Exploring the utility of joint morphological and syntactic learning from child-directed speech

Stella Frank
sfrank@inf.ed.ac.uk

Frank Keller
keller@inf.ed.ac.uk

Sharon Goldwater
sgwater@inf.ed.ac.uk

ILCC, School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK

Abstract

Children learn various levels of linguistic structure concurrently, yet most existing models of language acquisition deal with only a single level of structure, implicitly assuming a sequential learning process. Developing models that learn multiple levels simultaneously can provide important insights into how these levels might interact synergistically during learning. Here, we present a model that jointly induces syntactic categories and morphological segmentations by combining two well-known models for the individual tasks. We test on child-directed utterances in English and Spanish and compare to single-task baselines. In the morphologically poorer language (English), the model improves morphological segmentation, while in the morphologically richer language (Spanish), it leads to better syntactic categorization. These results provide further evidence that joint learning is useful, but also suggest that the benefits may be different for typologically different languages.

1 Introduction

Models of language acquisition seek to infer linguistic structure from data with minimal amounts of prior knowledge, in order to discover which characteristics of the input data are useful for learning, and thus potentially utilised by human learners. Most previous work has focused on learning individual aspects of linguistic structure. However, children clearly learn multiple aspects in parallel, rather than sequentially, implying that models of language acquisition should also incorporate joint learning. Joint models investigate the interaction between different levels of linguistic structure during learning. These interactions are often (but not necessarily) synergistic, enabling better, more robust, learning by making use of cues from multiple sources. Recent models using joint learning to model language acquisition have spanned various domains including phonology, word segmentation, syntax and semantics (Feldman et al., 2009; Elsner et al., 2012; Doyle and Levy, 2013; Johnson, 2008; Kwiatkowski et al., 2012).

In this paper we examine the joint learning of syntactic categories and morphology, which are acquired by children at roughly the same age (Clark, 2003b), implying possible interactions in the learning process. Both morphology and word order depend on categorising words based on their morphosyntactic function. However, previous models of syntactic category learning have relied principally on surrounding context, i.e., word order constraints, whereas models of morphology use word-internal cues. Our joint model integrates both sources of information, allowing the model to flexibly weigh them according to their utility.

Languages differ in the richness of their morphology and strictness of word order. These characteristics appear to be (anti)correlated, with rich morphology co-occurring with free word order and vice versa (Blake, 2001; McFadden, 2003). The timecourse of acquisition is also influenced by language typology: learners of morphologically rich languages become productive in morphology earlier (Xanthos et al., 2011), suggesting that richer morphology may be more salient for learners than impoverished morphology. Sentence comprehension in children also shows cross-linguistic differences in the cues used to make sense of non-canonical sentence structure: learners of a morphologically rich language (Turkish) disregard word order in

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 30–41,
Seattle, Washington, USA, 18-21 October 2013.

favour of morphology, whereas learners of English favour word order (Slobin, 1982; MacWhinney et al., 1984). These interactions between morphology and word order suggest that a joint model will be better able to support the differences in cue strength (rich morphology versus strict word order), and thus be more language-general, than single-task models.

Both syntactic category and morphology induction have been the focus of much recent work. (See Hammarström and Borin (2011) for an overview of unsupervised morphology learning, likewise Christodoulopoulos et al. (2010) for a comparison of part of speech/syntactic category induction systems.) However, given the tightly coupled nature of these two tasks, there has been surprisingly little work in *joint* learning of morphology and syntactic categories. Systems for inducing syntactic categories often make use of morpheme-like features, such as word-final characters (Smith and Eisner, 2005; Haghighi and Klein, 2006; Berg-Kirkpatrick et al., 2010; Lee et al., 2010), or model words at the character-level (Clark, 2003a; Blunsom and Cohn, 2011), but do not include morphemes explicitly. Other systems (Dasgupta and Ng, 2007; Christodoulopoulos et al., 2011) use morphological segmentations learned by a separate morphology model as features in a pipeline approach.

Models of morphology induction generally operate over a lexicon, i.e. a list of word types, rather than token corpora (Goldsmith, 2006; Creutz and Lagus, 2007; Kurimo et al., 2010). These models find morphological categories on the basis of word-internal features, without taking syntactic context into account (which is of course not available in a lexicon).

Lee et al. (2011) and Sirts and Alumäe (2012) present models that infer morphological segmentations and syntactic categories jointly, although Lee et al. (2011) do not evaluate the inferred syntactic categories. Both make use of a word-type constraint which limits each word form to a single analysis (i.e., all instances of *ducks* are assigned to a single category and will have the same morpheme analysis, ignoring the gold standard distinction between a plural noun and third person singular verb). This can make inference more tractable, and often increases performance, but does not respect the ambiguity inherent in natural language, both over syntactic categories and morphological analyses. The degree of ambiguity is language dependent, so that even if a type-constraint is perhaps relatively unproblematic in English, it will pose problems in morphologically richer languages. Furthermore, these two models make use of an array of heuristics that may not allow them to be easily generalisable across languages and datasets (e.g., likelihood scaling (Sirts and Alumäe, 2012), sequential suffix matching (Lee et al., 2011)).

In this paper, we present a joint model composed of two well-known individual models. This allows us to cleanly investigate the effects of joint learning and its potential benefits over the single task models. The simplicity of our models also allows us to avoid modelling and inference heuristics.

Previous models have used adult-directed written texts, which differs significantly from the type of language available to child learners. We test our joint model on child-directed utterances in English (a morphologically poor language) and Spanish (with richer morphology)[1]. Our results indicate that our joint model is able to flexibly accommodate languages with differing levels of morphological richness. The joint model matches the performance of single task models on both tasks, demonstrating that the additional complexity is not a problem (i.e., it does not add noise). Moreover, the joint model improves performance significantly on the task corresponding to the language's weaker cue, indicating a transfer of information from the stronger cue. The fact that the nature of this improvement varies by language provides evidence that joint learning can effectively accommodate typological diversity.

2 Model

The task is to assign word tokens to part of speech categories and simultaneously segment the tokens into morphemes. We assume a relatively simple yet commonly used concatenative morphology which models a word as a stem plus (possibly null) suffix[2].

[1] There are languages with much richer morphology than Spanish, but none with a child-directed corpus suitably annotated for evaluation.

[2] Fullwood and O'Donnell (2013) recently presented a model of non-concatenative morphology that could be integrated into this model; however, it does not perform well on English (and presumably other mostly concatenative languages).

Since this is an unsupervised model, the inferred categories and morphemes lack meaningful labels, but ideally will correspond to gold standard categories and morphemes.

2.1 Word Order

We model a sequence of words as a Hidden Markov Model (HMM) with a non-parametric emission distribution. As usual, the latent states of the HMM represent syntactic categories. The tag sequence is generated by a trigram Dirichlet-multinomial distribution, where transition parameters τ are drawn from a symmetric Dirichlet distribution with the hyperparameter α_t. Each tag t_i in the sequence is then drawn from the transition distribution conditioned on the previous two tags:

$$\begin{aligned} \tau_{(t,t')} \sim & \quad \mathrm{Dir}(\alpha_t) \\ t_i = t | t_{i-1} = t', t_{i-2} = t'', \tau \sim & \quad \mathrm{Mult}(\tau_{(t',t'')}) \end{aligned}$$

This model is token-based, permitting different tokens of the same word type to have different syntactic categories. Most recent models have included a constraint forcing all tokens of a given type into the same category, which improves performance but often complicates inference. The Bayesian HMM's performance is therefore not state-of-the-art, but is comparable to other token-based models (Christodoulopoulos et al., 2010) and the model is easy to extend within the Bayesian framework, allowing us to compare multiple versions.

This part of the model is parametric, operating over a fixed number of tags T, and is identical to the formulation of tag transitions in the Bayesian HMM (Goldwater and Griffiths, 2007). However, we replace the BHMM's emission distribution with the morphologically-informed distributions described below. As in the BHMM, the emission distributions are conditioned on the tag, i.e., each tag has its own morphology.

2.2 Morphology

The morphology model introduced by Goldwater et al. (2006) generates morphological analyses for a set of tokens. These analyses consist of a tag plus a stem and suffix pair, which are concatenated to form the observed words. Both stem s and suffix f are generated from Dirichlet-multinomials conditioned on the tag t:

$$\begin{aligned} \kappa \sim & \quad \mathrm{Dir}(\alpha_\kappa) \\ t|\kappa \sim & \quad \mathrm{Mult}(\kappa) \\ \sigma \sim & \quad \mathrm{Dir}(\alpha_s) \\ s|t,\sigma \sim & \quad \mathrm{Mult}(\sigma_t) \\ \phi \sim & \quad \mathrm{Dir}(\alpha_f) \\ f|t,\phi \sim & \quad \mathrm{Mult}(\phi_t) \end{aligned}$$

The αs are hyperparameters governing the Dirichlet distributions from which the multinomials κ, σ, ϕ are drawn. In turn, t, s, and f are drawn from these multinomials.

The probability of a word under this model is the sum of the probabilities of all possible analyses $l = (t, s, f)$:

$$P_0(w) = \sum_l P_0(l) = \sum_{\substack{t,s,f \text{ s.t.} \\ s \oplus f = w}} P(s|t)P(f|t)P(t) \quad (1)$$

where $s \oplus f = w$ denotes that the concatenation of stem and suffix results in the word w.

On its own, this distribution over morphological analyses makes independence assumptions that are too strong: most word tokens of a word type have the same analysis, but P_0 will re-generate that analysis for every token. To resolve this problem, a Pitman-Yor process (PYP) is placed over the generating distribution above. The Pitman-Yor process has been found to be useful for representing the power-law distributions common in natural language (Teh, 2006; Goldwater and Griffiths, 2007; Blunsom and Cohn, 2011).

The distribution of draws from a Pitman-Yor process (which, in our case, determines the distribution of word tokens with each morphological analysis) is commonly described using the metaphor of a Chinese restaurant. A series of customers (tokens $\boldsymbol{z} = z_1 \ldots z_N$) enter a restaurant with an infinite number of initially empty tables. Upon entering, each customer is seated at a table k with probability

$$p(z_i = k | z_1 \ldots z_{i-1}, a, b) = \quad (2)$$
$$\begin{cases} \frac{n_k - a}{i-1+b} & \text{if } 1 \leq k \leq K \\ \frac{Ka+b}{i-1+b} & \text{if } k = K+1 \end{cases}$$

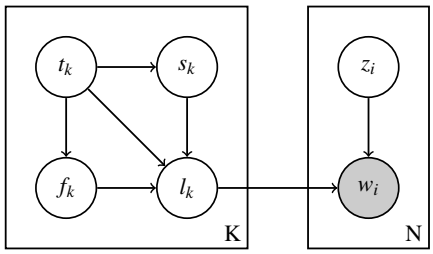

Figure 1: Plate diagram depicting the morphology model (adapted from Goldwater et al. (2006)). Hyperparameters have been omitted for clarity. The left-hand plate depicts the base distribution P_0; note that the morphological analyses l_k are generated deterministically as (t_k, s_k, f_k). The observed words w_i are also deterministic given $z_i = k$ and l_k, since $w_i = s_k \oplus f_k$.

where n_k is the number of customers already sitting at table k, K is the total number of tables occupied by the $i-1$ previous customers, and $0 \leq a < 1$ and $b \geq 0$ are hyperparameters of the process. The probability of being seated at a table increases with the number of customers already seated at that table, creating a 'rich-get-richer' power-law distribution of tokens to tables; a and b control the amount of reuse of existing tables, with smaller values leading to more reuse.

Crucially, each table serves a dish generated by the base distribution P_0—i.e., the dish is a morphological analysis $l_k = (t, s, f)$—and all the customers seated at the same table share the same dish, which is generated only once (at the point when that table is first occupied). The model can thus reuse the analysis for a particular word and avoid regenerating the same analysis multiple times. Note that multiple tables may have identical analyses, $l_k = l_{k'}$. Figure 1 illustrates how the full PYP morphology model generates the observed sequence of word tokens.

2.3 Combined Model

The full model (Figure 2) combines the latent tag sequence with the morphology model. Tag tokens are generated conditioned on local context, not the base distribution, as in the morphology model. Instead of a single PYP generating morphological analyses for all tokens, as in the Goldwater et al. (2006) model, we have a separate PYP for each tag type, i.e., each tag has its own restaurant with its own customers (the tokens labeled with that tag) and its own morphological analyses. The distribution of customers

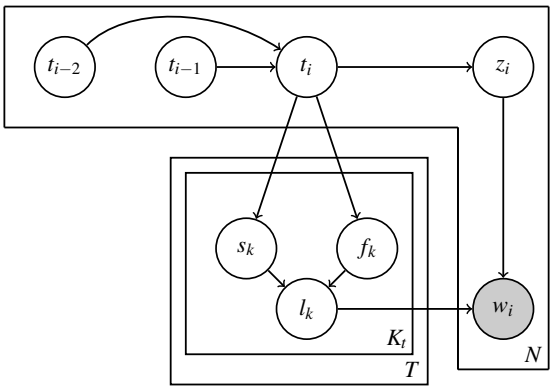

Figure 2: Plate diagram depicting the joint model. Hyperparameters have been omitted for clarity. The L-shaped plate contains the tokens, while the square plates contain the morphological analyses. The t are latent tags, z_i is an assignment to a morphological analysis $l_k = (s_k, f_k)$, and w_i is the observed word. T is the number of distinct tags, and K_t the number of tables used by tag type t.

in each of the tag-specific restaurants is still determined by Equation 2, except that all of the counts and indices are with respect to only the tokens and tables assigned to that tag.

Each tag-specific PYP (restaurant) also has a separate base distribution, $P_0^{(t)}$, resulting in distinct distributions over stems and suffixes for each tag. The analyses generated by the base distributions consist of (stem, suffix) pairs; the tag is given by the identity of the generating PYP.

$$P_0^{(t)}(w) = \sum_l P_0^{(t)}(l = (s, f)) = \sum_{\substack{s,f \text{ s.t.} \\ s \oplus f = w}} P(s|t)P(f|t) \tag{3}$$

The full joint posterior distribution of a sequence of words, tags, and morpheme analyses is shown in Figure 3. Note that all tag-specific morphology models share the same Pitman-Yor parameters a and b.

3 Inference

We use Gibbs sampling for inference over the three sets of discrete variables: tags $\boldsymbol{t}$, their assignments to morphological analyses (tables) $\boldsymbol{z}$, and the analyses themselves $\boldsymbol{l}$.

Each iteration of the sampler has two stages: First the morphological analyses $\boldsymbol{l}$ are sampled, and then each token samples a new tag and a new assignment to an analysis/table. Because the table assignments

$$P(\boldsymbol{t},\boldsymbol{l},\boldsymbol{z}|\alpha_t,a,b,\alpha_s,\alpha_f) = P(\boldsymbol{t}|\alpha_t)P(\boldsymbol{l}|\boldsymbol{t},\alpha_s,\alpha_f)P(\boldsymbol{z}|a,b) \tag{4}$$

$$P(\boldsymbol{t}|\alpha_t) = \prod_{i=2}^{N} P(t_i|t_{i-1},t_{i-2},\boldsymbol{t}_{1\ldots i-1},\alpha_t) = \prod_{t,t'=1}^{T} \frac{\Gamma(T\alpha_t)}{\Gamma(n_{tt'}+T\alpha_t)} \prod_{t''=1}^{T} \frac{\Gamma(n_{tt't''}+\alpha_t)}{\Gamma(\alpha_t)} \tag{5}$$

$$P(\boldsymbol{l}|\boldsymbol{t},\alpha_s,\alpha_f) = \prod_{t=1}^{T}\prod_{k=1}^{K_t} P_t(l_k=(s,f)|\boldsymbol{l}_{1\ldots k-1},\alpha_s,\alpha_f) \tag{6}$$

$$= \prod_{t=1}^{T} \frac{\Gamma(S\alpha_s)}{\Gamma(m_t+S\alpha_s)} \frac{\Gamma(F\alpha_f)}{\Gamma(m_t+F\alpha_f)} \prod_{s=1}^{S} \frac{\Gamma(m_{ts}+\alpha_s)}{\Gamma(\alpha_s)} \prod_{f=1}^{F} \frac{\Gamma(m_{tf}+\alpha_f)}{\Gamma(\alpha_f)} \tag{7}$$

$$P(\boldsymbol{z}|a,b) = \prod_{t=1}^{T}\prod_{i=1}^{N_t} P(z_i|t,\boldsymbol{z}_{1\ldots i-1},a,b) \tag{8}$$

$$= \prod_{t=1}^{T} \frac{\Gamma(1+b)}{\Gamma(n_t+b)} \prod_{k=1}^{K_t} (ka+b) \frac{\Gamma(n_k-a)}{\Gamma(1-a)} \tag{9}$$

Figure 3: The posterior distribution of our joint model. Because the sequence of words $\boldsymbol{w}$ is deterministic given analyses $\boldsymbol{l}$ and assignments to analyses (tables) $\boldsymbol{z}$, the joint posterior over all variables $P(\boldsymbol{w},\boldsymbol{t},\boldsymbol{l},\boldsymbol{z}|\alpha_t,a,b,\alpha_s,\alpha_f)$ is equal to $P(\boldsymbol{t},\boldsymbol{l},\boldsymbol{z}|\alpha_t,a,b,\alpha_s,\alpha_f)$ when $l_{z_i}=w_i$ for all i, and 0 otherwise. We give equations for the non-zero case. ns refer to token counts, ms to table counts. We add two dummy tokens at the start, end, and between sentences to pad the context history.

are conditioned on tags (i.e., a token must be assigned to a table in the correct PYP restaurant) resampling the tag requires immediate resampling of the table assignment as well.

3.1 Initialization

The tags are initialized uniformly at random. For each token, a segmentation point is chosen uniformly at random (we disallow segmentations with a null stem). If this segmentation is new within the PYP associated with that token's tag, a new table is created for the token in that PYP. If it matches an existing analysis, z_i is sampled from the existing tables $\boldsymbol{k}$ plus a possible new table k'.

3.2 Morphological Analyses

Each l_k represents the morphological analysis for the set of tokens assigned to table k. Resampling the segmentation point (stem and suffix identity) of the analysis changes the segmentation of all of the word tokens assigned to that analysis. Note that the tag is not included in l_k in the combined model, because the tag identity is dependent on the local contexts of all the tokens seated at the table.

Analyses are sampled from a product of Dirichlet-multinomial posteriors as follows:

$$p(l_k=(s,f)|t,\boldsymbol{l}^{\backslash k}) = \frac{m_s^{\backslash k}+\alpha_s}{m^{\backslash k}+S\alpha_s} \frac{m_f^{\backslash k}+\alpha_f}{m^{\backslash k}+F\alpha_f} \tag{10}$$

where m_s and m_f are the number of analyses for this tag that share a stem or suffix with l_k, and m is the total number of analyses for this tag. S and F are the total number of stems and suffixes in the model. $\boldsymbol{l}^{\backslash k}$ indicates that the current analysis l_k has been removed from the distribution and the appropriate counts, to create the correct conditioning distribution for the Gibbs sampler.

3.3 Tags

Tags are sampled from the product of posteriors of the transition and emission distributions. The transition distribution is a standard Dirichlet-multinomial posterior. Calculating the emission distribution probability, i.e. the marginal probability of the word given the tag, involves summing over the probability of all the existing tables in the given PYP that emit the correct word, plus the probability of a new table being created, which also includes the probability of a new analysis from $P_0^{(t)}$.

More precisely, tags are sampled from the following distribution:

$$
\begin{aligned}
& p(t_i = t | w_i = w, \boldsymbol{t}^{\setminus i}, \boldsymbol{z}^{\setminus i}, \boldsymbol{l}, \alpha_t, a, b) \qquad (11) \\
& \propto p(t_i = t | t_{i-1}, t_{i-2}, \boldsymbol{t}^{\setminus i}, \alpha_t) \times p(w | t, \boldsymbol{z}^{\setminus i}, \boldsymbol{l}) \\
& = p(t_i = t | t_{i-1}, t_{i-2}, \boldsymbol{t}^{\setminus i}, \alpha_t) \\
& \times \Big(\sum_{k \text{ s.t. } l_k = w} p(z_i = k | t, w, \boldsymbol{z}^{\setminus i}) + p(z_i = k_{\text{new}} | t, w, \boldsymbol{z}^{\setminus i}) \Big) \\
& = \frac{n_{t_{i-2} t_{i-1} t} + \alpha_t}{n_{t_{i-2} t_{i-1}} + T \alpha_t} \\
& \times \Big(\sum_{k \text{ s.t. } l_k = w} \frac{n_k - a}{n_t + b} + \frac{K_t a + b}{n_t + b} P_0^{(t)}(w) \Big)
\end{aligned}
$$

where $l_k = w$ matches tables compatible with w, i.e., the concatenation of stem and suffix form the word, $s_{l_k} \oplus f_{l_k} = w$. n_k is the number of words assigned to the table k and K_t is the total number of tables in the PYP for tag t. Note that all counts are obtained after the removal of the current t_i and z_i, i.e., from $\boldsymbol{t}^{\setminus i}$ and $\boldsymbol{z}^{\setminus i}$.

3.4 Table Assignments

Once a new tag has been sampled for a token, the table assignment must be resampled conditioned on the new tag. The assignment z_i is drawn over all compatible tables in the tag's PYP (that is, where $l_k = w$), plus a possible new table:

$$
p(z_i = k | t_i = t, \boldsymbol{w}, \boldsymbol{z}^{\setminus i}, a, b) \propto \qquad (12)
$$
$$
\begin{cases} \frac{n_k - a}{n_t + b} & \text{if } 1 \leq k \leq K_t \\ \frac{K_t a + b}{n_t + b} \, P_0^{(t)}(w) & \text{if } k = K_t + 1 \end{cases}
$$

$P_0^{(t)}$ is calculated by summing over the probability of all possible segmentations for a new analysis for word w_i, using Equation 3. If a new table is drawn ($k > K_t$) then we also sample a new analysis for that table from $P_0^{(t)}$.

4 Preliminary Experiments

An important argument for joint learning is that it affords increased flexibility and robustness across a wider range of input data. A model that relies on word order cannot learn syntactic categories from a morphologically complex language with free word order; likewise a model attempting to categorise words using morphology alone will fail on a language without morphology. An effective joint model will be able to make use of the different cues in both language types in a flexible way.

Language A
abdc fefh pomo rtut usst
cdcc bcba gghh npop npoo
cdca aaaa fefh hfeg pnon
Language B
noom.no usrs.st bbdb.ac cbab.cc cdaa.cc
rttt.uu cbab.aa mnom.oo ccda.bc onmm.om
rruu.ts npop.mm gehg.fh trrt.uu tssu.uu

Table 1: Example sentences in the synthetic languages. Words in Category 1 are made of characters `a-d`, Category 2 `e-h`, Category 3 `m-p`, Category 4 `r-u`. Suffixes in Language B are separated with periods (.) for illustrative purposes only.

In order to test the proposed model, we run two experiments on synthetic languages, which simulate languages in which either word order or morphology is the sole cue. Most natural languages fall between these extremes, but these experiments show that our model can capture the full spectrum.

Language A is a strict word order language lacking morphology. It has a vocabulary of 200 word types, split into four different categories. The 50 word types in each category are created by combining four letters, with replacement, into four-letter words, with a different set of letters used in each category[3]. Words within a category may thus share beginning or ending characters, which could be posited as stems or suffixes by the model, but since only 50 of 256 possible strings are used, there will be no strong evidence for consistent stem and suffixes (i.e. stems appearing with multiple suffixes and vice versa). Each sentence in Language A consists of five words in one of twenty possible category sequences. In these sequences, each category is either followed by itself or the next category (i.e. [2,2,2,3,4] is valid but [2,4,3,1,4] is not). Word order is thus strongly constrained by category membership.

Language B has free word order, with category membership signalled by suffixes. Words are cre-

[3] We achieved the same results with a language using the same four characters in all categories, but using different characters makes the categories human-readable. The model does not have a orthographic/phonological component and so will not recognise the within-category similarity, other than possibly positing spurious stems or suffixes.

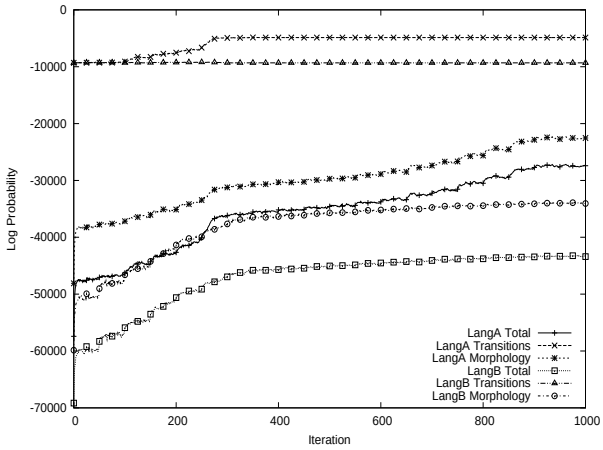

Figure 4: Log probability of the sampler state over 1000 iterations on Languages A and B.

ated by the concatenation of a stem and a suffix, where the stems are the same as the words in language A (50 stems in each of four categories). One of six category-specific suffixes is appended to each stem, resulting in 300 word types per category. Each suffix is two letters long, created by combining three possible letters (the same letters used to create the stems), thus making mis-segmentation possible (for instance, up to three of the suffixes could have the same final letter). Sentences are again five words long, but the sequence of categories is drawn at random, resulting in uniformly random word order. See Table 1 for example sentences in both languages.

We create a 5000 word corpus for each language, and run our model on these corpora. Hyperparameters are set to the same values in both languages[4].

We run the sampler on each dataset for 1000 iterations with simulated annealing. In both cases, the correct solution is found by iteration 500. Figure 4 shows that the morphology component continues to increase the log probability by increasing the number of tokens seated at a table. Note that the correct solution in Language A involves learning a very peaked transition distribution as well as an even more extreme distribution over suffixes (where only the null suffix has high probability), whereas the same distributions in Language B are much flatter. The fact that the same hyperparameter setting is able to correctly identify the two language extremes indicates that the model is robust to hyperparameter values.

These experiments demonstrate that our joint model is able to learn correctly even when only either morphology or word order is informative in a language. We now turn to acquisition data from natural languages in which both morphology and word order are useful cues but to varying degrees.

5 CDS Experiments

5.1 Data

We use two corpora, Eve (Brown, 1973) and Ornat (Ornat, 1994), from the CHILDES database (MacWhinney, 2000). These corpora consist of the child-directed utterances heard by two children, the former learning English and the latter Spanish. These have been annotated for part of speech categories and morphemes.

The CHILDES corpora are tagged with a very rich set of part of speech tags (74 tags), which we collapse to a smaller set of tags[5]. The Eve corpus has 61224 tokens and is thus larger than the Spanish corpus, which has 40497 tokens. However, the English corpus has only 17 gold suffix types, while Spanish has 83. The increased richness of Spanish morphology also has an effect on the number of word types in the corpus: the Spanish dataset has 3046 word types, whereas the larger English dataset has only 1957.

Morphology is annotated using a stem-affix encoding which does not directly correspond to our segmentation-based model. The word *running* is annotated as *run*-`ING`, *jumping* as *jump*-`ING`; the annotation is thus agnostic about ortho-morphemic segmentation (i.e., whether to segment as *run.ning* or *runn.ing*), whereas the model is forced to choose a segmentation point. Syncretic suffixes (sharing an identical surface form) are disambiguated: *sings* is annotated as *sing*-`3S`, *plums* as *plum*-`PL`. Conversely, the annotation scheme merges allomorphs into a single suffix: infinitive verbs in Spanish, for instance, are encoded as ending with `-INF`, corresponding to *-ar*, *-er*, and *-ir* surface forms.

[4]The PYP parameters are set to $a = 0.1, b = 1.0$ and the HMM transition parameter $\alpha_t = 1.0$; the parameters in the base distribution are $\alpha_s, \alpha_f = 0.001, \alpha_k = 0.5$.

[5]These are 13 for English (`ADJ`, `ADV`, `AUX`, `CONJ`, `DET`, `INF`, `NOUN`, `NEG`, `OTH`, `PART`, `PREP`, `PRO`, `VERB`) and 10 for Spanish, since the gold standard does not distinguish `AUX`, `PART` or `INF`.

We ignore irregular/non-affixing forms annotated with *&* (e.g. *was*, annotated as *be&PAST*) and use only hyphen-separated suffixes to evaluate. Where multiple suffixes are concatenated together (e.g., *dog*-DIM-PL) we treat this as a single suffix (-DIM-PL) for evaluation purposes.

In Spanish, many words are annotated as having a suffix of effectively zero length, e.g. the imperative *gusta* is annotated as *gusta*-2S&IMP. We replace these suffixes (where the stem is equal to the word) with a null suffix, excluding them from evaluation, as they are impossible for a segmentation-based model to find.

5.2 Evaluation

Tags are evaluated using VM (Rosenberg and Hirschberg, 2007), as has become standard for this task (Christodoulopoulos et al., 2010). VM is a measure of the normalised cross-entropy between gold and proposed clusters; it ranges between 0 and 100, with higher scores being better.

We also use VM to evaluate the morphological segmentation: all tokens with a common suffix are clustered together, and these clusters are compared against the gold suffix clusters[6]. Using a clustering metric avoids the need to evaluate against a gold segmentation point (which the annotation lacks). Tag membership is added to the non-null model suffixes, so that a final *-s* suffix found in tag 2 is distinguished from the same suffix found in tag 8 (creating suffixes -s-T8 and -s-T2), analogous to the gold annotation distinction between syncretic morphemes -PL and -3S.

Note that ceiling performance of our model on Suffix VM will be below 100, since our model cannot cluster allomorphs, which are represented by a single abstract morpheme in the gold standard.

5.3 Baselines

We test the full model, MORTAG, against a number of variations to investigate the advantages of jointly modelling the two tasks.

Two variants remove the transition distributions, and thus local syntactic context, from the model. MORTAGNOTRANS is the full model without transitions between tag tokens; morphology PYP draws remain conditioned on token tags. We add a Dirichlet prior over tags ($\alpha_t = 0.1$) to encourage tag sparsity (analogous to the transition distribution in the full model). MORCLUSTERS is the original model of Goldwater et al. (2006), in which tags (called clusters in the original) are drawn by P_0.

MORTAGNOSEG is a variant in which the only available suffix is the null suffix; thus segmentations are trivial and only tags are inferred. This model is approximately equivalent to a simple Bayesian HMM but with the addition of PYPs within the emission distribution. We also evaluate against tags found by the BHMM, with a Dirichlet-multinomial emission distribution and no morphology.

MORTAGTRUETAGS is the full model but with all tags fixed to their gold values. This model gives us oracle-type results for morphology. (Due to the annotation scheme used in CHILDES, oracle morphological segmentations are unavailable, so we were unable to test a model with gold morphology and inferred tags.)

5.4 Experimental Procedure

Hyperparameter values for the Pitman-Yor process were found using grid search on a development set (Section 10 of Eve and Section 8 of Ornat; these sections are removed from the dataset we report results on). We use the values which give the best Suffix VM performance on the development data; however we stress that the development results did not vary greatly over a wide range of hyperparameter values, and only deteriorated significantly at extreme values of a.

There are a number of other hyperparameters in the model which we set to fixed values. The transition hyperparameter α_t is set to 0.1 in all models. We set the hyperparameters for the stem and suffix distributions in the morphology base distribution P_0 to 0.001 for both α_s and α_f; α_k over tags in the MORCLUSTERS model is set to 0.5. The number of possible stems and suffixes is given by the dataset: in the Eve dataset there are 5339 candidate stems and 6617 candidate suffixes; in the Ornat dataset these numbers are 8649 and 6598, respectively. The number of tags available to the model is set to the number of gold tags in the data.

[6] We also evaluated stem morpheme clusters and found near-ceiling performance due to the high number of null-suffix words in both corpora.

	Tag VM	Suffix VM
MORTAG	59.1(1.9)	41.9(10.0)
MORCLUSTERS	22.4(1.0)*	28.0(11.9)*
MORTAGNOTRANS	19.3(1.2)*	24.4(5.2)*
MORTAGNOSEG	59.4(1.7)	–
BHMM	56.2(2.3)*	–
MORTAGTRUETAGS	–	42.5(5.2)

Table 2: English Eve corpus results. Standard deviations are in parentheses; $*$ denotes a significant difference from the MORTAG model.

	Tag VM	Suffix VM
MORTAG	43.4(2.6)	41.4(2.5)
MORCLUSTERS	20.3(2.5)*	46.5(3.2)
MORTAGNOTRANS	14.4(1.7)*	36.4(2.0)*
MORTAGNOSEG	39.6(3.7)*	–
BHMM	36.4(0.7)*	–
MORTAGTRUETAGS	–	59.8(0.4)*

Table 3: Spanish Ornat corpus results. Standard deviations are in parentheses; $*$ denotes a significant difference from the MORTAG model.

Sampling is run for 5000 iterations with annealing. Inspection of the posterior log-likelihood indicates that the models converge after about 1000 iterations. We run inference over all models ten times and report the average performance. Significance is reported using the non-parametric Wilcoxon rank-sum test with a significance level of $\rho < 0.05$.

5.5 Results: English

Results on the English Eve corpus are shown in Table 2. We use PYP parameters $a = 0.3$ and $b = 10$, though we found similar performance over a wide range of values of a and b. Our results show a clear improvement in the morphological segmentations found by the joint model and stable tagging performance across all models with context information.

The syntactic clusters found by models using only morphological patterns, MORTAGNOTRANS and MORCLUSTERS, are clearly inferior and lead to low Tag VM results. The models with local syntactic context all perform approximately equally well in terms of finding tags. We find no improvement on tagging performance in English when adding morphology, compared to the MORTAGNOSEG baseline in which words are not segmented. However, we do see a small but significant improvement over the BHMM for both of these models, due to the replacement of the multinomial emission distribution in the BHMM with the PYP.

Morphological segmentations, as measured by Suffix VM, clearly improve with the addition of local contexts (and the ensuing better tags): the full model outperforms the baselines without syntactic contexts. On this dataset, the joint MORTAG model even matches the performance of the model using oracle tags. The standard deviation over Suffix VM scores is quite large for MORTAG and MORCLUSTERS; this is due to frequent words having two high probability segmentations (most notably *is*, which in some runs was segmented as *i.s*).

5.6 Results: Spanish

For the Spanish Ornat corpus, we found slightly different optimal PYP hyperparameters and set $a = 0.1$ and $b = 0.1$. Results are shown in Table 3.

The Spanish results pattern in the opposite way as English. Here we see a statistically significant improvement in tagging performance of the full joint model over both models without morphology (MORTAGNOSEG and BHMM). Models without context information again find much worse tags, mainly because (as in English) function words are not identifiable by suffixes.

However, the full model does not find better morphological segmentations than the MORCLUSTERS model, despite better tags (the two models' Suffix VM scores are not statistically significantly different). We also see that the difference between the segmentations found by the model using gold tags and estimated tags is quite large. This is due to the oracle model finding the rarer suffixes which were not distinguished by the models with noisier tags. This demonstrates the importance of syntactic categorisation for the morpheme induction task, and suggests that a more sophisticated tagging model (with better performance) may yet improve morpheme segmentation performance in Spanish.

6 Conclusion

We have presented a model of joint syntactic category and morphology induction. Operating within a generative Bayesian framework means that combining single-task components is straightforward and well-founded. Our model is token-based, allowing for syntactic and morphemic ambiguity.

To our knowledge, this is the first joint model to be tested on child-directed speech data, which is less complex than the newswire corpora used by previous joint models. Child-directed speech may be simple enough for joint learning not to be necessary: our results indicate the contrary, namely that joint learning is indeed helpful when learning from realistic acquisition data.

We tested this model on two languages with different morphological characteristics. On English, a language with relatively little morphology, especially in child directed speech, we found that better categorisation of words yielded much better morphology in terms of suffixes learned. Conversely, in Spanish we saw less difference on the morphology task between models with categories inferred solely from morphemic patterns and models that also used local syntactic context for categorisation. However, in Spanish we saw an improvement in the tagging task when morphology information was included.

This suggests that English and Spanish make different word-order and morphology trade-offs. In English, local context provides at least as much information as morphology in terms of determining the correct syntactic category, but knowing a good estimate of the correct syntactic category is useful for determining a word's morphology. In Spanish, a word's morphology can more easily be determined simply by looking at frequent suffixes within a purely morphological system. On the other hand, word order is freer, making local syntactic context unreliable, so taking morphological information into account can improve tagging. These differences between languages demonstrate the benefits of joint learning, which enables the learner to more flexibly utilise the information available in the input data.

References

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, and Dan Klein. Painless unsupervised learning with features. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, 2010.

Barry J. Blake. *Case*. Cambridge University Press, 2001.

Phil Blunsom and Trevor Cohn. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.

Roger Brown. *A first language: The early stages*. Harvard University Press, Cambridge, MA, 1973.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. A Bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

Alexander Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th annual Meeting of the European Association for Computational Linguistics (EACL)*, 2003a.

Eve V. Clark. *First Language Acquisition*. Cambridge University Press, 2003b.

Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):1–34, 2007.

Sajib Dasgupta and Vincent Ng. Unsupervised part-of-speech acquisition for resource-scarce languages. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

Gabriel Doyle and Roger Levy. Combining multiple information types in Bayesian word segmentation. In *Proceedings of NAACL-HLT 2013*, pages 117–126, 2013.

Micha Elsner, Sharon Goldwater, and Jacob Eisenstein. Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

Naomi Feldman, Thomas Griffiths, and James Morgan. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci)*, 2009.

Michelle A. Fullwood and Timothy J. O'Donnell. Learning non-concatenative morphology. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2013.

John Goldsmith. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4):353–371, December 2006.

Sharon Goldwater and Thomas L. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18*, 2006.

Aria Haghighi and Dan Klein. Prototype-driven grammar induction. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

Harald Hammarström and Lars Borin. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350, 2011.

Mark Johnson. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

Mikko Kurimo, Sami Virpioja, and Ville T. Turunen. Proceedings of the MorphoChallenge 2010 workshop. Technical Report TKK-ICS-R37, Aalto University School of Science and Technology, Espoo, Finland, 2010.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettelmoyer, and Mark Steedman. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. Simple type-level unsupervised POS tagging. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. Modeling syntactic context improves morphological segmentation. In *Proceedings of Fifteenth Conference on Computational Natural Language Learning*, 2011.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Brian MacWhinney, Elizabeth Bates, and Reinhold Kliegl. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23:127–150, 1984.

Thomas McFadden. On morphological case and word-order freedom. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 29, pages 295–306, 2003.

S. Lopez Ornat. *La adquisicion de la lengua espagnola*. Siglo XXI, Madrid, 1994.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

Kairit Sirts and Tanel Alumäe. A hierarchical Dirichlet process model for joint part-of-speech and morphology induction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2012.

Dan Slobin. Universal and particular in the acquisition of language. In Eric Wanner and Lila R. Gleitman, editors, *Language acquisition: the state*

of the art, pages 128–170. Cambridge University Press, 1982.

Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

Aris Xanthos, Sabine Laaha, Steven Gillis, Ursula Stephany, Ayhan Aksu-Koç, Anastasia Christofidou, Natalia Gagarina, Gordana Hrzica, F. Nihan Ketrez, Marianne Kilani-Schoch, Katharina Korecky-Kröll, Melita Kovačević, Klaus Laalo, Marijan Palmović, Barbara Pfeiler, Maria D. Voeikova, and Wolfgang U. Dressler. On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 31(4):461–479, 2011.