Using Sentence Type Information for Syntactic Category Acquisition

Stella Frank (s.c.frank@sms.ed.ac.uk)
Sharon Goldwater (sgwater@inf.ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

In this paper we investigate a new source of information for syntactic category acquisition: sentence type (question, declarative, imperative). Sentence type correlates strongly with intonation patterns in most languages; we hypothesize that these intonation patterns are a valuable signal to a language learner, indicating different syntactic patterns. To test this hypothesis, we train a Bayesian Hidden Markov Model (and variants) on child-directed speech. We first show that simply training a separate model for each sentence type decreases performance due to sparse data. As an alternative, we propose two new models based on the BHMM in which sentence type is an observed variable which influences either emission or transition probabilities. Both models outperform a standard BHMM on data from English, Cantonese, and Dutch. This suggests that sentence type information available from intonational cues may be helpful for syntactic acquisition cross-linguistically.

1 Introduction

Children acquiring the syntax of their native language have access to a large amount of contextual information. Acquisition happens on the basis of speech, and the acoustic signal carries rich prosodic and intonational information that children can exploit. A key task is to separate the acoustic properties of a word from the underlying sentence intonation. Infants become attuned to the pragmatic and discourse functions of utterances as signalled by intonation extremely early; in this they are helped by the fact that intonation contours of child and infant directed speech are especially well differentiated between sentence types (Stern et al., 1982; Fernald, 1989). Children learn to use appropriate intonational melodies to communicate their own intentions at the one word stage, before overt syntax develops (Snow and Balog, 2002).

It follows that sentence type information (whether a sentence is declarative, imperative, or a question), as signaled by intonation, is readily available to children by the time they start to acquire syntactic categories. Sentence type also has an effect on sentence structure in many languages (most notably on word order), so we hypothesise that sentence type is a useful cue for syntactic category learning. We test this hypothesis by incorporating sentence type information into an unsupervised model of part of speech tagging.

We are unaware of previous work investigating the usefulness of this kind of information for syntactic category acquisition. In other domains, intonation has been used to identify sentence types as a means of improving speech recognition language models. Specifically, (Taylor et al., 1998) found that using intonation to recognize dialogue acts (which to a significant extent correspond to sentence types) and then using a specialized language model for each type of dialogue act led to a significant decrease in word error rate.

In the remainder of this paper, we first present the Bayesian Hidden Markov Model (BHMM; Goldwater and Griffiths (2007)) that is used as the baseline model of category acquisition, as well as our extensions to the model, which incorporate sentence type information. We then discuss the distinctions in sentence type that we used and our evaluation measures, and finally our experimental results. We perform experiments on corpora in four different languages: English, Spanish, Cantonese, and Dutch. Our results on Spanish show no difference between the baseline and the models incorporating sentence type, possibly due to the small size of the Spanish corpus. Results on all other corpora show a small improvement in performance when sentence type is included as a cue to the learner. These cross-linguistic results suggest that sentence type may be a useful source of information to children acquiring syntactic categories.

2 BHMM Models

2.1 Standard BHMM

We use a Bayesian HMM (Goldwater and Griffiths, 2007) as our baseline model. Like a standard trigram HMM, the BHMM assumes that the probability of tag t_i depends only on the previous two tags, and the probability of word w_i depends only on t_i. This can be written as

$$t_i | t_{i-1} = t, t_{i-2} = t', \tau^{(t,t')} \sim \text{Mult}(\tau^{(t,t')}) \quad (1)$$

$$w_i | t_i = t, \omega^{(t)} \sim \text{Mult}(\omega^{(t)}) \quad (2)$$

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 BHMM assumes that these parameters are in turn drawn from symmetric Dirichlet priors with parameters α and β, respectively:

$$\tau^{(t,t')}|\alpha \sim \text{Dirichlet}(\alpha) \quad (3)$$

$$\omega^{(t)}|\beta \sim \text{Dirichlet}(\beta) \quad (4)$$

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

$$P(t_i|\mathbf{t}_{-i},\alpha) = \frac{C(t_{i-2},t_{i-1},t_i)+\alpha}{C(t_{i-2},t_{i-1})+T\alpha} \quad (5)$$

$$P(w_i|t_i,\mathbf{t}_{-i},\mathbf{w}_{-i},\beta) = \frac{C(t_i,w_i)+\beta}{C(t_i)+W_{t_i}\beta} \quad (6)$$

where $\mathbf{t}_{-i} = t_1 \ldots t_{i-1}$, $\mathbf{w}_{-i} = w_1 \ldots w_{i-1}$, $C(t_{i-2},t_{i-1},t_i)$ and $C(t_i,w_i)$ are the counts 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_i} is the number of word types emitted by t_i.

Based on these predictive distributions, (Goldwater and Griffiths, 2007) develop a Gibbs sampler for the model, which samples from the posterior distribution over tag sequences $\mathbf{t}$ given word sequences $\mathbf{w}$, i.e., $P(\mathbf{t}|\mathbf{w},\alpha,\beta) \propto P(\mathbf{w}|\mathbf{t},\beta)P(\mathbf{t}|\alpha)$. This is done by using Equations 5 and 6 to iteratively resample each tag t_i given the current values of all other tags.[1] The results show that the BHMM with Gibbs sampling performs better than the standard HMM using expectation-maximization. In particular, the Dirichlet priors in the BHMM constrain the model towards sparse solutions, i.e., solutions in which each tag emits a relatively small number of words, and in which a tag transitions to few following tags. This type of model constraint allows the model to find solutions which correspond to true syntactic parts of speech (which follow such a sparse, Zipfian distribution), unlike the uniformly-sized clusters found by standard maximum likelihood estimation using EM.

In the experiments reported below, we use the Gibbs sampler described by (Goldwater and Griffiths, 2007) for the BHMM, and modify it as necessary for our own extended models. We also follow (Goldwater and Griffiths, 2007) in using Metropolis-Hastings sampling for the hyperparameters, which are inferred automatically in all experiments. A separate β parameter is inferred for each tag.

2.2 BHMM with Sentence Types

We wish to add a sentence type feature to each timestep in the HMM, signalling the current sentence type. We treat sentence type (s) as an observed variable, on the assumption that it is observed via intonation or punctuation features (not part of our model), and these features are informative enough to reliably distinguish sentence types (as speech recognition tasks have found to be the case, see Section 1).

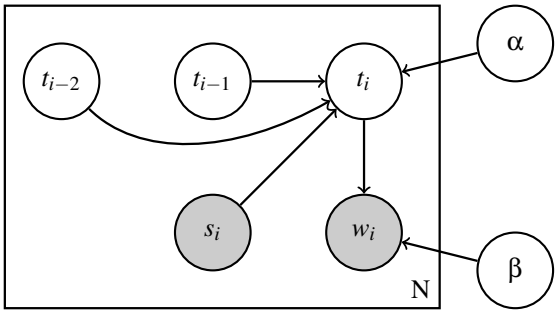

Figure 1: Graphical model representation of the BHMM-T, which includes sentence type as an observed variable on tag transitions (but not emissions).

In the BHMM, there are two obvious ways that sentence type could be incorporated into the generative model: either by affecting the transition probabilities or by affecting the emission probabilities. The first case can be modeled by adding s_i as a conditioning variable when choosing t_i, replacing line 1 from the BHMM definition with the following:

$$t_i|s_i = s, t_{i-1} = t, t_{i-2} = t', \tau^{(t,t')} \sim \text{Mult}(\tau^{(s,t,t')}) \quad (7)$$

We will refer to this model, illustrated graphically in Figure 1, as the BHMM-T. It assumes that the distribution over t_i depends not only on the previous two tags, but also on the sentence type, i.e., that different sentence types tend to have different sequences of tags.

In contrast, we can add s_i as a conditioning variable for w_i by replacing line 2 from the BHMM with

$$w_i|s_i = s, t_i = t, \omega^{(t)} \sim \text{Mult}(\omega^{(s,t)}) \quad (8)$$

This model, the BHMM-E, assumes that different sentence types tend to have different words emitted from the same tag.

The predictive distributions for these models are given in Equations 9 (BHMM-T) and 10 (BHMM-E):

$$P(t_i|t_{-i},s_i,\alpha) = \frac{C(t_{i-2},t_{i-1},t_i,s_i)+\alpha}{C(t_{i-2},t_{i-1},s_i)+T\alpha} \quad (9)$$

$$P(w_i|t_i,s_i,\beta) = \frac{C(t_i,w_i,s_i)+\beta}{C(t_i,s_i)+W_{t_i}\beta} \quad (10)$$

Of course, we can also create a third new model, the BHMM-B, in which sentence type is used to condition both transition and emission probabilities. This model is equivalent to training a separate BHMM on each type of sentence (with shared hyperparameters). Note that introducing the 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 actually degrade performance if sentence type is not a useful variable.

[1] Slight corrections need to be made to Equation 5 to account for sampling tags from the middle of the sequence rather than from the end; these are given in (Goldwater and Griffiths, 2007) and are followed in our own samplers.

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 → PRO 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 PRO 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 principally constrained by sentences type (especially in the short and grammatically simple utterances found in child-directed speech).

3 Sentence Types

We experiment with a number of sentence-type categories, leading to increasingly fine grained distinctions.

The primary distinction is between *questions* (Q) and *declaratives* (D). Questions are marked by punctuation (in writing) or by intonation (in speech), as well as by word order or other morpho-syntactic markers in many languages.

Questions may be separated into categories, most notably *wh-questions* and yes/no-questions. Many languages (including several English dialects) have distinct intonation patterns for *wh-* and yes/no-questions (Hirst and Cristo, 1998).

Imperatives are a separate type from declaratives, with distinct word order and intonation patterns.

Declaratives may be further subdivided into *fragments* and full sentences. We define fragments as utterances without a verb (including auxiliary verbs).

As an alternate sentence-level feature to sentence type, we use length. Utterances are classified according to their length, as either shorter or longer than average. Shorter utterances are more likely to be fragments and may have distinct syntactic patterns. However these patterns are likely to be less strong than in the above type-based types. In effect this condition is a pseudo-baseline, testing the effects of less- or non-informative sentence features on our proposed models.

4 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. We report three evaluation measures in our experiments, in order to avoid the weaknesses inherent in any single measure and in an effort to be comparable to previous work.

Matched accuracy (MA), also called many-to-one accuracy, is a commonly used measure for evaluating unlabeled clusterings in part of speech tagging. Each unlabeled cluster is given the label of the gold category with which it shares the most members. Given these labels, accuracy can be measured as usual, as the percentage of tokens correctly labeled. Note that multiple clusters may have the same label if several clusters match the same gold standard category. This can lead to a degenerate solution if the model is allowed an unbounded number of categories, in which each word is in a separate cluster. In less extreme cases, it makes comparing MA across clustering results with different numbers of clusters difficult. Another serious issue with MA is the "problem of matching" (Meila, 2007): matched accuracy only evaluates whether or not the items in the cluster match the majority class label. The non-matching items within a cluster might all be from a second gold class, or they might be from many different classes. Intuitively, the former clustering should be evaluated as better, but matched accuracy is the same for both clusterings.

Variation of Information (VI) (Meila, 2007) is a clustering evaluation measure that avoids the matching problem. It measures the amount of information lost and gained when moving between two clusterings. More precisely:

$$\begin{aligned} VI(C,K) &= H(C) + H(K) - 2I(C,K) \\ &= H(C|K) + H(K|C) \end{aligned}$$

A lower score implies closer clusterings, since each clustering has less information not shared with the other: two identical clusterings have a VI of zero. However, VI's upper bound is dependent on the maximum number of clusters in C or K, making it difficult to compare clustering results with different numbers of clusters.

As a third, and, in our view, most informative measure, we use *V-measure* (VM; Rosenberg and Hirschberg (2007)). Like VI, VM uses the conditional entropy of clusters and categories to evaluate clusterings. However, it also has the useful characteristic of being analogous to the precision and recall measures commonly used in NLP. Homogeneity, the precision analogue, is defined as

$$VH = 1 - \frac{H(C|K)}{H(C)}.$$

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 is low. Completeness (recall) is defined symmetrically to VH as:

$$VC = 1 - \frac{H(K|C)}{H(K)}.$$

VC measures the conditional entropy of the clusters within each gold standard category, and is highest if each category maps to a single cluster so that each

		Eve		Manchester	
Sentence type		Counts	$\|w\|$	Counts	$\|w\|$
Total		13494	4.39	13216	4.23
D	Total	8994	4.48	8315	3.52
	I	623	4.87	757	4.22
	F	2996	1.73	4146	1.51
Q	Total	4500	4.22	4901	5.44
	wh	2105	4.02	1578	4.64
	Short utts	5684	1.89	6486	1.74
	Long utts	7810	6.21	6730	6.64

Table 1: Counts of sentence types in the Eve and Manchester training set. (Test and dev sets are approximately 10% of the size of training.) $|w|$ is the average length in words of utterances of this type. D: declaratives, I: imperatives, F: fragments, Q: questions, wh: *wh*-questions.

model cluster completely contains a category. The V-measure VM is simply the harmonic mean of VH and VC, analogous to traditional F-score. Unlike MA and VI, VM is invariant with regards to both the number of items in the dataset and to the number of clusters used, and consequently it is best suited for comparing results across different corpora.

5 English experiments

5.1 Corpora

We use the Eve corpus (Brown, 1973) and the Manchester corpus (Theakston et al., 2001) from the CHILDES collection (MacWhinney, 2000). The Eve corpus is a longitudinal study of a single US American child from the age of 1.5 to 2.25 years, whereas the Manchester corpus follows a cohort of 12 British children from the ages of 2 to 3. Using both corpora ensures that any effect is not due to a particular child, and is not specific to a type of English.

From both corpora we remove all utterances spoken by a child; the remaining utterances are nearly exclusively child-directed speech (CDS). We use the full Eve corpus and a similarly sized subset of the Manchester corpus, consisting of the first 70 CDS utterances from each file. Files from the chronological middle of each corpus are set aside for development and testing (Eve: file 10 for testing, 11 for dev; Manchester: file 17 from each child for testing, file 16 for dev).

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*, or pronouns (*who* and the rest).

Table 1 show the sizes of the training sets, and the breakdown of sentence types within them. Each sentence type can be identified using a distinguishing characteristic. Sentence-final punctuation is used to differentiate between questions and declaratives; *wh*-questions are then further differentiated by the presence of a *wh*-word. Imperatives are separated from the declaratives by a heuristic (since CHILDES does not have an imperative verb tag): if an utterance includes a base verb within the first two words, without a pronoun proceeding it (with the exception of *you*, as in *you sit down right now*), the utterance is coded as an imperative. Fragments are also identified using the tag annotations, namely by the lack of a verb or auxiliary tag in an utterance.

The CHILDES annotation guide specifies that the question mark is to be used with any utterance with "final rising contour", even if syntactically the utterance might appear to be a declarative or exclamation. The question category consequently includes echo questions (*Finger stuck?*) and non-inverted questions (*You want me to have it?*).

5.2 Inference and Evaluation Procedure

Unsupervised models do not suffer from overfitting, so generally it is thought unnecessary to use separate training and testing data, with results being reported on the entire set of input data. However, there is still a danger, in the course of developing a model, of overfitting in the sense of becoming too finely attuned to a particular set of input data. To avoid this, we use separate test and development sets. The BHMM is trained on (train+dev) or (train+test), but evaluation scores are computed based on the dev or test portions of the data only. [2]

We run the Gibbs sampler for 2000 iterations, with hyperparameter resampling and simulated annealing. Each iteration produces an assignment of tags to the tokens in the corpus; the final iteration is used for evaluation purposes. Since Gibbs sampling is a stochastic algorithm, we run all models multiple times (three, except where stated otherwise) and report average values for all evaluation measures, as well as confidence intervals. We run our experiments using a variety of sentence type features, ranging from the coarse question/declarative (Q/D) distinction to the full five types. For reasons of space we do not report all results here, instead confining ourselves to representative samples.

5.3 BHMM-B: Type-specific Sub-Models

When separate sub-models are used for each sentence type, as in the BHMM-B, where both transition and emission probabilities are conditioned on sentence type, the hidden states (tags) in each sub-model do not correspond to each other, e.g., a hidden state 9 in one sub-model is not the same state 9 in another sub-model. Consequently, when evaluating the tagged output, each sentence type must be evaluated separately (otherwise the evaluation would equate declaratives-tag-9 with questions-tag-9).

[2]The results presented in this paper are all evaluated on the dev set; preliminary test set results on the Eve corpus show the same patterns.

Model	VM	VC	VH	VI	MA
wh-questions					
BHMM:	63.0 (1.0)	59.8 (0.4)	66.6 (1.8)	1.63 (0.03)	70.7 (2.7)
BHMM-B:	58.7 (2.0)	58.2 (2.1)	59.2 (2.0)	1.74 (0.09)	59.7 (2.0)
Other Questions					
BHMM:	65.6 (1.4)	62.7 (1.3)	68.7 (1.5)	1.62 (0.06)	74.5 (0.5)
BHMM-B:	64.4 (3.6)	62.6 (4.4)	66.2 (2.8)	1.65 (0.19)	70.8 (2.5)
Declaratives					
BHMM:	60.9 (1.3)	58.7 (1.1)	63.3 (1.6)	1.84 (0.06)	73.5 (0.8)
BHMM-B:	58.0 (1.2)	55.5 (1.1)	60.7 (1.5)	1.99 (0.06)	69.0 (1.5)

Table 2: Results for BHMM-B on W/Q/D sentence types (dev set evaluation) in the Manchester corpus, compared to the standard BHMM. The confidence interval is indicated in parentheses. Note that lower VI is better.

Model	VM	VC	VH	VI	MA
BHMM:	59.4 (0.2)	56.9 (0.2)	62.3 (0.2)	1.96 (0.01)	72.2 (0.2)
Q/D:	61.2 (1.2)	58.6 (1.2)	64.0 (1.4)	1.88 (0.06)	72.1 (1.5)
W/Q/D:	61.0 (1.7)	59.0 (1.5)	63.0 (2.0)	1.86 (0.08)	69.6 (2.2)
F/I/D/Q/W:	61.7 (1.7)	58.9 (1.8)	64.8 (1.6)	1.80 (0.09)	70.5 (1.3)

Table 3: Results for BHMM-E on the Eve corpus (dev set evaluation), compared to the standard BHMM. The confidence interval is indicated in parentheses.

Table 2 shows representative results for the W/Q/D condition on the Manchester corpus, separated into *wh*-questions, other questions, and declaratives. For each sentence type, the BHMM-B performs significantly worse than the BHMM. The *wh*-questions sub-model, which is trained on the smallest subset of the input corpus, performs the worst across all measures except VI. This suggests that lack of data is why these sub-models perform worse than the standard model.

5.4 BHMM-E: Type-specific Emissions

Having demonstrated that using entirely separate sub-models does not improve tagging performance, we turn to the BHMM-E, in which emission probability distributions are sentence-type specific, but transition probabilities are shared between all sentence types.

The results in Table 3 show that BHMM-E does result in slightly better tagging performance as evaluated by VI (lower VI is better) and VM and its components. Matched accuracy does not capture this same trend. Inspecting the clusters found by the model, we find that clusters for the most part do match gold categories. The tokens that do not fall into the highest matching gold categories are not distributed randomly, however; for instance, nouns and pronouns often end up in the same cluster. VI and VM capture these secondary matches while MA does not. Some small gold categories (e.g. the single word infinitival-*to* and negation-*not* categories) are often merged by the model into a single cluster, with the result that MA considers nearly half the cluster as misclassified.

The largest increase in performance with regards to the standard BHMM is obtained by adding the distinction between declaratives and questions. Thereafter, adding the *wh*-question, fragment and imperative sentence types does not worsen performance, but also does not significantly improve performance on any measure.

5.5 BHMM-T: Type-specific Transitions

Lastly, the BHMM-T shares emission probabilities among sentence types and uses sentence type specific transition probabilities.

Results comparing the standard BHMM with the BHMM-T with sentence-type-specific transition probabilities are presented in Table 4. Once again, VM and VI show a clear trend: the models using sentence type information outperform both the standard BHMM and models splitting according to utterance length (shorter/longer than average). MA shows no significant difference in performance between the different models (aside from clearly showing that utterance length is an unhelpful feature). The fact that the MA measure shows no clear change in performance is likely a fault of the measure itself; as explained above, VI and VM take into account the distribution of words within a category, which MA does not.

As with the BHMM-E, the improvements to VM and VI are obtained by distinguishing between questions and declaratives, and then between *wh*- and other questions. Both of these distinctions are marked by intonation in English. In contrast, distinguishing fragments and imperatives, which are less easily detected by intonation, provides no obvious benefit in any case. Using sentence length as a feature degrades performance considerably, confirming that improvements in performance are due to sentence types capturing useful information about the tagging task, and not simply due to splitting the input in some arbitrary way.

One reason for the improvement when adding the *wh*-question type is that the models are learning to identify and cluster the *wh*-words in particular. If we evaluate the *wh*-words separately, VM rises from 32.3

Model	VM	VC	VH	VI	MA
Eve					
BHMM:	59.4 (0.2)	56.9 (0.2)	62.3 (0.2)	1.96 (0.01)	72.2 (0.2)
Q/D:	60.9 (0.5)	58.3 (0.4)	63.7 (0.6)	1.89 (0.02)	72.7 (0.3)
W/Q/D:	**62.5** (1.2)	60.0 (1.3)	65.2 (1.0)	1.81 (0.06)	72.9 (0.8)
F/I/D/Q/W:	62.2 (1.5)	59.5 (1.6)	65.2 (1.3)	1.77 (0.08)	71.5 (1.4)
Length:	57.9 (1.2)	55.3 (1.1)	60.7 (1.3)	2.04 (0.06)	69.7 (2.0)
Manchester					
BHMM:	60.2 (0.9)	57.6 (0.9)	63.1 (1.0)	1.92 (0.05)	72.1 (0.7)
Q/D:	61.5 (0.7)	59.2 (0.6)	63.9 (0.9)	1.84 (0.03)	71.6 (1.5)
W/Q/D:	**62.7** (0.2)	60.6 (0.2)	65.0 (0.3)	1.78 (0.01)	71.2 (0.6)
F/I/D/Q/W:	62.5 (0.4)	60.3 (0.5)	64.9 (0.4)	1.79 (0.02)	71.3 (0.9)
Length:	58.1 (0.7)	55.6 (0.8)	60.8 (0.6)	2.02 (0.04)	71.0 (0.6)

Table 4: Results on the Eve and Manchester corpora for the various sentence types in the BHMM and BHMM-T models. The confidence interval is indicated in parentheses.

in the baseline BHMM to 41.5 in the W/Q/D condition with the BHMM-T model and 46.8 with the BHMM-E model. Performance for the non-*wh*-words also improves in the W/Q/D condition, albeit less dramatically: from 61.1 in the baseline BHMM to 63.6 with BHMM-T and 62.0 with BHMM-E. The *wh*-question type enables the models to pick up on the defining characteristics of these sentences, namely *wh*-words.

We predicted the sentence-type specific transition probabilities in the BHMM-T to be more useful than the sentence-type specific emission probabilities in the BHMM-E. The BHMM-T does perform slightly better than the BHMM-E, however, the effect is small. Word or tag order may be the most overt difference between questions and declaratives in English, but word choice, especially the use of *wh*-words varies sufficiently between sentence types for sentence-type specific emission probabilities to be equally useful.

6 Crosslinguistic Experiments

In the previous section we found that sentence type information improved syntactic categorisation in English. In this section, we evaluate the BHMM's performance on a range of languages other than English, and investigate whether sentence type information is useful across languages. To our knowledge this is the first application of the BHMM to non-English data.

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

The CHILDES collection includes tagged corpora for Spanish and Cantonese: the Ornat corpus (Ornat, 1994) and the Lee Wong Leung (LWL) corpus (Lee et al., 1994) respectively. To cover a greater variety of word order patterns, a Dutch corpus of adult dialogue (not CDS) is also tested. We describe each corpus in turn below; Table 5 describes their relative sizes.

	Total	Ds	all Qs	*wh*-Qs
Spanish	8759	4825	3934	1507
$\|w\|$	4.29	4.41	4.14	3.72
Cantonese	12544	6689	5855	2287
$\|w\|$	4.16	3.85	4.52	4.80
Dutch	8967	7812	1155	363
$\|w\|$	6.16	6.19	6.00	7.08

Table 5: Counts of sentence types in the training sets for Spanish. Cantonese and Dutch. (Test and dev sets are approximately 10% of the size of training.) $|w|$ is the average length in words of utterances of this type. D: declaratives, Qs: questions, *wh*-Qs: *wh*-questions.

6.1 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 08/09 are used testing/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).

6.2 Cantonese

The LWL corpus consists of transcripts from a set of children followed over the course of a year, totalling 128 files. The ages of the children are not matched, but they range between one and three years old. Our training set consists of the first 500 utterances of all training files, 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 are the development set files (again, the first 500 utterances from each of these make up the test/dev corpus).

The tagset used in the LWL is larger than the English corpus. It consists of 20 tags: adjective, adverb, aspectual 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. Despite the use of tones in Cantonese, questions are marked with rising final intonation.

6.3 Dutch

The Corpus of Spoken Dutch (CGN) contains Dutch spoken in a variety of settings. We use the "spontaneous conversation" component, consisting of 925 files, since it is the most similar to CDS. However, the utterances are longer, and there are far fewer questions, especially *wh*-questions (see Table 5).

The corpus does not have any meaningful timeline, so we designated all files with numbers ending in 0 as test files and files ending in 9 as dev files. The first 60 utterances from each file were used, to create training/test/dev sets similar in size to the other corpora.

The coarse CGN tagset consists of 11 tags, which we used directly: adjective, adverb, conjunction, determiner, interjection, noun, number, pronoun/determiner, preposition, and verb.

Dutch follows verb-second word order in main clauses and SOV word order in embedded clauses. Yes/no-questions are created by verb-fronting, as in Spanish. *wh*-questions involve a *wh*-word at the beginning of the utterance followed by the verb in second position.

6.4 Results

We trained standard BHMM, BHMM-T and BHMM-E models in the same manner as with the English corpora. Given the poor performance of the BHMM-B, we did not test it here.

Due to inconsistent annotation and lack of familiarity with the languages we tested only two sentence type distinctions, Q/D and W/Q/D. Punctuation was used to distinguish between questions and declaratives. *wh*-questions were identified by using a list of *wh*-words for Spanish and Dutch; for Cantonese we relied on the *wh*-word tag annotation.

Results are shown in Table 6. Since the corpora are different sizes and use tagsets of varying sizes, VI and MA results are not comparable between corpora. VM (and VC and VH) are more robust, but even so cross-corpora comparisons should be made carefully. The English corpora VM scores are significantly higher (around 10 points higher) than the non-English corpora scores.

In Cantonese and Dutch, the W/Q/D BHMM-T model performs best; in both cases significantly better than the BHMM. In Cantonese, the separation of *wh*-questions improves tagging significantly in both the BHMM-T and BHMM-E models, whereas simply separating questions and declaratives helps far less. In the Dutch corpus, *wh*-questions improved only in the BHMM-T, not in the BHMM-E.

The Spanish models have higher variance, due to the small size of the corpus. Due to the high variance, there are no significant differences between any of the conditions; it is also difficult to spot a trend.

7 Future Work

We have shown sentence type information to be useful for syntactic tagging. However, the BHMM-E and BHMM-T models are successful in part however because they also share information as well as split it; the completely split BHMM-B does not perform well. Many aspects of tagging do not change significantly between sentence types. Within a noun phrase, the ordering of determiners and nouns is the same whether it is in a question or an imperative, and to a large extent the determiners and nouns used will be the same as well. Learning these patterns over as much input as possible is essential. Therefore, the next step in this line of work will be to add a general (corpus-level) model alongside type-specific models. Ideally, the model will learn when to use the type-specific model (when tagging the beginning of questions, for instance) and when to use the general model (when tagging NPs). Such a model would make use of sentence-type information in a better way, hopefully leading to further improvements in performance. A further, more sophisticated model could learn the useful sentence types distinctions automatically, perhaps forgoing the poorly performing imperative or fragment types we tested here in favor of a more useful type we did not identify.

8 Conclusions

We set out to investigate a hitherto unused source of information for models of syntactic category learning, namely intonation and its correlate, sentence type. We showed that this information is in fact useful, and including it in a Bayesian Hidden Markov Model improved unsupervised tagging performance.

We found tagging performance increases if sentence type information is used to generate either transition probabilities or emission probabilities in the BHMM. However, we found that performance decreases if sentence type information is used to generate both transi-

Model	VM	VC	VH	VI	MA
Spanish					
BHMM:	49.4 (1.8)	47.2 (1.9)	51.8 (1.8)	2.27 (0.09)	61.5 (2.1)
BHMM-E Q/D:	49.4 (1.5)	47.0 (1.4)	52.1 (1.7)	2.28 (0.06)	60.9 (2.6)
BHMM-E W/Q/D:	48.7 (2.5)	46.4 (2.4)	51.2 (2.6)	2.31 (0.11)	60.2 (3.0)
BHMM-T Q/D:	49.0 (1.7)	46.7 (1.6)	51.6 (1.7)	2.30 (0.07)	60.9 (2.5)
BHMM-T W/Q/D:	49.5 (2.5)	47.2 (2.3)	52.1 (2.8)	2.27 (0.11)	61.0 (3.0)
Cantonese					
BHMM:	49.4 (0.8)	44.5 (0.7)	55.4 (1.0)	2.60 (0.04)	67.2 (1.0)
BHMM-E Q/D:	50.7 (1.6)	45.4 (1.5)	57.5 (1.7)	2.55 (0.09)	69.0 (1.0)
BHMM-E W/Q/D:	**52.3** (0.3)	46.9 (0.3)	59.3 (0.4)	2.46 (0.02)	69.4 (0.9)
BHMM-T Q/D:	50.3 (0.9)	45.0 (0.9)	57.0 (1.0)	2.57 (0.05)	68.4 (0.8)
BHMM-T W/Q/D:	**52.2** (0.8)	46.8 (0.9)	59.1 (0.7)	2.47 (0.05)	69.9 (1.9)
Dutch					
BHMM:	48.4 (0.7)	47.1 (0.8)	49.7 (0.7)	2.38 (0.04)	62.3 (0.3)
BHMM-E Q/D:	48.4 (0.4)	47.3 (0.4)	49.7 (0.5)	2.37 (0.02)	62.2 (0.3)
BHMM-E W/Q/D	47.6 (0.3)	46.3 (0.4)	48.8 (0.2)	2.41 (0.02)	61.2 (1.1)
BHMM-T Q/D:	47.9 (0.5)	46.7 (0.4)	49.1 (0.5)	2.40 (0.02)	61.5 (0.4)
BHMM-T W/Q/D:	**49.6** (0.2)	48.5 (0.2)	50.8 (0.2)	2.31 (0.10)	64.1 (0.2)

Table 6: Results for BHMM, BHMM-E, and BHMM-T on non-English corpora.

tion and emission probabilities (which is equivalent to training a separate BHMM for each sentence type).

To test the generality of our findings, we carried out a series of cross-linguistic experiments, integrating sentence type information in unsupervised tagging models for Spanish, Cantonese, and Dutch. The results for Cantonese and Dutch mirrored those for English, showing a small increase in tagging performance for models that included sentence type information. For Spanish, no improvement was observed.

References

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

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

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

Daniel Hirst and Albert Di Cristo, editors. 1998. *Intonation systems: a survey of twenty languages*. Cambridge University Press.

Thomas H.T. Lee, Colleen H Wong, Samuel Leung, Patricia Man, Alice Cheung, Kitty Szeto, and Cathy S P Wong. 1994. The development of grammatical competence in cantonese-speaking children. Technical report, Report of RGC earmarked grant 1991-94.

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

Marina Meila. 2007. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, 98:873–895.

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

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP*.

David Snow and Heather Balog. 2002. Do children produce the melody before the words? A review of developmental intonation research. *Lingua*, 112:1025–1058.

Daniel N. Stern, Susan Spieker, and Kristine MacKain. 1982. Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, 18(5):727–735.

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

Anna Theakston, Elena Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28:127–152.