# Adding sentence types to a model of syntactic category acquisition

Stella Frank

Institute for Language

Cognition and Computation, School of Informatics

University of Edinburgh

Sharon Goldwater

Institute for Language

Cognition and Computation, School of Informatics

University of Edinburgh

Frank Keller

Institute for Language

Cognition and Computation, School of Informatics

University of Edinburgh

Keywords: Language acquisition, Syntactic categories, Bayesian modeling, Computational modeling, Sentence prosody

## Author Note

Corresponding Author: Stella Frank

Email: sfrank@inf.ed.ac.uk

Telephone: +44 131 650 4418

## Abstract

The acquisition of syntactic categories is a crucial step in the process of acquiring syntax. At this stage, before a full grammar is available, only surface cues are available to the learner. Previous computational models have demonstrated that local contexts are informative for syntactic categorization. However, local contexts are affected by sentence-level structure. In this paper, we add sentence type as an observed feature to a model of syntactic category acquisition, based on experimental evidence showing that pre-syntactic children are able to distinguish sentence type using prosody and other cues. The model, a Bayesian Hidden Markov Model, allows for adding sentence type in a few different ways; we find that sentence type can aid syntactic category acquisition if it is used to characterize the differences in word order between sentence types. In these models, knowledge of sentence type permits similar gains to those found by extending the local context.

# Adding sentence types to a model of syntactic category acquisition

## Introduction

An essential early step in syntax acquisition is learning to group words into categories such as nouns or verbs according to their syntactic functions. Like other aspects of syntax acquisition, this task is unsupervised: a child is not told that a given utterance consists of, say, a determiner followed by a noun. Instead categories must be formed on the basis of implicit information about the syntactic similarity of words. Characterizing the nature and amount of the implicit information available to children as they acquire language is essential for any theory of language acquisition. Computational modeling plays an important role here, by demonstrating that there is (or is not) sufficient information in a learner's input for an idealized learner to reach certain conclusions about the structure of language, such as the underlying syntactic categories of words.

Syntactic categories are defined as a set of words that can fill the same (or highly similar) grammatical roles, i.e., that are syntactically similar. A key aspect of syntactic similarity is the distributional similarity between words, i.e., words that appear in the same surface contexts (similar distributions) tend to belong to the same syntactic categories. Distributional similarity is clearly weaker than full syntactic similarity, but it is used to avoid the circular problem of needing a grammar defined over syntactic categories to decide the syntactic category of a given word (although this process can be used once a fairly comprehensive grammar is in place to bootstrap individual unknown words).

The distributional hypothesis (Maratsos & Chalkley, 1980) posited that children use distributional information to bootstrap grammar. Empirical studies have provided evidence for this view, demonstrating that infants are sensitive to distributional context before they are likely to have an adult-like grammar (before 18 months): infants can distinguish nonce nouns from adjectives based on the distributional context alone (Booth & Waxman, 2003) and can construct distribution-dependent grammatical paradigms for both natural and artificial languages (Gerken, Wilson, & Lewis, 2005; Gómez & Lakusta, 2004; Thothathiri, Snedeker, & Hannon, 2011).

Likewise, computational methods for automatically inducing syntactic categories use distributional information heavily, when not exclusively (Brown, Pietra, deSouza, Lai, and Mercer (1992), Cartwright and Brent (1997), Clark (2003), Harris (1946), Mintz (2003), and many others). The distributional context of a word in these models is defined using a few surrounding items (words or categories). This captures the immediate context, which is sufficient to learn reasonable categories, and avoids exploding the number of model parameters.

However, non-local syntactic effects play a role as well. In this paper, we add a form of non-local, sentence-level context to models using only local context, to investigate whether the added context improves performance. Specifically, we add *sentence type* as a known feature to the local context, i.e., whether the local context is within a question, declarative sentence, short fragment, etc. Sentence type often affects sentence structure and word order, and thereby can change the nature of local contexts. Taking sentence type into account may thus lead to clustering on the basis of more informative context distributions. An improvement in performance would indicate that this new information is useful to language learners, but it could also decrease performance if it is too noisy or does not correlate with syntactic category sequences.

Our enhanced models assume that children are aware of different sentence types and can make use of them at the stage of learning syntactic categories. A great deal of evidence from language development supports this assumption. Sentence types are strongly signaled by prosody in most languages (Hirst & Cristo, 1998). Prosody is, along with phonotactics, the first step in language learning; areas in the brains of three month olds are already sensitive to the prosody of the surrounding language (Homae, Watanabe, Nakano, Asakawa, & Taga, 2006) and experiments with newborn infants have demonstrated their ability to distinguish their native language using prosody alone (Mehler et al., 1988). Two month olds use prosody to remember heard utterances (Mandel, Jusczyka, & Kemler Nelson, 1994). Notably Mandel, Kemler Nelson, and Jusczyk (1996) showed that natural sentence level prosody aided memory of word order in two month old infants, which is essential for remembering and using distributional information.

Infants are aided in language learning by the fact that intonation (pitch) contours of child and infant directed speech (CDS) are especially well differentiated between sentence types, more than in adult directed speech (Fernald, 1989; Stern, Spieker, & MacKain, 1982). It is specifically the pitch contours of CDS that infants prefer over adult directed speech (Fernald & Kuhl, 1987) — the same contours that signal sentence type. CDS tends to be more interactive (as measured by the proportion of questions) than adult directed speech (Fernald & Mazzie, 1991; Newport, Gleitman, & Gleitman, 1977), resulting in a greater variety of frequent sentential prosody patterns and potentially making sentential prosody a more salient feature at the beginning of language learning (Stern, Spieker, Barnett, & MacKain, 1983). Visual cues, particularly the speaker's facial expression, can also be used to distinguish between questions and statements (Srinivasan & Massaro, 2003).

Infants' awareness of sentence types can be demonstrated by their sensitivity to the pragmatic function signaled by sentence type. For example, mothers will stop asking questions if infants do not react appropriately, as when the mother is interacting with time-delayed video feed of the infant (Murray & Trevarthen, 1986). Since CDS is characterized by a high proportion of questions, this demonstrates that in normal caretaker-child interactions infants as young as three months are 'holding up' their side of the conversation in some basic sense. Infants produce appropriate intonation melodies to communicate their own intentions at the one word stage, before they develop productive syntax (Balog & Brentari, 2008; Galligan, 1987; Snow & Balog, 2002). Children also exhibit adult-like behavior when using prosody to distinguish between otherwise identical Mandarin questions and declarative sentences in an on-line fashion (Zhou, Crain, & Zhan, 2012).

Based on these experimental results, we conclude that children who are at the point of learning syntax — at two to three years of age — are well equipped to use sentential prosody as part of their armory of potentially relevant input features. The current work investigates whether it would be advantageous for them to do so, given a classic computational model of syntactic category learning. To this end we annotate a corpus of child directed speech with sentence types, and extend the model to enable it to use these features.

We are not aware of previous work investigating the usefulness of sentence type information for syntactic category acquisition models. However, sentence types (identified by prosody) have been used to improve the performance of speech recognition systems. Specifically, Taylor, King, Isard, and Wright (1998) found that using intonation to recognize dialog acts (which to a large extent correspond to sentence types) and then using a specialized language model for each type of dialog act led to a significant decrease in word error rate.

In this paper we first examine the corpus data motivating our use of sentence types in syntactic category learning, and describe how we label sentence types. We then experiment with three different ways of incorporating sentence type into a token-based tagging model, the Bayesian Hidden Markov Model (BHMM). Our results demonstrate that sentence type is a beneficial feature for representing word order (or more precisely, syntactic category order).

## Data

One of the first detailed investigations of CDS (Newport et al., 1977) found that it differs from adult directed speech in a number of key ways; for example, child directed utterances are significantly shorter and more intelligible than adult directed speech, with fewer false starts and corrections. This emphasizes the need to use realistic (i.e., CDS) corpora when modeling acquisition: the linguistic environment in which children acquire language is unlike the standard corpora used in computational linguistics.

More immediately relevant to our current work is the fact that CDS is far more diverse in terms of sentence type than either adult written or spoken language. Whereas adult directed speech is largely made up of declarative utterances, CDS includes many more questions and imperative statements (Fernald & Mazzie, 1991; Newport et al., 1977). Indeed, one of the arguments for the utility of CDS (Gleitman, Newport, & Gleitman, 1984) is that it is the range and the complexity of input that enables a learner to delimit the linguistic space, that is, to successfully separate grammatical sentences from non-grammatical. If a learner was given an overly constrained language to begin with, she could construct wrong hypotheses that would not admit the more complex adult language she would be faced with later on.

The data we we use come from CHILDES (MacWhinney, 2000), a collection of corpora shared by language development researchers. We use the Eve corpus (Brown, 1973) and the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001). The Eve corpus is a longitudinal study of a single US American child from the age of 1;6 to 2;3 years, whereas the Manchester corpus follows a cohort of 12 British children from the ages of 2 to 3. We remove all utterances with any unintelligible words or words tagged as `quote` (about 5% of all utterances and between 2 and 3% of CDS utterances). Corpus statistics are presented in Table 1. The Manchester corpus is over twenty times as large as the Eve corpus; by inferring and evaluating our models on both corpora we can investigate the effects of data set size.

——————— Insert Table 1 about here ———————

Sentence types are not annotated in these corpora, so we use simple heuristics to label the sentences with their sentence type. The Cambridge Grammar of the English Language (Huddleston & Pullum, 2002) identifies the following five clause types:

**Declarative** *I have eaten the plums in the icebox.*

**Closed Interrogative/*yes*/*no* questions** *Were you saving them for breakfast?*

**Open Interrogative/*wh*-questions** *Why were they so cold?*

**Exclamatory** *What a sweet fruit!*

**Imperative** *Please forgive me!*

We use these clause types as a starting point. (To stress that we are labeling a full utterance/sentence, we will use the term *sentence type* rather than *clause type.*) We do not use the exclamatory sentence type due to its scarcity in the corpus; it is also difficult to identify automatically. Additionally, we add a short utterance category to distinguish probable fragments (verb-less clauses). The resulting sentence types with their identifying characteristics are:

**Open Interrogative/*wh*-questions (W):** Utterances ending with a question mark and beginning (in the first two words) with a '*wh*-word' (one of *who*, *what*, *where*, *when*, *why*, *how*, *which*).

**Closed Interrogative/*yes*/*no* questions (Q):** Utterances ending with a question mark but not beginning with a *wh*-word. This includes tag questions and echo questions with declarative (unmarked) word order.

**Imperative (I):** Utterances with an imperative-mood verb in the first two words[1].

**Short (S):** One- or two-word non-question utterances, typically interjections and fragments.

**Declarative (D):** All remaining utterances.

It would be preferable to use audio cues to categorize utterances, especially with regard to the difference between declaratives, closed interrogatives, and imperatives (since short utterances are easily identified by their length, and *wh*-words are a reliable cue for open interrogatives). Unfortunately the speech data is not available for our corpora, so we must approximate the audio cues available to children with the above orthographic

[1] Since the corpus we use, CHILDES, does not annotate the imperative mood, we use all utterances with a 'base' verb in the first two words without a pronoun or noun preceding it (e.g. *well go and get your telephone*).

(punctuation) and lexical-syntactic (*wh*-word and verb identification) cues. Note that the CHILDES annotation guidelines state that all utterances with question-characteristic intonation must be transcribed with a question mark, even if the utterance is syntactically declarative (*You ate the plums?*). In any case, even prosody data would not include all the cues available to the child such as visual cues, facial expressions, and so on.

Table 2 gives the number of each type of utterance in each corpus. Notably, while declaratives are the largest category, they make up only about a third of total utterances, while questions make up a further third of all utterances. (In contrast, questions make up only 3% of the Wall Street Journal corpus of news text (Marcus, Santorini, Marcinkiewicz, & Taylor, 1999) and 7% of the Switchboard corpus of adult conversations (Godfrey, Holliman, & McDaniel, 1992).)

———————— Insert Table 2 about here ————————

Note that these sentence-labeling heuristics are rather coarse and prone to noise. This is in line with the noisy input that children receive, presumably leading to errors on their part as well. Any model must be equally robust against noise and miscategorization. We hand-checked a subset of the Eve (section 10, the dev section) to verify our sentence type heuristics. Of 497 sentences, only 10 (2%) were misclassified[2].

## BHMM models with sentence types

In this section we add sentence type as an observed variable to a simple part of speech induction model, the Bayesian Hidden Markov Model (BHMM) (Goldwater & Griffiths, 2007). The HMM has long been a standard model for part of speech induction (Merialdo, 1994); reformulating it as a Bayesian model avoids some of the problems with standard maximum likelihood estimation and leads to more accurate clusters. (See Griffiths, Kemp, and Tenenbaum (2008) for an introduction to Bayesian methods from a cognitive modeling

[2] The misclassified sentences were: 8 imperatives classified as declaratives (mostly sentences of the form *Eve please listen*); one declarative sentence that is clearly a question in context but was not annotated with a question mark; one declarative sentence that was mistagged as an imperative due to an annotation error (*people* tagged as a verb).

perspective.) Both the Bayesian and classic HMM assign part of speech labels to a sequence of word tokens, rather than word types (such as the models of Redington, Chater, and Finch (1998) and Clark (2001)). This enables the model to deal with homographs in a natural way. It also ensures complete coverage of the data (all tokens must be tagged), whereas many type-based models limit themselves to only the most frequent items.

Modeling the underlying structure of a sentence as a sequence of part of speech labels is a particularly simplified view of syntax, removing the hierarchical structure inherent to language. However, this simplification seems appropriate for the first stages of acquisition, when learners encounter short sentences with simple syntax. Most computational models of syntax acquisition or grammar induction assume that syntactic categories are available (either annotated or learned by a separate method) and induce a grammar over syntactic categories rather than lexical items (Klein & Manning, 2004; Perfors, Tenenbaum, & Regier, 2011). A simpler model such as the HMM and its variants can demonstrate what can be learned before full grammatical structure is in place.

We first present an overview of the BHMM and introduce our notation (but see Goldwater and Griffiths (2007) for complete details); we then add sentence type evidence to the BHMM in three ways and describe the ensuing models.

## BHMM

HMMs are a classic model used to assign 'labels' (hidden state identities, in this case parts of speech) to a sequence of observations (word tokens). They are defined by two probability distributions: the transition distribution, which gives the probability of transitioning to a tag given the surrounding context (tags), and the emission distribution, which gives the probability of a particular word given its tag. The transition distribution characterizes the sequence of words, or rather tags, and thus can represent a low-level, local syntax concerned with word order (but not movement or long distance dependencies). The emission distributions characterize the set of words that are likely to appear as a given part of speech, i.e., the categorization of the words into categories.

Like a standard HMM, the Bayesian HMM makes the independence assumption that the probability of word $w_i$ depends only on the current tag $t_i$, and the probability of tag $t_i$ depends only on the previous tags. The number of previous tags included in the context is known as the order of the model: a first-order model, also known as a bigram model, uses a single previous tag as conditioning context to determine the transition probability, whereas a second-order (trigram) model uses the two previous tags as context.

The trigram transition and emission distributions are written as:

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

where $\tau_{(t',t'')}$ are the parameters of the multinomial distribution over following tags given previous tags $(t', t'')$ and $\omega_{(t)}$ are the parameters of the distribution over outputs given tag $t$. The generative process associated with the BHMM assumes that these parameters have in turn been drawn from symmetric Dirichlet priors with hyperparameters $\alpha$ and $\beta$, respectively:

$$\begin{aligned} \tau_{(t',t'')}|\alpha &\sim \mathrm{Dirichlet}(\alpha) && (3)\\ \omega_{(t)}|\beta &\sim \mathrm{Dirichlet}(\beta) && (4)\end{aligned}$$

Using these Dirichlet priors allows the multinomial distributions to be integrated out, leading to the following conditional posterior distributions:

$$\begin{aligned} P(t_i|\mathbf{t}_{-i},\alpha) &= \frac{n_{t_{i-2},t_{i-1},t_i}+\alpha}{n_{t_{i-2},t_{i-1}}+T\alpha} && (5)\\ P(w_i|t_i,\mathbf{t}_{-i},\mathbf{w}_{-i},\beta) &= \frac{n_{t_i,w_i}+\beta}{n_{t_i}+W_{t_i}\beta} && (6)\end{aligned}$$

where $\mathbf{t}_{-i} = t_1 \ldots t_{i-1}$, all tags but $t_i$, and likewise $\mathbf{w}_{-i} = w_1 \ldots w_{i-1}$, all words but $w_i$.

$n_{t_{i-2},t_{i-1},t_i}$ and $n_{t_i,w_i}$ are the number of occurrences of the trigram $(t_{i-2}, t_{i-1}, t_i)$ and the tag-word pair $(t_i, w_i)$ in $\mathbf{t}_{-i}$ and $\mathbf{w}_{-i}$. $T$ is the size of the tagset and $W_t$ is the number of word types emitted by $t$. The hyperparameters function as a type of smoothing, with $\alpha$ and $\beta$ providing pseudo-counts.

Clearly, the Markov independence assumption made in the HMM/BHMM, namely that a small context window is sufficient to determine the hidden label, is too strong in natural language contexts, but adding more context leads to an unwieldy and sparse model. One of the important questions for the models presented here is whether sentence types can proxy for larger context windows.

## BHMM with sentence types

The BHMM depends solely on local information for tagging. However global information — such as sentence types — can play a role in syntax at the tag sequence level, by requiring shifts in word order. Hence they are likely to be informative for a tagger by enriching the impoverished local context representation.

In order to incorporate sentence type information into the BHMM, we add an observed variable to each time-step in the model with the value set to the current sentence type[3]. Given that the BHMM consists of two principal distributions, there are two straightforward ways that sentence type could be incorporated into the BHMM: either by influencing the transition probabilities or the emission probabilities. The former would reflect the effect of sentence type on word order, whereas the latter would investigate whether sentence type affects the set of words categorized as a single part of speech. We discuss both, as well as their combination.

**BHMM-T.** In the first case, transitions are conditioned not only on previous context, as in the BHMM, but also on the context's sentence type. This leads different

[3] Arguably sentence type only needs to be included in the model once per sentence, rather than at each time-step, since sentence type never changes within a sentence. However, since sentence type is an observed variable, replicating it has no effect, and it makes the notation clearer.

sentence types to assign different probabilities to the same sequence of tags, so that, for example, `PRONOUN` will be more likely to be followed by a `VERB` in declaratives than in imperatives. (Note however that the estimated tag clusters will not necessarily correspond to gold tags.) By separating out the transitions, the model will have more flexibility to accommodate word order changes between sentence types.

Formally, the observed sentence type $s_{i-1}$ is added as a conditioning variable when choosing $t_i$, i.e., we replace line 1 from the BHMM definition with the following:

$$t_i | s_{i-1} = s, t_{i-1} = t', t_{i-2} = t'', \tau_{(s,t',t'')} \;\sim\; \mathrm{Mult}(\tau_{(s,t',t'')}) \tag{7}$$

We refer to this model, illustrated graphically in Fig. 1, as the BHMM-T (for transitions).

——————— Insert Figure 1 about here ———————

The BHMM-T has a larger number of parameters than the BHMM, which has $T^{o+1} + TV$ (where $T$ is the number of tags, $o$ is the model order, and $V$ is the size of the vocabulary) parameters, whereas the BHMM-T has $ST^{o+1} + TV$ ($S$ being the number of sentence types)[4].

**BHMM-E.** Analogously, we can add sentence type as a conditioning variable in the emission distribution by replacing line 2 from the BHMM with

$$w_i | s_i = s, t_i = t, \omega_{(s,t)} \;\sim\; \mathrm{Mult}(\omega_{(s,t)}) \tag{8}$$

This model, the BHMM-E (for emissions), results in models in which each sentence type has a separate distribution of probable words for each tag, but the transitions between those tags are shared between all sentence types, as in the BHMM. This does not correspond well to the word-order effect that sentence type has in many languages, but may capture vocabulary differences between sentence types, if these exist.

[4] Since probability distributions are constrained to sum to one, the last parameter in each distribution is not a free variable, and so the true number of necessary emission parameters is $T(V-1)$ (and likewise for transition parameters), but we omit this technicality in favor of clarity.

The model size is $T^{o+1} + STV$, which in practice is significantly larger than the BHMM-T model, given $V \gg T > S$ and model orders of one and two (bigram and trigram models).

**BHMM-ET.** The combination of the two, BHMM-T plus BHMM-E, is also possible. In the BHMM-ET, sentence type conditions both transition and emission probabilities. Each sentence type now has a separate set of transition and emission distributions (both transitions and emissions are conditionally independent given sentence type). Without any shared information, tags are not in any sense equivalent between sentence types, so this model is equivalent to inferring a separate BHMM on each type of sentence, albeit with shared hyperparameters.

Introducing the sentence type parameter as an extra conditioning variable in these models has the consequence of splitting the counts for transitions, emissions, or both. The split distributions will therefore be estimated using less data, which could degrade performance if sentence type is not a useful predictor of tag sequences or tag-word pairings. This will be especially vital to the performance of the BHMM-ET, without any shared information at all. If the separate models in the BHMM-ET match the BHMM's performance, this would indicate that sentence type is as reliable an indicator of tagging information as a large amount of additional data from other sentence types. However, it is cognitively implausible for there to be no sharing of information at all between sentence types: this model serves principally as a measure of sentence type informativeness.

Our prediction is that sentence type is more likely to be useful as a conditioning variable for transition probabilities (BHMM-T) than for emission probabilities (BHMM-E). For example, the auxiliary inversion in questions is likely to increase the probability of the `AUX` → `PRONOUN` transition, compared to declaratives. Knowing that the sentence is a question may also affect emission probabilities, e.g. it might increase the probability the word *you* given a `PRONOUN` and decrease the probability of *I*; one would certainly expect *wh*-words to have much higher probability in *wh*-questions than in declaratives. However, many other variables also affect the particular words used in a sentence (principally, the current semantic and pragmatic context). We expect that sentence type plays a relatively small role compared to these other factors. The ordering of tags within an utterance, on the other hand, is primarily constrained by sentence type, especially in the short and grammatically simple utterances found in child-directed speech.

## English experiments

### Procedure

**Corpora.** We use the Eve and Manchester corpora from CHILDES for our experiments. From both corpora we remove all utterances spoken by a child; the remaining utterances are nearly exclusively CDS.

Although our model is fully unsupervised (meaning the gold standard tags are never visible to the model), files from the chronological middle of each corpus are set aside for development and testing evaluation (Eve: file 10 for development, 11 for testing; Manchester: file 16 from each child for development, file 17 for testing). The remainder of each corpus is used for inference only, and is never evaluated. The BHMM is thus inferred using either (dev+remainder) or (test+remainder) datasets, with only the inferred tags in either the dev or test portion being evaluated.

The motivation behind this evaluation regime is to ensure that the model structures we investigate generalize to multiple settings. In most previous unsupervised modeling work, results are reported for the entire corpus. However, since different model structures and (hyper)parameterizations may be explored during development, this methodology still leaves open the possibility that the final model structure may be better suited to the particular corpus being used than to others. To avoid this issue, we set aside a separate test dataset that is only used for the final evaluation.

Both corpora have been tagged using the relatively rich CHILDES tagset, which we collapse to a smaller set of thirteen tags: adjectives, adverbs, auxiliaries, conjunctions, determiners, infinitival-to, nouns, negation, participles, prepositions, pronouns, verbs and other (communicators, interjections, fillers and the like). *Wh*-words are tagged as adverbs (*why*, *where*, *when* and *how*), pronouns (*who* and *what*), or determiners (*which*).

Each sentence is labeled with its sentence type using the heuristics described earlier. Dummy inter-sentence markers are added, so transitions to the beginning of a sentence will be from the inter-sentence hidden state (which is fixed). The inter-sentence markers are assigned a separate dummy sentence type.

In our experiments, we experiment with coarser sentence type categorizations as well as the full five types. This enables us to discover which sentence types are most informative for the tagging task. Specifically, we try:

**QD** in which all questions (*wh*- and other) are collapsed to one question category and all other utterances are collapsed to declaratives.

**WQD** in which the question categories are separated but the non-questions are in a single category.

**SWQD** as above, but with short (declarative) utterances distinguished from other non-question utterances.

**ISWQD** in which all five sentence types are separated: imperatives, short utterances, *wh*-questions, other questions, and declaratives.

**Inference.** We follow Goldwater and Griffiths (2007) in using a collapsed Gibbs sampler to perform inference over the BHMM and its variants. Gibbs samplers are a standard batch inference method for Bayesian models and are designed to produce samples from the posterior distribution of interest—here, the distribution over tag assignments given the observed word sequences $P(\mathbf{t}|\mathbf{w}, \alpha, \beta) \propto P(\mathbf{w}|\mathbf{t}, \beta)P(\mathbf{t}|\alpha)$. The sampler is initialized by assigning a random tag to each word, and then runs iteratively over the corpus, resampling the tag assignment for each word from the posterior distribution defined by the current tag assignments for all other words, $P(t_i|t_{-i}, w, \alpha, \beta)$ (where $t_{-i}$ represents the other tag assignments). It can be shown that this procedure will eventually converge to producing samples from the true posterior distribution over all tags (Geman & Geman, 1984). The equations used to resample each word's tag assignment in each model are shown in Fig. 2.

——————— Insert Figure 2 about here ———————

In order to minimize the amount of prior information given to the model, we also optimize the hyperparameters $\alpha$ and $\beta$ using using a Metropolis-Hastings sampling step, following Goldwater and Griffiths (2007). The hyperparameter over emission distributions, $\beta$, is estimated separately for each tag distribution, whereas $\alpha$ is constrained to be the same for all transition distributions.

We set the number of hidden states (corresponding to tag clusters) to the number of gold tags in the corpus (i.e., 13).

We run the Gibbs sampler for 10000 iterations, with hyperparameter resampling and simulated annealing. (On the Manchester corpus, due to its size, we only manage 5000 iterations.) The final sample is used for evaluation purposes. Since Gibbs sampling is a stochastic algorithm, we run all models multiple times and report average values for all evaluation measures as well as confidence intervals at the 95% level. Significance is measured using the non-parametric Wilcoxon signed-rank test.

**Evaluation measures.** Evaluation of fully unsupervised part of speech tagging is known to be problematic, due to the fact that the part of speech clusters found by the model are unlabeled, and do not automatically correspond to any of the gold standard part of speech categories. Commonly used evaluation measures, such as one-to-one and many-to-one accuracy, suffer from "the problem of matching" (Meila, 2007): their greedy 'winner takes all' approach means that all unmatched items are rated as equally bad, despite the fact that a cluster with members of only two gold classes is intuitively better than a cluster that is made up of many gold classes.

For evaluation we thus use *V-measure* (VM; Rosenberg and Hirschberg, 2007), which avoids this issue[5]. VM uses the conditional entropy of clusters and categories to evaluate clusterings. It is analogous to the F-measure commonly used in NLP, in that it is the harmonic mean of two measures analogous to precision and recall, homogeneity (VH) and completeness (VC). VH is highest when the distribution of categories within each cluster is highly skewed towards a small number of categories, such that the conditional entropy of categories $C$ given the clusters $K$, $H(C|K)$, is low. Conversely, VC measures the conditional entropy of the clusters within each gold standard category, $H(K|C)$, and is highest if each category maps to a single cluster so that each model cluster completely contains a category.

$$VH = 1 - \frac{H(C|K)}{H(C)} \qquad VC = 1 - \frac{H(K|C)}{H(K)} \qquad VM = \frac{2 \times VH \times VC}{VH + VC}$$

Like the standard F-measure, VM ranges from zero to one (rescaled to 0-100). It is invariant with regards to both the number of items in the dataset and to the number of clusters used, and consequently it is suitable for comparing results across different corpora.

## Results and discussion

We now present results for the three BHMM variants discussed in the previous section: a model in which both transitions and emission distributions are separated by sentence type, the BHMM-ET; a model in which only emission distributions are separated by sentence type and transitions are shared, the BHMM-E; and the converse, the BHMM-T, in which transitions are separated and emission distributions are shared. We find that these models perform in very different ways, demonstrating the effect of sentence type on word order rather than word usage.

[5] We also calculated many-to-one accuracy and found that it was highly correlated with VM.

### BHMM-ET: sentence-type-specific sub-models

By including a sentence type indicator in both the transition and emission distributions, the BHMM-ET separates both transition and emission probabilities by sentence type, effectively creating separate sub-models for each sentence type. The resulting tags are thus not equivalent between sentence types, i.e., the tag `TAG-9` used in a declarative sentence is not the 'same' `TAG-9` as used in questions, since they have distinct transition and emission distributions. Consequently, when evaluating the tagged output each sentence type must be evaluated separately, to avoid conflating incompatible clusterings.

——————— Insert Figure 3 about here ———————

——————— Insert Figure 4 about here ———————

Baseline BHMM and BHMM-ET results for the Eve corpus, split by sentence type, are in Fig. 3 and 4. We see that in this relatively small corpus, as expected, splitting the sentence types results in decreased performance as compared to the baseline BHMM, in which all counts are shared. Only the declaratives, the most frequent sentence type, provide enough information on their own to match (and in the dev settings exceed) the baseline performance.

——————— Insert Figure 5 about here ———————

——————— Insert Figure 6 about here ———————

Fig. 5 and 6 show results for the much larger Manchester corpus. The BHMM-ET models here perform much closer to the baseline, due to the larger amount of data available to each sentence type sub-model. Each of the sentence types contain enough information to learn approximately equivalent taggers using either only single sentence type data or the full data set. The exceptions are the short utterances models, which continue to suffer from the lack of sufficiently informative contexts.

### BHMM-E: sentence-type-specific emissions

We now turn to the BHMM-E, in which emission probability distributions are sentence-type-specific, but transition probabilities are shared between all sentence types. In this model a given sequence of tags is equally likely among all sentence types, but those tags can correspond to different words in different sentence types.

Returning to Fig. 3 and 4 (Eve corpus) and Fig. 5 and 6 (Manchester corpus), we see that for almost every sentence type, the BHMM-E performs worse than both the BHMM-ET and the baseline BHMM. The negative effect of the split emissions is most striking on the Manchester corpus, where small dataset size cannot be a problem. Whereas with the BHMM-ET we might posit that, given enough data, sentence-type-specific models would learn an equivalent model to the shared baseline model (apart from the short utterance distinction), here we see that adding the sentence type feature to only the emissions is actively harmful.

### BHMM-T: sentence-type-specific transitions

Lastly, we evaluate the BHMM-T, which shares emission probabilities among sentence types and uses sentence-type-specific transition probabilities. This describes a model in which all sentence types use the same set of tags, but those tags can appear in different orders in different sentence type. Since this situation corresponds best to the behavior of sentence type in English and many other languages — word order changes according to sentence type, but word usage does not — we expect the BHMM-T to perform the best of all the sentence type BHMM models, and to improve over the baseline BHMM.

In the bigram models (Fig. 3 for the Eve corpus and Fig. 5 for the Manchester corpus) we see the BHMM-T either match or improve upon the baseline BHMM in all sentence types, apart from short utterances. Models trained on the Manchester corpus improve significantly over the baseline on all other sentence types. Models trained on the smaller Eve corpus show more mixed results, with the most constrained sentence types (*wh*-questions and imperatives) showing significant improvement over the baseline.

Trigram models (Fig. 4 and 6) also differ by dataset size. On the smaller dataset, we find no consistent difference in performance between BHMM-T and baseline; notably short utterances also match baseline performance (unlike in bigram models). Given more data, the Manchester-trained models significantly outperform the baseline on two sentence types, and match on the others. Imperatives, the least frequent sentence type, suffer a decline in performance; there may simply not be enough data to estimate good trigram transition distributions, even with shared emissions.

Trigram models must estimate significantly more parameters than bigram models; adding sentence type increases the number of parameters still further. Where BHMM-T matches and exceeds baseline performance, adding sentence type is creating split transition distributions that are more or equally accurate to the fully shared distribution, despite being estimated on far less data. This demonstrates the potential effectiveness of sentence type information for transitions.

Not all sentence types seem to be useful: bigram models perform poorly on short utterances, due to the limited context; the baseline model makes use of transfer from other sentence types. Rare sentence types such as imperatives are difficult to learn, particularly for trigam models. These differences indicate that it may be advantageous to only distinguish certain sentence types.

——————— Insert Figure 7 about here ———————

——————— Insert Figure 8 about here ———————

Fig. 7 (Eve) and Fig. 8 (Manchester) show the results of BHMM-T models trained on data sets using a variety of less fine-grained sentence types, as described earlier. The shared emission distributions in the BHMM-T allow us to evaluate the corpus as a whole, given a single clustering for all sentence types.

On the Eve dataset we see a difference in baseline performance between the dev and test sets that is not reflected in the BHMM-T models; the test set is easier for the baseline than the dev set. In the BHMM-T, all sentence types apart from declaratives do slightly better on the test set as well, but because declaratives are so frequent, BHMM outperforms BHMM-T significantly on the test set. We suspect this is due to peculiarities in the test data portion: if we evaluate (as is common with unsupervised models) on the full dataset used for inference at test time (i.e., the full corpus apart from the development section), we get results that are more similar to the dev-only results, where nearly all BHMM-T variants significantly improve over the baseline, both in trigram and bigram models.

We were unable to ascertain what the relevant characteristics of the dev and test set was that caused the difference in performance, but note that both the dev and test set are small (approximately two thousand tokens) so some variation is to be expected.

Examining the sentence type variants, on the Eve datasets we see bigram models do best with data annotated with three sentence types, two question types and a general declaratives sentence type (WQD), whereas trigrams benefit from a simpler question/other (QD) distinction. On the Manchester corpus, bigram models also perform best with the WQD split, but nearly all bigram BHMM-T models improve significantly over the baseline, regardless of the exact sentence type distinctions used. Trigram BHMM-T are less consistently better than the baseline, but also do not perform worse. The best performance is achieved when using the full set of sentence types (ISWQD). However, the larger amount of data used to infer trigram tag sequences leads to all models performing approximately equally well, with or without sentence type information.

When trained on the Eve corpus, the trigram BHMM-T does not have sufficient data to accurately infer categories when the transitions are split between too many sentence types, and performs best with only two sentence types. On the other hand, when more data is available, the trigram BHMM-T is able to estimate good parameters for all five sentence types: the models trained on the Manchester corpus with all sentence types outperform all the others. The improvement over the baseline and other models is slight, however, indicating that sentence type is providing minimal additional information in this case.

In contrast, when evaluated on the Eve development corpus, the bigram model with the best set of sentence type labels (WQD) performs as well as the trigram model without sentence types. In this case, sentence-type-specific bigram transitions are as informative as transitions with twice as much local context information, leading to a model with fewer parameters but equal performance. However, it is important to note that even in cases where sentence type does not seem to add additional information in the transition distribution, it never decreases performance (as it did when added to the emission distribution). This indicates that at worst, sentence type carries the same information (but no more) as the context history already available in the BHMM.

In summary, based on experiments using English corpora we have found that separate emission distributions between sentence type are harmful (BHMM-ET and BHMM-E), whereas separate transitions for sentence types may be helpful. This is in line with our predictions, based on the fact that in English sentence types primarily affect word order.

Cognitively, separate emission distributions would be hard to justify, since they result in non-corresponding syntactic categories between sentence types. In these models, each sentence type has a separate set of syntactic categories, which means that e.g. *cat* and *mouse* must be clustered together separately for each sentence type. Such models, in which categories are replicated multiple times and differ between sentence types, clearly do not make efficient use of limited input data.

Unlike the words making up syntactic categories, word order does change between sentence types in many languages, and taking this into account by learning separate word orders for each sentence type seems to be an effective strategy. Here we found that the choice of sentence type categories matters, and is dependent on the amount of input data available: with larger amounts of data, finer sentence type categories can be used.

## Cross-linguistic experiments

In the previous section we found that sentence type information improved syntactic categorization in English. In this section, we evaluate the BHMM's performance on two languages other than English, and investigate whether sentence type information is useful across languages.

Nearly all human languages distinguish between closed *yes/no*-questions and declaratives in intonation. Open questions are most commonly marked by rising intonation (Hirst & Cristo, 1998). *Wh*-questions do not always have a distinct intonation type, but they are signaled by the presence of members of the small class of *wh*-words.

We use tagged corpora for Spanish and Cantonese from the CHILDES collection: the Ornat corpus (Ornat, 1994) and the Lee Wong Leung (LWL) corpus (Lee et al., 1994) respectively. We describe each corpus in turn below; Table 3 lists their relative sizes.

——————— Insert Table 3 about here ———————

**Spanish.** The Ornat corpus is a longitudinal study of a single child between the ages of one and a half and nearly four years, consisting of 17 files. Files 08 and 09 are used for testing and development. We collapse the Spanish tagset used in the Ornat corpus in a similar fashion to the English corpora. There are 11 tags in the final set: adjectives, adverbs, conjuncts, determiners, nouns, prepositions, pronouns, relative pronouns, auxiliaries, verbs, and other.

Spanish *wh*-questions are formed by fronting the *wh*-word (but without the auxiliary verbs added in English); *yes/no*-questions involve raising the main verb (again without the auxiliary inversion in English). Spanish word order in declaratives is generally freer than English word order. Verb- and object-fronting is more common, and pronouns may be dropped (since verbs are marked for gender and number). Note that verb-fronted declaratives will have the same structure as closed questions. This suggests that there will be fewer clear differences between transition distributions in the various sentence types.

**Cantonese.** The LWL corpus consists of transcripts from a set of eight children followed over the course of a year, totaling 128 files. The ages of the children are not matched, but they range between one and three years old. Our dataset consists of the first 500 utterances of each file, in order to create a data set of similar size as the other corpora used. Files from children aged two years and five months are used as the test set; files from two years and six months make up the development set.

The tagset used in the LWL, which we use directly, is larger than the collapsed English tagset. It consists of 20 tags: adjective, adverb, aspect marker, auxiliary or modal verb, classifier, communicator, connective, determiners, genitive marker, preposition or locative, noun, negation, pronouns, quantifiers, sentence final particle, verbs, *wh*-words, foreign word, and other. We remove all sentences that are encoded as being entirely in English but leave single foreign, mainly English, words (generally nouns) in a Cantonese context.

Cantonese follows the same basic SVO word order as English, but with a much higher frequency of topic-raising. Questions are not marked by different word order. Instead, particles are inserted to signal questioning. These particles can signal either a yes/no-question or a *wh*-question; in the case of *wh*-questions they replace the item being questioned (e.g., *playing-you what?*), without *wh*-raising as in English or Spanish. Strictly speaking the only syntactic change in transitions would thus be an increase in transitions to and from the *wh*-particles in questions. However, there may be other systematic differences between questions and declaratives.

**Results.** We inferred BHMM and BHMM-T models in the same manner as with the English corpora (10 runs each, 10000 iterations, with simulated annealing and hyperparameter estimation).

Due to inconsistent annotation and lack of familiarity with the languages, we used only three sentence types: open/*wh*-questions, other questions, and declaratives. Punctuation was used to distinguish between questions and declaratives. *wh*-questions were identified by using a list of *wh*-words for Spanish; the Cantonese corpus included a *wh*-word tag.

—————— Insert Figure 9 about here ——————

In English, the BHMM-T was able to improve performance by taking into account the distinct word orders characteristic of the different sentence types. Spanish does not show the same improvement (Fig. 9). The estimated BHMM-T models do not differ significantly from the baseline BHMM; however they have much higher variance. This indicates that the BHMM-T is harder to estimate, presumably because the separate transitions merely introduce more parameters without offering the same benefits as in English. Given that sentence type does not affect word order in the same way in Spanish as in English, this is an unsurprising result.

—————— Insert Figure 10 about here ——————

In Cantonese, we see a significant improvement for the bigram BHMM-T (Fig. 10). This is despite the fact that Cantonese has relatively little word order marking of questions; the BHMM-T was able to make use of the extra information. The tagging of *wh*-questions improves most in the BHMM-T in bigram Cantonese models, but declaratives and other questions also improve slightly. Trigram BHMM-T models do not outperform the baseline; as in the larger English models, sentence type does not add significant new information to the trigram context history.

As in English, none of the Spanish or Cantonese BHMM-T models perform significantly worse than the BHMM baseline. Even when sentence type is not an entirely reliable signal to word order, the separately estimated transitions still match the performance of shared transitions.

## Conclusion

We have investigated whether sentence type can be a useful cue for models of syntactic category acquisition. The structure of the BHMM made it possible to distinguish between adding sentence types to either the parameters governing the order of syntactic categories in an utterance (transition distributions), or to the parameters describing the words making up the syntactic categories (emission distributions). We found that, as expected, adding sentence type to emission distributions resulted in degraded performance due to the decreased amount of data available to estimate each separate sentence type emission distribution; additionally these models are awkward since they create separate sets of clusters for each sentence type. This is contrary to syntactic categories as they are generally understood and leads to a representation with high levels of redundancy. Hence we dismiss this model structure for syntactic category acquisition.

Our models with sentence-type-specific transitions demonstrate that sentence type can be an informative cue for word and tag order, especially in models with limited amounts of local context. The amount of input data available to the model and the number of model parameters affected which set of sentence types performed best; further research is necessary to characterize these interactions more precisely. However, we showed that at the very least sentence-type-specific distributions never found worse syntactic clusters than shared distributions, and in many cases found better clusters. Arbitrarily adding parameters — and thereby splitting distributions — in unsupervised models is unlikely to improve performance, due to ensuing data sparsity, unless the parameter is genuinely useful (unlike supervised models, in which unhelpful parameters will be ignored). This problem arose when the sentence type parameter was added to the emission distribution. On the other hand, adding this parameter to the transition distribution resulted in an larger number of distributions estimated on smaller amounts of data, but these models were able to recover or improve on the original performance. This demonstrates that the sentence type parameter added useful additional information to the model in this setting.

Sentence type as an additional feature for syntactic category learning is not limited to the BHMM structure. Models that use frames (sets of surrounding words) as features for distributional categorization (Mintz, 2003; St Clair, Monaghan, & Christiansen, 2010) could extend the frame description to include the sentence type the frame appears in. Other models represent individual context words independently, using a vector of word co-occurrences as features for clustering (Christodoulopoulos, Goldwater, & Steedman, 2011; Redington et al., 1998; Toutanova & Johnson, 2007). In this case each context word would have to be annotated with the current sentence type, in order to separate the contexts by sentence type. Adding sentence type to these models could improve accuracy. For example, one of the most common frames in Mintz (2003) is *do_want*, which in declaratives is nearly always filled with *not* and in questions with *you*; separating this frame by sentence type would prevent these words from being clustered together.

Recently, two models of online learning of categories have been presented (Chrupala & Alishahi, 2010; Parisien, Fazly, & Stevenson, 2008), both of which allow for learning an unlimited number of categories. These models both use independent occurrences of context words as features and could be easily extended to use context words plus sentence type, as outlined above. Adding sentence type to these models could allow us to track the usefulness of sentence type during the learning process, unlike the static model we presented here. Based on our results on data sets of different sizes, we hypothesize that initially using coarser sentence types will be advantageous, but as more data is seen, more fine-grained sentence types will become helpful.

Our study has shown that computationally there may be an advantage to a representation of word/syntactic category order that includes sentence type information. This result naturally leads to the question of whether human learners do in fact have such a linguistic representation. Experimental work has demonstrated infants' awareness of adjacent contexts as cues for categorization at 12 months (Gómez & Lakusta, 2004). At this stage, they are also clearly aware of prosody, both in its function to signal dialog pragmatics as well as using it as a feature for storage of heard phrases (Mandel et al., 1996). However, it has yet to be shown that infants' representation of early syntax, when they are learning syntactic categories, allows for a link between prosody and word order, i.e., that infants can make use of the fact that different prosodies may signal different word orders. An artificial language learning experiment for learning categories, similar to Gómez and Lakusta's (2004) study, but in which transitions probabilities are manipulated along with prosody cues, could provide evidence for infants' ability (or otherwise) to link word order and prosody. Crucially, infants would also have to be able to ignore prosody when it is not informative, given that the link between word order and prosody is language dependent and only informative in certain contexts.

The model we have presented for making this link between sentence type prosody and word order is brittle: the different sentence types share no information about transitions. A more realistic model would interpolate between sentence-type-specific transition distributions, where these are more informative, and general distributions, where sentence type information is lacking, in order to be more robust to issues of data sparsity. It would also make use of the fact that some transitions are often shared between sentence types (i.e., nouns are often preceded by determiners in both questions and declaratives). In future work we will investigate adding this flexibility to the model. However, our work has demonstrated that adding a link between sentence type prosody and word order at minimum does not add noise, and in many cases is a source of additional information.

The Bayesian probabilistic modeling methodology gives us a principled way to add components to a model, either as observed variables representing input readily available to the learner, or as latent variables to be learned. Here, we explored a model in which the additional information (sentence type) was treated as observed, on the assumption that infants could use prosody to determine sentence type before beginning to learn syntactic categories. However, it is probably more plausible to assume that, since the relationship between prosody and sentence type needs to be learned, sentence type information may be only partially available when syntactic category acquisition begins. This suggests that ultimately we will need to consider learning sentence types and syntactic categories as a *joint* learning problem, where partial information about each task constrains the learning of the other. Bayesian models such as ours are a particularly apt way to model joint learning, with different latent variables used to represent different types linguistic information. Recent Bayesian models have shown that joint learning can be highly successful in areas such as syntax/semantics (Kwiatkowski, Goldwater, Zettelmoyer, & Steedman, 2012; Maurits, Perfors, & Navarro, 2009), syntax/morphology (Sirts & Alumäe, 2012) and phonetic/lexical learning (Elsner, Goldwater, & Eisenstein, 2012; Feldman, Griffiths, & Morgan, 2009; Feldman, Griffiths, Goldwater, & Morgan, in submission). We are also beginning to see experimental evidence supporting the cognitive plausibility of joint learning (Feldman, Myers, White, Griffiths, & Morgan, 2011). Although our models as they stand do not perform joint learning, nevertheless we feel they support this general program by emphasizing that multiple sources of information can improve learning, and suggesting what some of those sources could be. Ultimately, a complete model of language acquisition will need to consider interactions between all aspects of language. The models presented in this paper provide an example of the potential utility of this holistic, integrative approach to language acquisition.

# References

Balog, H. L., & Brentari, D. (2008). The relationship between early gestures and intonation. *First Language*, *28*(2), 141–163.

Booth, A. E., & Waxman, S. R. (2003). Mapping words to the world in infancy: infants' expectations for count nouns and adjectives. *Journal of Cognition and Development*, *4*(3), 357–381.

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Brown, R. (1973). *A first language: the early stages.* Harvard University Press.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2011). A Bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of the conference on empirical methods in natural language processing.*

Chrupala, G., & Alishahi, A. (2010). Online entropy-based model of lexical category acquisition. In *Proccedings of the 14th conference on natural language learning.*

Clark, A. (2001). Unsupervised induction of stochastic context-free grammars using distributional clustering. In *Proceedings of the workshop on computational natural language learning.*

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th annual meeting of the European association for computational linguistics* (pp. 59–66).

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th annual meeting of the association of computational linguistics.*

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the cognitive science society (cogsci)*.

Feldman, N., Myers, E., White, K., Griffiths, T., & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th boston university conference on language development*.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (in submission). A role for the developing lexicon in phonetic category acquisition.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Development*, *60*(6), 1497–1510.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279–293.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221.

Galligan, R. (1987). Intonation with single words: purposive and grammatical use. *Journal of Child Language*, *14*, pp 1–21.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of Child Language*, *11*, 43–79.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th annual meeting of the association of computational linguistics.*

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, *7*(5), 567–580.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling.* Cambridge University Press.

Harris, Z. (1946). From morpheme to utterance. *Language*, *22*(3), 161–183.

Hirst, D., & Cristo, A. D. (Eds.). (1998). *Intonation systems: a survey of twenty languages.* Cambridge University Press.

Homae, F., Watanabe, H., Nakano, T., Asakawa, K., & Taga, G. (2006). The right hemisphere of sleeping infant perceives sentential prosody. *Neuroscience Research*, *54*(4), 276 –280.

Huddleston, R. D., & Pullum, G. K. (2002). *The Cambridge grammar of the English language.* Cambridge University Press.

Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of the conference on empirical methods in natural language processing.*

Kwiatkowski, T., Goldwater, S., Zettelmoyer, L., & Steedman, M. (2012). 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.*

Lee, T. H., Wong, C. H., Leung, S., Man, P., Cheung, A., Szeto, K., & Wong, C. S. P. (1994). *The development of grammatical competence in Cantonese-speaking children.*

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mandel, D. R., Jusczyka, P. W., & Kemler Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, *53*, 155–180.

Mandel, D. R., Kemler Nelson, D. G., & Jusczyk, P. W. (1996). Infants remember the order of words in a spoken sentence. *Congnitive Development*, *11*, 181–196.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: the ontogenesis and representation of syntactic categories. *Children's language*, *2*, 127–214.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). Treebank-3. Linguistic Data Consortium.

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In *Proceedings of the 42nd annual conference of the cognitive science society.*

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Meila, M. (2007). Comparing clusterings — an information-based distance. *Journal of Multivariate Analysis*, *98*, 873–895.

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, *20*(2), 155–172.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Murray, L., & Trevarthen, C. (1986). The infant's role in mother-infant communications. *Journal of Child Language*, *13*, 15–29.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: language input and acquisition* (pp. 109–149). Cambridge, UK: Cambridge University Press.

Ornat, S. L. (1994). *La adquisicion de la lengua espagnola*. Madrid: Siglo XXI.

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. In *Proceedings of the 12th conference on computational natural language learning* (pp. 89–96). Manchester.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306 –338.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425 –469.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: a conditional entropy-based external cluster evaluation measure. In *Proceedings of the conference on empirical methods in natural language processing.*

Sirts, K., & Alumäe, T. (2012). 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.*

Snow, D., & Balog, H. (2002). Do children produce the melody before the words? A review of developmental intonation research. *Lingua*, *112*, 1025–1058.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. *Language and Speech*, *46*, 1–22.

St Clair, M., Monaghan, P., & Christiansen, M. (2010). Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, *116*(3), 341–360.

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: infant age and context related changes. *Journal of Child Language*, *10*, 1–15.

Stern, D. N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, *18*(5), 727–735.

Taylor, P. A., King, S., Isard, S. D., & Wright, H. (1998). Intonation and dialogue context as constraints for speech recognition. *Language and Speech*, *41*(3), 493–512.

Theakston, A., Lieven, E., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, *28*, 127–152.

Thothathiri, M., Snedeker, J., & Hannon, E. (2011). The effect of prosody on distributional learning in 12- to 13-month-old infants. *Infant and Child Development*.

Toutanova, K., & Johnson, M. (2007). A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Advances in neural information processing systems 20*.

Zhou, P., Crain, S., & Zhan, L. (2012). Sometimes children are as good as adults: the pragmatic use of prosody in childrenâăźs on-line sentence processing. *Journal of Memory and Language*, *67*, 149–164.

Table 1
*Summary of Eve and Manchester corpora*

| | Eve | | Manchester | |
|---|---|---|---|---|
| | All | CDS only | All | CDS only |
| Utterances | 25295 | 14450 | 582375 | 318349 |
| Mean Length | 4.68 | 5.38 | 4.66 | 4.31 |
| Word Tokens | 118372 | 77766 | 2713672 | 1371936 |
| Word Types | 2235 | 1995 | 11739 | 11030 |

Table 2

*Number of child-directed utterances by sentence type and average utterance length in parentheses.*

| Sentence Type | Eve | Manchester |
|---|---|---|
| *wh*-Questions | 2273 (4.03) | 33461 (4.72) |
| Other Questions | 2577 (4.41) | 74327 (5.80) |
| Declaratives | 6181 (5.79) | 99318 (5.83) |
| Short Utterances | 2752 (1.27) | 95518 (1.28) |
| Imperatives | 665 (5.30) | 15725 (5.17) |

Table 3

*Counts of sentence types (average utterance length in words) in Spanish and Cantonese data sets (without test and dev sets).*

| Sentence Type | Spanish (Ornat) | Cantonese (LWL) |
|---|---:|---:|
| Total | 8759 (4.29) | 12544 (4.16) |
| *wh*-Questions | 1507 (3.72) | 2287 (4.80) |
| Other Questions | 2427 (4.40) | 3568 (4.34) |
| Declaratives | 4825 (4.41) | 6689 (3.85) |

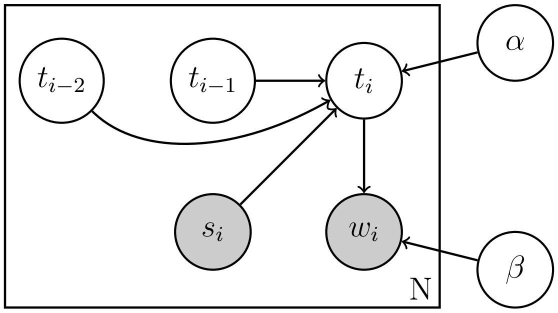

*Figure 1.* Graphical model representation of the BHMM-T, which includes sentence type as an observed variable on tag transitions (but not emissions).

$$P_{BHMM}(t|\mathbf{t}_{-i},\mathbf{w},\alpha,\beta_t) \propto \frac{n_{t,w_i}+\beta_t}{n_t+V\beta_t} \times \frac{n_{t_{i-2},t_{i-1},t}+\alpha}{n_{t_{i-2},t_{i-1}}+T\alpha} \tag{9}$$

$$P_{BHMM-E}(t|\mathbf{t}_{-i},\mathbf{w},\mathbf{s},\alpha,\beta_t) \propto \frac{n_{s_i,t,w_i,}+\beta_t}{n_{s_i,t}+V\beta_t} \times \frac{n_{t_{i-2},t_{i-1},t}+\alpha}{n_{t_{i-2},t_{i-1}}+T\alpha} \tag{10}$$

$$P_{BHMM-T}(t|\mathbf{t}_{-i},\mathbf{w},\mathbf{s},\alpha,\beta_t) \propto \frac{n_{t,w_i}+\beta_t}{n_t+V\beta_t} \times \frac{n_{t_{i-2},t_i,s_{i-1},t_i}+\alpha}{n_{s_{i-1},t_{i-2},t_{i-1}}+T\alpha} \tag{11}$$

$$P_{BHMM-ET}(t|\mathbf{t}_{-i},\mathbf{w},\mathbf{s},\alpha,\beta_t) \propto \frac{n_{s_i,t,w_i,}+\beta_{t_i}}{n_{s_i,t}+V\beta_t} \times \frac{n_{t_{i-2},t_{i-1},s_{i-1},t}+\alpha}{n_{s_{i-1},t_{i-2},t_{i-1}}+T\alpha} \tag{12}$$

*Figure 2*. Trigram Gibbs sampling equations for the four model variants. $n_{t,w_i}$ is the number of occurrences of tag $t$ with word $w_i$ (and analogously for trigram transition counts); $\mathbf{t}_{-i}$ indicates that these counts do not include the current value of $t_i$. Transition factors for trigrams $(t_{i-1},t,t_{i+1})$ and $(t,t_{i+1},t_{i+2})$ are not shown but must be included (see Goldwater and Griffiths (2007)).

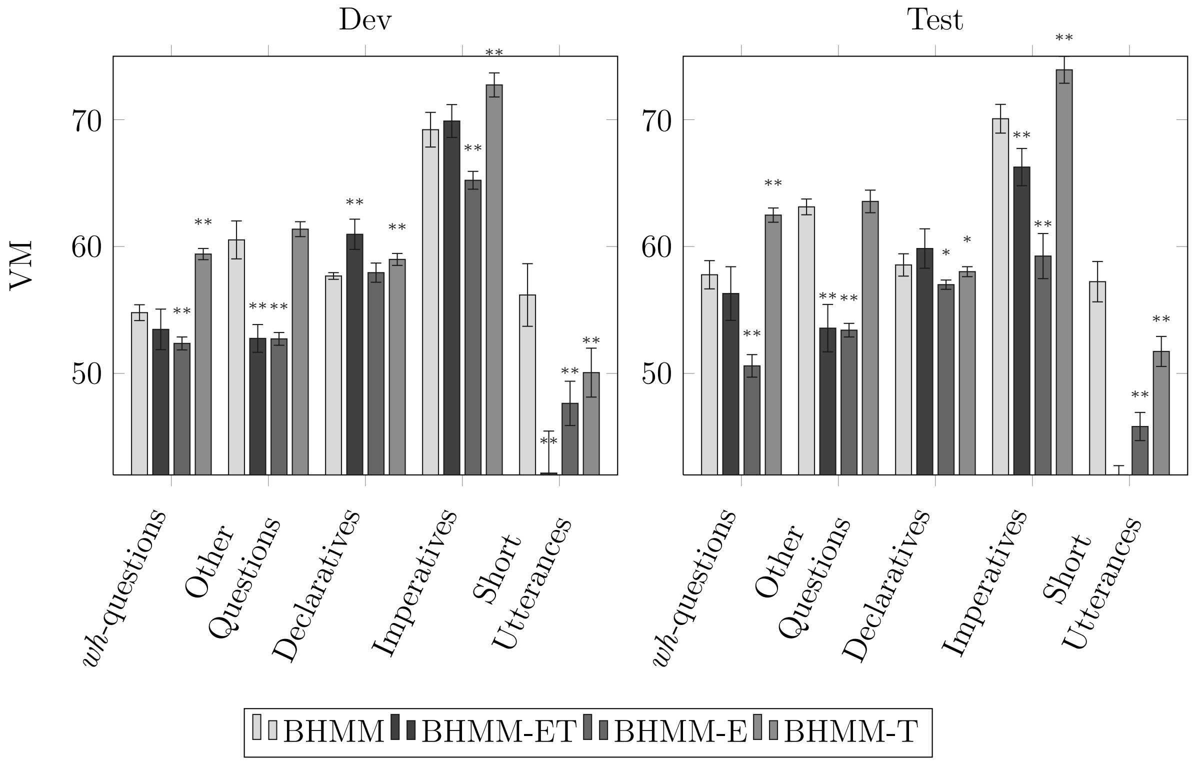

*Figure 3*. Performance of bigram BHMM and variants by sentence type on the Eve corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

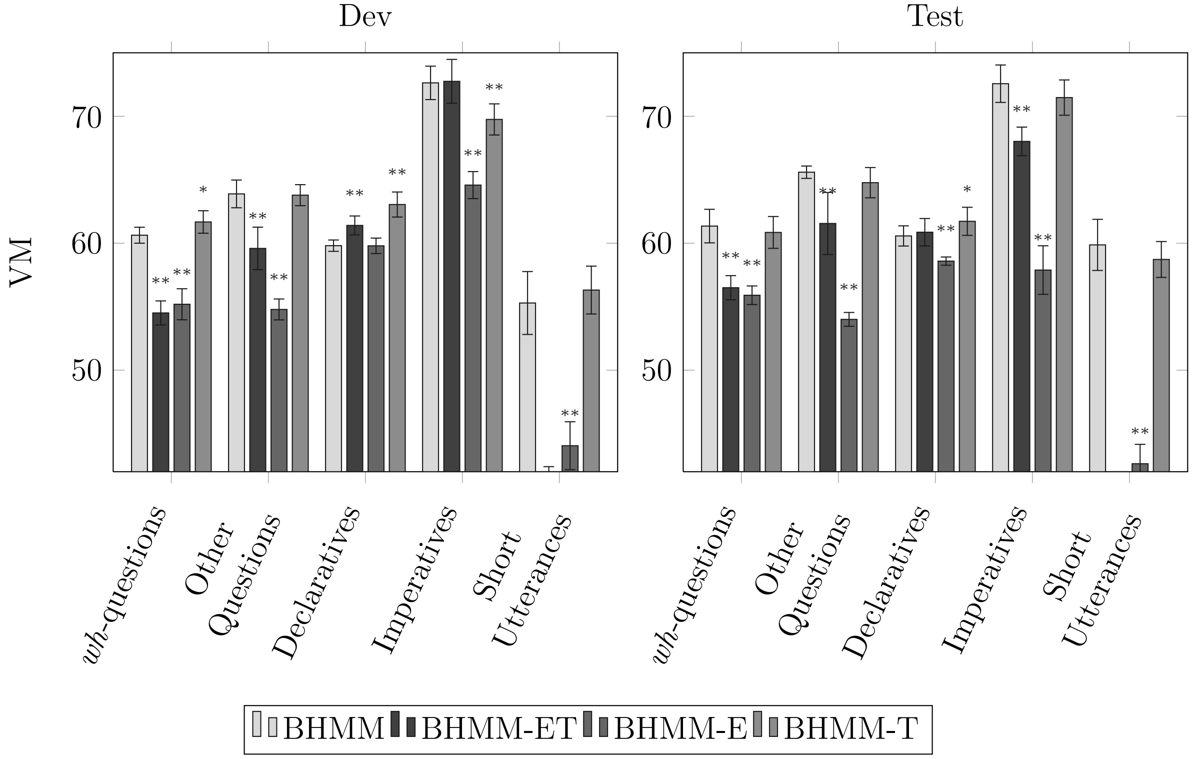

*Figure 4.* Performance of trigram BHMM and variants by sentence type on the Eve corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

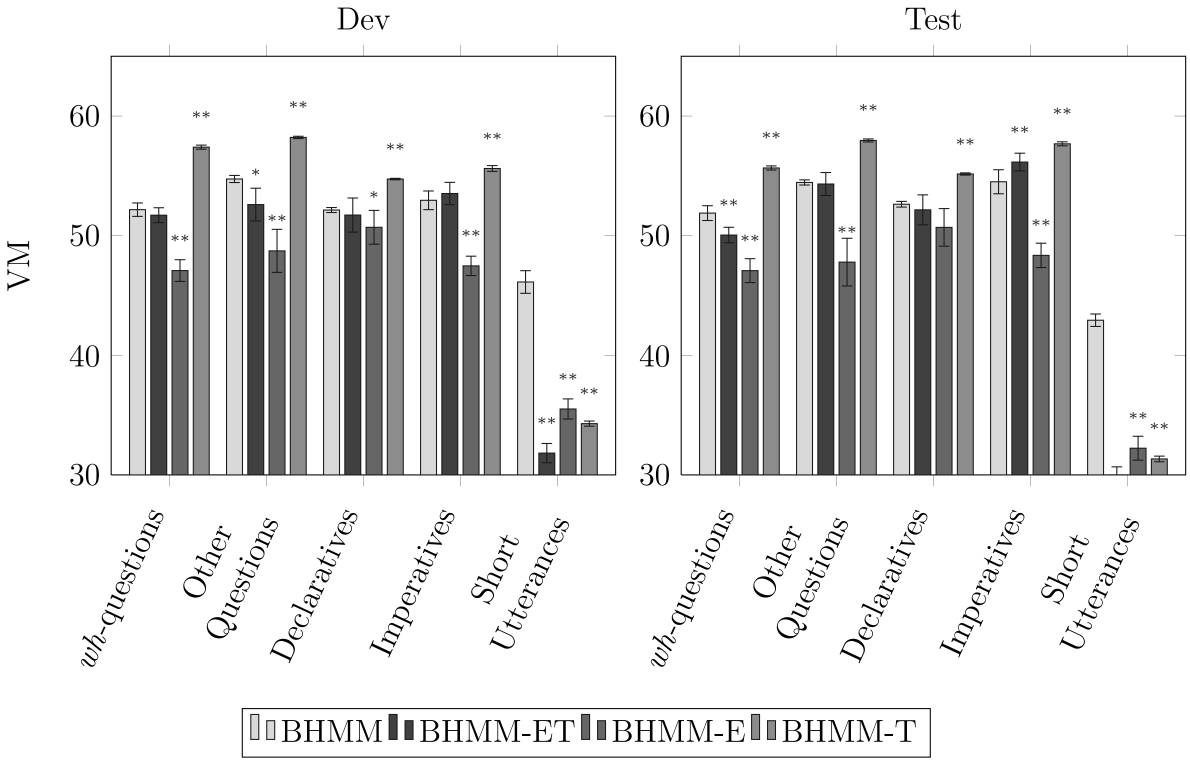

*Figure 5*. Performance of bigram BHMM and variants by sentence type on the Manchester corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

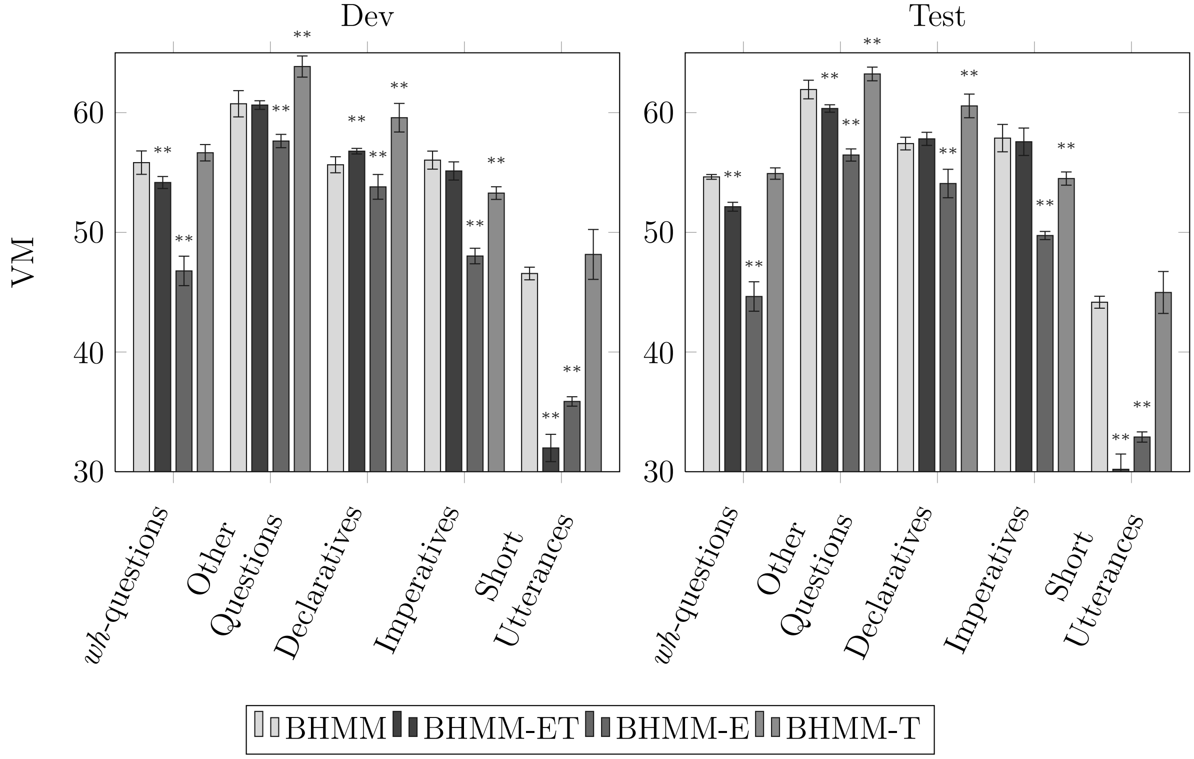

*Figure 6*. Performance of trigram BHMM and variants by sentence type on the Manchester corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

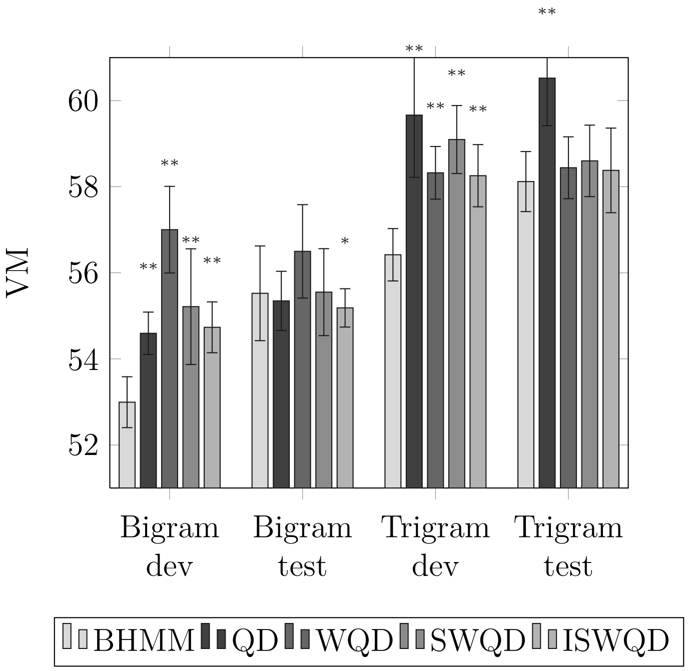

*Figure 7*. BHMM-T performance on the Eve corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

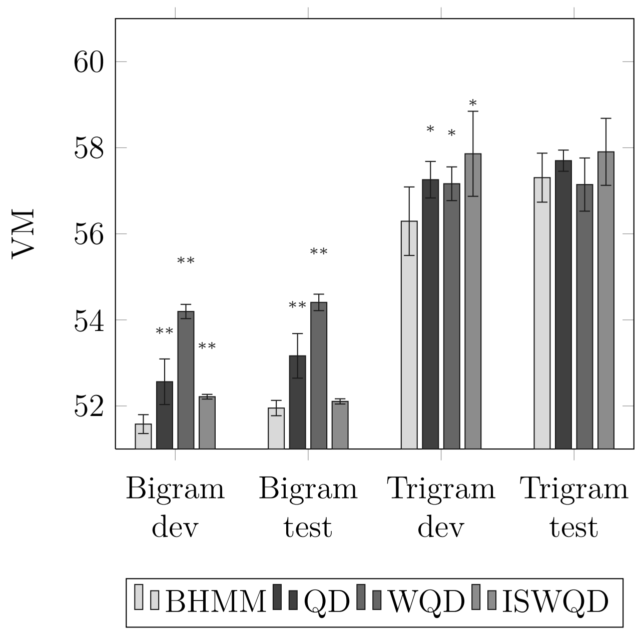

*Figure 8*. BHMM-T performance on the Manchester corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

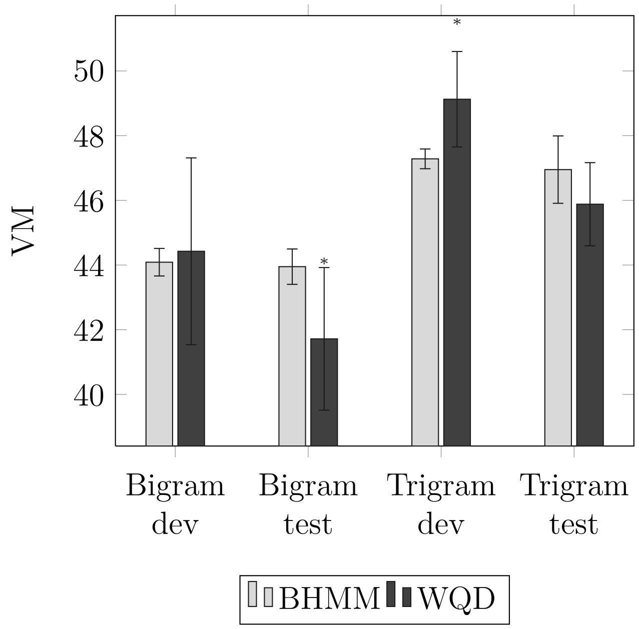

*Figure 9*. BHMM-T performance on the Spanish (Ornat) corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).

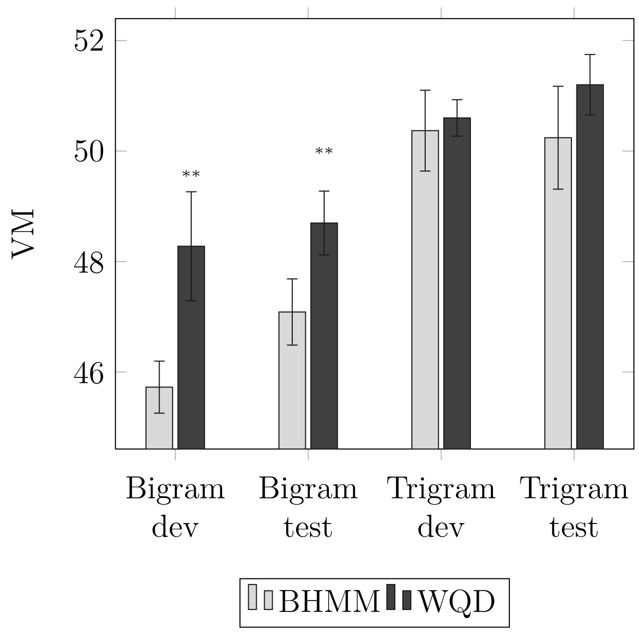

*Figure 10*. BHMM-T performance on the Cantonese (LWL) corpus: Mean VM and 95% Confidence Intervals ($N = 10$). Values that differ significantly from baseline (BHMM) are marked with $*$ ($p < 0.05$, two-tailed) or $**$ ($p < 0.01$, two-tailed).