$k \leftarrow 3$
Functional category learning:
iterate f over the 14 functional morphemes:
`primeCluster`(k, morph_f)
$k \leftarrow k+1$
`lockCluster`(k)
select utterances for functional morpheme f
estimate HMM parameters with Online EM

Figure 2: Online learning with order, unlocking, and priming. The implementations of `lockCLuster` and `primeCLuster` are explained in §**Gradual unlocking** and §**Priming**.

The baseline model is Online EM coupled with a first-order HMM as implemented by Liang and Klein (2009). Unless otherwise stated, all experiments use the full dataset of 295K sentences, one iteration, batches of size one, and 10 clusters, as those parameters are justified with respect to cognitive plausibility in earlier sections.

The fully augmented model[3] with ordering, gradual unlocking and priming is outlined in Figure 2 and described below. First we need to perform content word initialisation, following the "lexical-before-functional" pattern. Cluster 0 and cluster 1 are initialised to emit nouns and verbs with uniform probability correspondingly, and cluster 2 is left uninitialised as a hold-all. Online EM is performed using the portion of training utterances that do not contain Brown morphemes. For the initialisation with nouns we use Basic Level Categories (BLC) nouns, which children are observed to acquire first (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976),

[3] `https://github.com/milamarcheva/unsupervised-modeling-cds`

and specifically we use the list from Chen and Teufel (2021). There is no BLC list for verbs, so instead we use the Swadesh verb list, which provides a set of basic lexical items that persist over time (Swadesh, 1955).

After the content word initialisation, we proceed to functional category learning. We introduce distinct phases of learning during each of which the focus is on only one functional category at a time. The routine traverses successively over all functional morpheme types (see Table 1). After a cluster is primed for a given functional category via one morph that manifests it, only examples containing that type of functional morpheme are presented to the learner during the relevant phase.

Ordering training data

In machine learning, training models by ordering examples from simple to more complex is known as *curriculum learning* (Bengio, Louradour, Collobert, & Weston, 2009). Ordering examples by increasing complexity is intrinsic to "less-is-more" and cognitively justified in the distinction between *input* and *intake*. We explore various complexity approximations which result in distinct orderings and experiments. The linguistic complexity of a sentence can be approximated using heuristic (sentence length; Brown's order), probabilistic (sentence probability), or statistical (frequency) **complexity approximations (CAs)**.

Sentence length, measured as the total number of tokens in a sentence, is a simple CA to start with. It is motivated in how child speech production steadily increases in length (Dye et al., 2018), but note that production is known to underrepresent competence (Guasti, 2017).

A more sophisticated heuristic approximation is **Browns's attested order of acquisition of functional morphemes** (Brown, 1973). Brown found that functional morphemes emerge in correlation with MLU stages, as detailed in Table 1. The functional morphemes associated with a given MLU stage can appear in any order within that stage, but all the morphemes from a given MLU stage will be learned before the child moves on to morphemes from subsequent MLU stages. Thus, we can distinguish between Brown-compliant and non-Brown-compliant orders. For experiments where we need to use one order we use the mean order of acquisition (Brown, 1973, p. 317).

Normalised sentence log-probability $\mathbb{P}(s)$ is calculated by dividing the sentence log-probability (Sap, Horvitz, Choi, Smith, & Pennebaker, 2020) by the number of tokens in the sentence: $\mathbb{P}(s) \equiv N \log p(s) = \frac{\sum_j \log p(t_j|t_{1,j-1})}{j}$. We derive sentence log-probabilities from `GPT2LMHeadModel` (Radford et al., 2018; Wolf et al., 2019). A language model is built on the basis of large volumes of textual data and the likelihood of each token is evaluated in the context of all preceding tokens. Although a language model is not trained on CDS, it captures latent statistics in language, making it an appropriate CA.

The **frequency** of the different Brown morphemes appearing in the corpus can be used to rank these morphemes from simplest (most frequent) to most complex (least frequent). Brown (1973) also performed frequency analysis, but did not find frequency to determine the order of acquisition, and the lesser importance of frequency has since been replicated (Demuth, 2007; Lieven, 2010). We therefore include frequency for completeness only.

Gradual unlocking

Normally all clusters of an HMM are available for training. However, in the spirit of "less-is-more", in order to simulate the children's growing cognitive capacity during FLA, we perform gradual unlocking of the clusters. Gradual unlocking means that at the start of training only a few of the HMM hidden states are available, and more hidden states are unlocked one by one in the process of training. Here the number of states available for training corresponds to the number of syntactic categories a child is assumed to be aware of at a given time. In Figure 2 learning starts with 3 clusters only: N, V, and the hold-all X, as per the "lexical-before-functional" pattern. A cluster i is locked by setting to 0 its start probability S_i as well as the transition probabilities T_{ji} of all other clusters j to the locked cluster i. If a functional morpheme needs to be assigned to a new cluster, as opposed to the clusters already in use, then a new cluster will be unlocked.

Priming

Priming is linguistically motivated because FLA exhibits clear "explosive bursts" in which children acquire new words and structures, after what Snyder designates a *First Regular Use* (Snyder, 2021). The `primeCLuster` function from Figure 2 takes as input a cluster index k and a single morph m, which manifests the corresponding functional morpheme of current focus; then the cluster's emission probabilities E_{km} are updated to emit the morph m with increased probability.

Priming as implemented here assumes that during FLA children first acquire coarse-grained lexical categories, which is supported by what is known about FLA more generally (Dye et al., 2018). Afterwards, children begin to notice inflectional morphemes that frequently appear in stable positions in relation to the members of the early-acquired coarse-grained lexical categories (see i.a. Wexler (1998) on children as "little inflection machines"). Priming for a functional category occurs before the examples containing the functional category are presented to the learner. This equates to a child noticing a functional morpheme, assigning it a category, and then adjusting its *intake* to focus on input that contains the new morpheme.

Results and analysis

We evaluate the results using many-to-1 accuracy (Johnson, 2007), variation of information (Meilă, 2003), and the final training log-likelihood (Equation 1, *LL*). We further perform phenomenon-driven evaluation and ultimately find that Brown-compliant orders of phasing significantly outperform random ones (see §**Brown-compliant vs random orders**).

	O	**U**	**P**	*LL*	**ACC**	**VI**
1	−	−	−	-5.66	32.1%	4.33
2	+	−	−	-5.76	30.8%	3.75
3	−	+	−	-5.65	30.7%	3.56
4	−	−	+	-5.21; .0106	57.1%; .0161	3.53; .0914
5	−	+	+	-5.30; .00876	55.7%; .00674	3.46; .0301
6	+	+	−	-5.20	30.7%	3.56
7	+	−	+	**-4.00**; .0531	**59.9%**; .0336	**3.45**; .164
8	+	+	+	-4.16; .0464	54.8%; .0307	3.64; .163

Table 3: Final log-likelihood (*LL*), accuracy (ACC), and variation of information (VI) for combinations of Brown **O**rder, gradual **U**nlocking and **P**riming. High values of LL and ACC and low values of VI indicate good performance. Results with priming are in the form mean; st.d. agglomerating 500 experiments, reflecting there are various combinations of morphs.

The effects of all possible combinations of order, gradual unlocking and priming, are displayed in Table 3. **P**riming has the largest positive effect on the results, even without **O**rder or gradual **U**nlocking (see row 4). The highest *LL*, ACC, and VI (row 7) occur from the simultaneous use of **O**rdering and **P**riming, which coupled together simulate *intake*, a concept intrinsic to "less-is-more" (see §**Cognitive biases in language acquisition**). To elaborate, when Priming and Ordering are combined, the primed category—corresponding to a mentally represented abstract category—is supplied with appropriately representative examples during training. This primed category enables specific tokens in the training data (the input) to be recognized and interpreted as intake. Contrary to hypothesised, the three factors combined (row 8) yield worse accuracy and VI results in comparison with other combinations of the three factors (rows 4, 5, 7), but achieve the second highest *LL*; we suspect this is due to the effect of gradual **U**nlocking. Gradual unlocking does not work well without priming because without manipulating the emission probabilities (see §**Priming**) for a newly unlocked cluster, the cluster remains unused.

Results for experiments with priming (rows 4, 5, 7, 8 from Table 3) are the mean and st.d. of 500 experiments, as there are various combinations of morphs that can be used. For each abstract functional morpheme stage, we prime the cluster with only a single morph manifesting that functional morpheme. Due to computational limitations, it is not possible to run an exhaustive experimental suite for the 35K combinations of morphs possible, so instead we select 500 combinations at random and reuse them across the experiments. The strong positive effect of priming with just one morph indicates that priming enables generalisation based on the small amounts of primed information.

Table 4 summarises the final *LL* results for the four different complexity approximations (CA)s. These results go against the hypothesis that ordering training data with increasing complexity should improve performance, which might be an indicator that our CAs (except $\mathbb{P}(s)$ which supports the hypothesis) are not appropriate. Additionally, none of the CAs achieve a better *LL* than the baseline (row 1 in Table 3) where the training data is shuffled. The shuffled training data introduces randomness and better distribution of the morphemes into different clusters as an effect of that. The lack of improvement when only order is manipulated further emphasises the importance of *intake* in acquisition, i.e. order coupled with priming (rows 7 and 8 in Table 3).

	Length	$\mathbb{P}(s)$	Brown	F
↑	-5.952	**-5.942**	-5.756	-5.757
↓	**-5.940**	-5.968	**-5.724**	**-5.711**

Table 4: Final *LL* for Online EM using different complexity approximations (Length, Normalised sentence log-probability $\mathbb{P}(s)$, Brown, Frequency) resulting in different orders, with ↑ increasing or ↓ decreasing complexity.

Generalisation beyond the primed morphemes

The following analysis is based on the 10 most probable emissions for each cluster at the end of training, as illustrated in Table 5. We compare the results from 3 experiments: the baseline, the **O**rder+**P**riming model, and the **O**rder+**U**nlocking+**P**riming model, corresponding to rows 1, 7, 8 of Table 3. For the baseline, the learned emissions are the same for each cluster, and they are the 10 most frequent tokens in the training dataset (see row 0 of Table 5). For rows 1-7 in Table 5, the morphemes listed are the intersection of the top 10 emissions for a cluster in the fully augmented model and in the Order+Priming model. The same clusters emerge in the experiment where intake is simulated.

0 - top 10 ranked	*you 's the what it that a to i do*
1 - prenominals	*the* ***a*** *your that this his some my her very*
2 - lexical verbs	*go do gon get* ***eat*** *say play look try come*
3 - post-verbal	***ing s ed*** *na n't it with you out*
4 - prep., part., aux.	***in*** *on to up for it get have*
5 - pronouns, part.	*you i he she we that not just*
6 - start pronouns	*what he it she who oh*
7 - copular, aux.	***'s*** *'re is was are do then*

Table 5: Words with highest emission probabilities for a cluster at the end of training. Row 0 is the baseline, and rows 1-7 are the intersection between the augmented model's emissions. Brown's morphemes are in *blue*, and morphs used for priming are in ***bold***.

The more cohesive clusters consist of higher-frequency tokens. Due to the specifics of Online EM and HMMs, the clusters produced by our models are of equal size, as all tokens can be emitted from all clusters. A model which can represent clusters of different sizes, coupled with priming, will likely lead to even more cohesive clustering – this is because functional categories are closed-class, i.e. they have a smaller and fairly constant membership in comparison to lexical categories. Although the clusters in row 1-7 in Table 5 leave a lot to be desired, they are a big improvement on row 0 from Table 5, where the baseline model has not learnt anything beyond the rank of the tokens.

Brown-compliant vs random orders

Using the highest-performing setting, the model with Ordering and Priming from row 7 in Table 3, we ran 4,000 experiments to explore how Brown-compliant ordering of the functional category phases compares to random, non-Brown compliant ordering, the final *LL*s for which are visualised in Figure 3. Both Brown-compliant and non-Brown compliant experiments use priming and ordering the examples correspondingly to the priming order; but the order of priming is what differs between the settings. Brown-compliant experiments are ones where the functional morphemes can be learned in any order within the MLU stage, but all the morphemes from a given MLU stage are learned before the child moves on to morphemes from subsequent MLU stages. Non-Brown-compliant orders are ones where the functional morphemes are shuffled without any regard for the MLU stages. The Brown-compliant orders achieve mean final $LL = -4.25$, compared to $LL = -4.64$ for the random ones. Using a non-parametric test (Mann & Whitney, 1947), we show that the difference is statistically significant with $p < 10^{-140}$ and a large effect size $d = 1.11$. This is the first time that Brown's attested order, which is much-cited in literature, is computationally demonstrated to be superior to non-Brown-compliant orders.

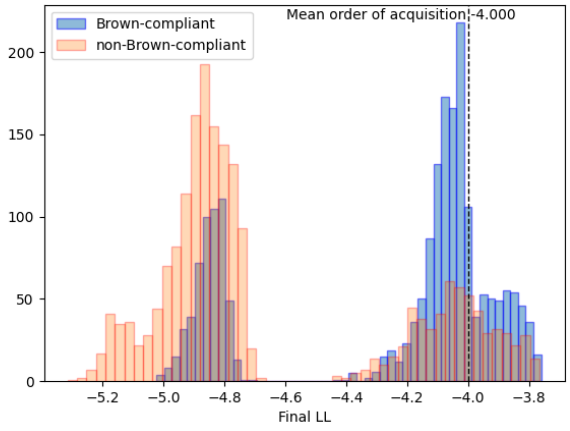

Figure 3: *LL* for 2,000 Brown-compliant orders and 2,000 non-Brown-compliant orders using the augmented model. The Brown-compliant orders yield significantly higher final *LL* ($p < 10^{-140}$).

The highest *LL* is achieved when either the uncontractible or contractible auxiliary phase is last (refer to Figure 2 for the computational operationalisation). The lowest-scoring orders for both groups are the ones where the irregular third person present phase is last. This phase includes only three verbs: *has, does, goes*. This might be of interest to linguists as it emphasises the necessity for the right environment, frequency, and context in order to be able to learn irregular forms.

Related Work

The works below provide enriching accounts of modelling FLA. However, none of them use data tokenised at our granularity and none explicitly focus on functional categories.

Imposing order is theoretically motivated, starting with Gold (1967)'s foundational work with subsequent computational operationalisations of aspects of FLA (Gibson & Wexler, 1994; Briscoe, 1997) also underlining the importance of order in understanding successful acquisition. Brown's order specifically is referenced in Buttery (2006) with respect to the types of the first produced sentences, but functional categories and morphemes are not mentioned.

Syntactic category acquisition is an existing research area, starting with Cartwright and Brent (1997). Parisien, Fazly, and Stevenson (2008) propose an incremental Bayesian model, but their focus is on lexical as opposed to functional categories. Chrupała and Alishahi (2010) present an entropy-based approach to lexical category acquisition, which results in more linguistically plausible category groupings than commonly used POS tagsets (in accordance with our motivation to produce a tagset specifically for FCI).

Grammar induction is the task of finding the latent structure of a natural language, its grammar, based on a set of raw sentences from the language. Prototype-driven grammar induction by Haghighi and Klein (2006) relies on a list of the target labels and a few words associated with each label, similarly to priming. A. Clark (2003) and Yatbaz, Sert, and Yuret (2012) also demonstrate how the incorporation of morphological features, based both on inflections and function words, improves POS induction. Character-based PCFG (Jin, Oh, & Schuler, 2021) utilises the information inside a word, but our work differs as we target the smallest linguistic unit, morphemes, instead of naively placing equal importance on all characters. Categorial grammar induction (C. Clark & Schuler, 2024) allows for the more transparent interaction of syntax and semantics. In Marcheva, Biberauer, and Sun (2025) we examine how morphemic tokenisation affects state-of-the art neural grammar induction. Automatic morphemic tokenisation (Smit, Virpioja, Grönroos, & Kurimo, 2014) may allow exploration of a wider range of morphemes, but is unnecessary for the scale of this pilot study.

Conclusion

We present the new computational task of functional category induction (FCI), supplied with a manually annotated dataset, to simulate an aspect of syntactic acquisition. Taking into account "less-is-more"-type cognitive biases, we identify three experiment control factors (order, gradual unlocking, and priming) which we incorporate in a computational operationalisation based on statistical learning. We perform experiments to explore the effect of the three factors on FCI, and, as hypothesised, find that the cognitively augmented models outperform the purely statistical Online EM baseline. Finally, for the first time in the literature we statistically confirm the superiority of Brown's order over random orders.

References

Ambridge, B., & Lieven, E. V. M. (2011, March). Child language acquisition: Contrasting theoretical approaches. In (p. 103 - 136). Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9780511975073.005 doi: 10.1017/cbo9780511975073

Baker, M. C. (2008). The macroparameter in a microparametric world. In *The limits of syntactic variation* (p. 351–373). John Benjamins Publishing Company. Retrieved from http://dx.doi.org/10.1075/la.132.16bak doi: 10.1075/la.132.16bak

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (p. 41–48). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1553374.1553380 doi: 10.1145/1553374.1553380

Biberauer, T. (2018, November). Less is more: On the tolerance principle as a manifestation of maximize minimal means. , *8*(6), 707–711. Retrieved from https://doi.org/10.1075/lab.18080.bib doi: 10.1075/lab.18080.bib

Biberauer, T. (2019a, October). Children always go beyond the input: The maximise minimal means perspective. *Theoretical Linguistics*, *45*(3–4), 211–224. Retrieved from http://dx.doi.org/10.1515/tl-2019-0013 doi: 10.1515/tl-2019-0013

Biberauer, T. (2019b, December). Factors 2 and 3: Towards a principled approach. *Catalan Journal of Linguistics*, 45. Retrieved from https://doi.org/10.5565/rev/catjl.219 doi: 10.5565/rev/catjl.219

Bosch, N., & Biberauer, T. (2024). Emergent syntactic categories and increasing granularity: evidence from a multilingual corpus study. *BUCLD 48 Proceedings*.

Brent, M., & Siskind, J. (2001, 10). The role of exposure to isolated words in early vocabulary. *Cognition*, *81*, B33-44. doi: 10.1016/S0010-0277(01)00122-6

Briscoe, T. (1997, July). Co-evolution of language and of the language acquisition device. In *35th annual meeting of the association for computational linguistics and 8th conference of the European chapter of the association for computational linguistics* (pp. 418–427). Madrid, Spain: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P97-1054 doi: 10.3115/976909.979671

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press. Retrieved from https://doi.org/10.4159/harvard.9780674732469 doi: 10.4159/harvard.9780674732469

Buttery, P. J. (2006, November). *Computational models for first language acquisition* (Tech. Rep. No. UCAM-CL-TR-675). University of Cambridge, Computer Laboratory. Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-675.pdf doi: 10.48456/tr-675

Cartwright, T. A., & Brent, M. R. (1997, May). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170. Retrieved from http://dx.doi.org/10.1016/S0010-0277(96)00793-7 doi: 10.1016/s0010-0277(96)00793-7

Chen, Y., & Teufel, S. (2021, November). Synthetic textual features for the large-scale detection of basic-level categories in English and Mandarin. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8294–8305). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.654 doi: 10.18653/v1/2021.emnlp-main.654

Chomsky, N. (2005, 01). Three Factors in Language Design. *Linguistic Inquiry*, *36*(1), 1-22. Retrieved from https://doi.org/10.1162/0024389052993655 doi: 10.1162/0024389052993655

Chrupała, G., & Alishahi, A. (2010, July). Online entropy-based model of lexical category acquisition. In M. Lapata & A. Sarkar (Eds.), *Proceedings of the fourteenth conference on computational natural language learning* (pp. 182–191). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W10-2922

Clark, A. (2003, April). Combining distributional and morphological information for part of speech induction. In *10th conference of the European chapter of the association for computational linguistics.* Budapest, Hungary: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E03-1009

Clark, C., & Schuler, W. (2024, May). Categorial grammar induction with stochastic category selection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 2893–2900). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.258/

Culbertson, J., Compostella, A., & Kirby, S. (2024, November). Language structure reflects biases in pattern learning across domains and modalities. *Quarterly Journal of Experimental Psychology*. Retrieved from http://dx.doi.org/10.1177/17470218241282404 doi: 10.1177/17470218241282404

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Demuth, K. (2007, December). The role of frequency in language acquisition. In *Frequency effects in language acquisition* (p. 383–388). DE GRUYTER MOU-

TON. Retrieved from http://dx.doi.org/10.1515/9783110977905.383 doi: 10.1515/9783110977905.383

Dye, C., Kedar, Y., & Lust, B. (2018, December). From lexical to functional categories: New foundations for the study of language development. *First Language*, *39*(1), 9–32. Retrieved from http://dx.doi.org/10.1177/0142723718809175 doi: 10.1177/0142723718809175

Elman, J. L. (1993, July). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*(1), 71–99. Retrieved from https://doi.org/10.1016/0010-0277(93)90058-4 doi: 10.1016/0010-0277(93)90058-4

Evers, A. E., & van Kampen, J. (2008). Parameter setting and input reduction. In *The limits of syntactic variation* (p. 483–515). John Benjamins Publishing Company. Retrieved from http://dx.doi.org/10.1075/la.132.22eve doi: 10.1075/la.132.22eve

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (p. 215–256). Cambridge University Press. doi: 10.1017/CBO9780511620669.010

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454. Retrieved 2023-10-19, from http://www.jstor.org/stable/4178869

Gold, E. M. (1967, May). Language identification in the limit. *Information and Control*, *10*(5), 447–474. Retrieved from https://doi.org/10.1016/s0019-9958(67)91165-5 doi: 10.1016/s0019-9958(67)91165-5

Guasti, M. T. (2017, February). Language acquisition. In (2nd ed., p. 1-27). Cambridge, MA: Bradford Books.

Haghighi, A., & Klein, D. (2006, July). Prototype-driven grammar induction. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 881–888). Sydney, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P06-1111 doi: 10.3115/1220175.1220286

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Paul H Brookes Publishing.

Haspelmath, M. (2020, May). The morph as a minimal linguistic form. *Morphology*, *30*(2), 117–134. Retrieved from http://dx.doi.org/10.1007/s11525-020-09355-5 doi: 10.1007/s11525-020-09355-5

Jin, L., Oh, B.-D., & Schuler, W. (2021, November). Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 4367–4378). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.371 doi: 10.18653/v1/2021.findings-emnlp.371

Johnson, M. (2007, June). Why doesn't EM find good HMM POS-taggers? In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 296–305). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D07-1031

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall. Retrieved from http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y

Liang, P., & Klein, D. (2009, June). Online EM for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the north American chapter of the association for computational linguistics* (pp. 611–619). Boulder, Colorado: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N09-1069

Lidz, J., & Gagliardi, A. (2015, January). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353. Retrieved from https://doi.org/10.1146/annurev-linguist-030514-125236 doi: 10.1146/annurev-linguist-030514-125236

Lieven, E. (2010, November). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*(11), 2546–2556. Retrieved from http://dx.doi.org/10.1016/j.lingua.2010.06.005 doi: 10.1016/j.lingua.2010.06.005

MacWhinney, B. (2014). *The childes project*. Psychology Press. Retrieved from http://dx.doi.org/10.4324/9781315805672 doi: 10.4324/9781315805672

Mann, H. B., & Whitney, D. R. (1947, March). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. Retrieved from http://dx.doi.org/10.1214/aoms/1177730491 doi: 10.1214/aoms/1177730491

Marcheva, M., Biberauer, T., & Sun, W. (2025, May). Profiling neural grammar induction on morphemically tokenised child-directed speech. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, J. Li, & B.-D. Oh (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 47–54). Albuquerque, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.cmcl-1.7/

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993, June). Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, *19*(2), 313–330.

Meilă, M. (2003). Comparing clusterings by the varia-

tion of information. In *Lecture notes in computer science* (p. 173–187). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-45167-9_14 doi: 10.1007/978-3-540-45167-9_14

Montani, I., Honnibal, M., Honnibal, M., Landeghem, S. V., Boyd, A., Peters, H., ... Patel, A. (2020, October). *explosion/spaCy: v3.0.0rc: Transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4091419 doi: 10.5281/zenodo.4091419

Newport, E. L. (1990, January). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28. Retrieved from https://doi.org/10.1207/s15516709cog1401_2 doi: 10.1207/s15516709cog1401_2

Parisien, C., Fazly, A., & Stevenson, S. (2008, August). An incremental Bayesian model for learning syntactic categories. In A. Clark & K. Toutanova (Eds.), *CoNLL 2008: Proceedings of the twelfth conference on computational natural language learning* (pp. 89–96). Manchester, England: Coling 2008 Organizing Committee. Retrieved from https://aclanthology.org/W08-2112

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and bayesian inference in language acquisition. In J. Lidz, W. Snyder, & J. Pater (Eds.), *Oxford handbook of developmental linguistics.* United Kingdom: Oxford University Press.

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*. Retrieved from https://ling.auf.net/lingbuzz/001493

Pearl, L., & Sprouse, J. (2021, March). The acquisition of linking theories: A tolerance and sufficiency principle approach to deriving UTAH and rUTAH. *Language Acquisition*, 1–32. Retrieved from https://doi.org/10.1080/10489223.2021.1888295 doi: 10.1080/10489223.2021.1888295

Petrov, S., Das, D., & McDonald, R. (2012, May). A universal part-of-speech tagset. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 2089–2096). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Phillips, L., & Pearl, L. (2015, February). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854. Retrieved from https://doi.org/10.1111/cogs.12217 doi: 10.1111/cogs.12217

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439. Retrieved from https://www.sciencedirect.com/science/article/pii/001002857690013X doi: https://doi.org/10.1016/0010-0285(76)90013-X

Saffran, J. R., & Kirkham, N. Z. (2018, January). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. Retrieved from https://doi.org/10.1146/annurev-psych-122216-011805 doi: 10.1146/annurev-psych-122216-011805

Saffran, J. R., & Thiessen, E. D. (2007). Domain-general learning capacities. In *Blackwell handbook of language development* (p. 68–86). Blackwell Publishing Ltd. Retrieved from http://dx.doi.org/10.1002/9780470757833.ch4 doi: 10.1002/9780470757833.ch4

Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. (2020, July). Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1970–1978). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.178 doi: 10.18653/v1/2020.acl-main.178

Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014, April). Morfessor 2.0: Toolkit for statistical morphological segmentation. In S. Wintner, M. Tadić, & B. Babych (Eds.), *Proceedings of the demonstrations at the 14th conference of the European chapter of the association for computational linguistics* (pp. 21–24). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E14-2006/ doi: 10.3115/v1/E14-2006

Snyder, W. (2021, July). A parametric approach to the acquisition of syntax. *Journal of Child Language*, *48*(5), 862–887. Retrieved from http://dx.doi.org/10.1017/S0305000921000465 doi: 10.1017/s0305000921000465

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, *21*(2), 121–137.

van Kampen, J. (2004). Learnability order in the french pronominal system. In B. K.-M. Reineke Bok-Bennema Bart Hollebrandse & P. Sleeman (Eds.), *Romance languages and linguistic theory 2002: Selected papers from 'going romance'* (p. 163). John Benjamins Publishing Company. Retrieved from http://dx.doi.org/10.1075/cilt.256.10kam doi: 10.1075/cilt.256.10kam

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Wexler, K. (1998, December). Very early parameter setting and the unique checking constraint: A new explana-

tion of the optional infinitive stage. *Lingua*, *106*(1–4), 23–79. Retrieved from http://dx.doi.org/10.1016/S0024-3841(98)00029-1 doi: 10.1016/s0024-3841(98)00029-1

Wijnen, F. (2000). Input, intake and sequence in syntactic development. *From sound to sentence - Studies on first language acquisition*, 163–186. Retrieved from https://www.researchgate.net/publication/46602182_Input_intake_and_sequence_in_syntactic_development

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, *abs/1910.03771*. Retrieved from http://arxiv.org/abs/1910.03771

Yang, C. D. (2004, October). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. Retrieved from https://doi.org/10.1016/j.tics.2004.08.006 doi: 10.1016/j.tics.2004.08.006

Yatbaz, M. A., Sert, E., & Yuret, D. (2012, July). Learning syntactic categories using paradigmatic representations of word context. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 940–951). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D12-1086