Evaluating models of syntactic category acquisition without using a gold standard

Citation for published version:

Published in:
Proceedings of the 31st Annual Conference of the Cognitive Science Society
Evaluating Models of Syntactic Category Acquisition without Using a Gold Standard

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

Abstract
A number of different measures have been proposed for evaluating computational models of human syntactic category acquisition. They all rely on a gold standard set of manually determined categories. However, children’s syntactic categories change during language development, so evaluating against a fixed and final set of adult categories is not appropriate. In this paper, we propose a new measure, substitutable precision and recall, based on the idea that words which occur in similar syntactic environments share the same category. We use this measure to evaluate three standard category acquisition models (hierarchical clustering, frequent frames, Bayesian HMM) and show that the results correlate well with those obtained using two gold-standard-based measures.

Introduction
By the time children reach school age, they have achieved the remarkable feat of acquiring most of their native language, typically without explicit instruction. This includes the acquisition of syntactic categories (nouns, verbs, adjectives, etc.). A number of computational models of category learning have been developed, most of which conceptualize the problem as one of grouping together words whose syntactic behavior is similar. Typically, the input for the model is taken from a corpus of child-directed speech, and clusters are created based on distributional information (Redington et al., 1998; Mitz, 2003; Parisien et al., 2008).

A problem common to all existing models is the evaluation of the model clusters. Often researchers have tested the output of their models against gold-standard category assignments, such as those available in the CHILDES database (MacWhinney, 2000). These gold-standard categories are based on the intuition of human annotators and are representative of adult morphosyntactic knowledge. Therefore, this type of evaluation is not ideal for assessing the syntactic categories of children, as these may include linguistically valid distinctions not recognized by the gold standard. Conversely, the gold standard may make distinctions that children do not have, or only acquire during language development. For example, at the age of two, English-learning children have not fully acquired the verb category (Olguín & Tomasello, 1993), and functional categories such as determiners are acquired even later (Kemp et al., 2005).

It is therefore highly desirable to develop an evaluation measure that does not make reference to an (adult) gold standard. On the other hand, the measure should give results that correlate with gold-standard-based measures, indicating that it is capable of capturing the linguistic distinctions inherent in the gold-standard. Finally, the ideal measure needs to be applicable to a wide range of different acquisition models (e.g., it should not be limited to probabilistic models).

This paper proposes a new evaluation measure which meets these criteria: substitutable precision and recall. It relies on a classical idea from linguistics, viz., that words which share the same syntactic category occur in similar syntactic environments. It does not require a gold standard, and therefore is suitable for evaluating pre-adult categories. At the same time, it yields results that correlate with gold-standard-based measures. We will show this by applying our new measure, as well as existing measures, to three standard models that discover syntactic categories in child-directed speech. This is the first time these models have been systematically compared; previous authors have used their own evaluation measures and only applied them to their own data sets, thus making a comparison across models difficult.

Gold-standard-based Evaluation Measures
In the following section we describe two evaluation measures that have been used to evaluate category acquisition models. Both require gold-standard labeled data, which is problematic from an acquisition standpoint for the reasons previously discussed. Hand-labeled data is also scarce, particularly for languages other than English.

Some of the models we investigate categorize word types (a type being a word such as duck), whereas others categorize tokens (particular instances of duck). In order to compare both kinds of models, the measures we describe are used to score tokens, not types.

Matched Accuracy This measure is widely used in the field of Natural Language Processing for unsupervised part-of-speech tagging, in which the tokens of a text are automatically annotated (“tagged”) with cluster numbers. To obtain the matched accuracy $MA$, the clusters induced by the model are mapped onto the gold-standard categories in order to provide a gold-standard part-of-speech label for each cluster. $MA$ is then defined as the percentage of word tokens with correct category labels. The crucial aspect is the mapping between the clusters and the gold standard categories. In this paper, we use many-to-one accuracy, where each model cluster is matched onto the gold-standard category with which it shares the most tokens. This can result in a situation where multiple clusters are mapped onto the same gold standard category. This means the model is not penalized for creating more fine-grained clusters than the gold standard.
Pairwise Precision and Recall  These measures are widely
used in the cognitive literature on category acquisition
(e.g., Redington et al. 1998; Mintz 2003), and are sometimes
referred to as accuracy and completeness. To compute them,
we consider all possible word pairs. If the words in a pair are
grouped together by the model correctly (i.e., they are in the
same gold-standard category and in the same model cluster),
a true positive ($tp$) is recorded; if they are not in the same
gold-standard category, a false positive ($fp$) is recorded. If
the two words are not grouped together by the model, but are
in the same gold-standard category, then a false negative ($fn$)
is recorded. Pairwise precision and recall is then defined as:

$$PP = \frac{tp}{tp + fp} \quad PR = \frac{tp}{tp + fn} \tag{1}$$

Note that $tp + fp$ is the total number of pairs within model
clusters, whereas $tp + fn$ is the total number of pairs within
the same category in the gold-standard. $PP$ thus measures
the proportion of correct pairs within the model clustering
(i.e., whether the model clusters together the correct words),
while $PR$ measures the number of correct pairs as a fraction
of all pairs in the gold standard (i.e., whether all correct pairs
have been found).

Substitutable Precision and Recall

Our goal is to capture the essential nature of syntactic cate-
gories without using the actual categories themselves. Distri-
butional analysis gives us the notion of substitutability (Har-
riss, 1946; Brown & Fraser, 1964) as the key aspect of syntac-
tic categories. Substitutable categories are made up of words
with identical “privileges of occurrence”, i.e., a syntactic cat-
egory consists of words which may be substituted for each
other within a sentence without making the sentence ungram-
matical. For example, he and she both belong to the same
category because he is happy and she is happy are both gram-
matical.

The measure we propose, substitutable precision and re-
call, evaluates category acquisition models by testing whether
substitutable words — words which appear in the same contexts — have been clustered together. Because nearly-
identical sentences (which would be necessary to strictly
evaluate substitutability) are rare in corpora, we restrict our
notion of context to frames: two words appearing in the cor-
pus with exactly one word in between. From these frames, we
create substitutable clusters (S-clusters) that consist of the set
of word types that appear within the same frame. There is a
one-to-one correspondence between S-clusters and frames.

Substitutable precision and recall are calculated similarly
to standard pairwise precision and recall. However, this does
not require a gold standard; instead, the set of clusters $C$
induced by the model is compared with the set of S-clusters $S$.
Substitutable precision $SP$ (Eq. 2) thus measures whether the
clusters consist of substitutable words, while substitutable re-
call $SR$ measures to what extent substitutable words have been
clustered together (Eq. 3).\footnote{Note that we retain the pairwise nature of pairwise precision and recall, which leads to the second term in the products (i.e., the number of non-identical pairs in a cluster is $|c|(|c| - 1)$.)

\begin{equation}
SP = \frac{\sum_{s \in S} \sum_{c \in C} |s \cap c|(|s \cap c| - 1)}{\sum_{c \in C} |c|(|c| - 1)} \tag{2}
\end{equation}

\begin{equation}
SR = \frac{\sum_{s \in S} \sum_{c \in C} |s \cap c|(|s \cap c| - 1)}{\sum_{s \in S} |s|(|s| - 1)} \tag{3}
\end{equation}

Because the models we are investigating use context infor-
mation that is similar to frames, there may be danger of over-
fitting the evaluation measure to the models and their training
data. To avoid this, we compute $SP$ and $SR$ using a separate
test corpus. The S-clusters used for evaluation are based on
frames found in both the training and the test corpus, but the
words within each S-cluster are from the test corpus only (the
test words must be in the training corpus vocabulary). Under
the distributional definition, syntactic categories can be inter-
preted as expectations of substitutability, regardless of whether
the members of the category have appeared in the
same syntactic context. By using separate, additional data to
measure substitutable precision and recall, we evaluate the
extent to which these learned expectations of substitutability
generalize to increasing amounts of data.

If a frame is made up of words with multiple (model) clus-
ter memberships, the model may have discovered a valid am-
biguity. For example, the frame to — cake is (using gold stan-
dard tags) ambiguous between $to_{INF} — cake_{N}$ (“We are going
to eat cake today”), which has an S-cluster consisting of
words such as bake and eat, and $top_{PREP} — cake_{N}$ (“Put the
juice next to his cake”), with a corresponding S-cluster con-
sisting of words such as his or that. For this reason, we add
cluster membership information (as found by the model be-
ing evaluated) to each frame word, as well as to the words in
the S-clusters.

By using a separate test corpus, we introduce a dependency
on the size of the test set. In our experiments, we use a test set
that is six times the size of the training set (we use the Man-
chester corpus (1.5M words) to train, and the rest of CHILDES
(9M words) to test). Additionally, we only evaluate on frames
that occur more than once within the test data, since a single
occurrence gives no information about which words should
be clustered together, and a single occurrence of a rare event
also gives little information about which words not to cluster
together.

Models of Syntactic Category Acquisition

In this section we briefly describe three models\footnote{We use the word model loosely; the authors of these systems
do not always assume that they are modeling human learning, but may
only be examining the possible usefulness of distributional cues.} of syntac-
tic category acquisition that we will use to test our evaluation
method. These models were chosen primarily for being rep-
resentative of the space of possible models: they differ, for
example, in their treatment of syntactically ambiguous words
and whether or not they categorize every word in the corpus.
**Frequent Frames**

Our first model is based on the frequent frames (FF) procedure for discovering syntactic categories described by Mintz (2003), which has been influential in the language acquisition community (see, e.g., Gómez & Maye 2005). Mintz’s approach is inspired by behavioral experiments suggesting that human learning of syntactic categories is strongly aided by the presence of frequently occurring frames (Mintz, 2002). In this case, a frame is defined as any ordered pair of words \((a, b)\) that occurs in the corpus with a single intervening word. (Note that this differs from our use in the context of evaluation, where the categories assigned to the words are also included in the frame.) The most frequent frames in the corpus are recorded, and for each one, all words that occur within that frame are assigned to the same cluster.

Our implementation follows Mintz in initially defining a cluster for each frame whose frequency is at least \(0.09%\) of the total number of frames in the corpus. Pairs of clusters with the highest overlap in word types, proportionally to the largest of the two clusters, are then iteratively merged until the target number of clusters is reached.

One drawback of FF is that only a very small percentage of tokens are clustered (4%–8% in Mintz’s experiments with corpora of child-directed speech), and these are almost exclusively nouns and verbs. The clusters do, however, have very high accuracy (i.e., words that are grouped together almost always belong to the same gold standard category), and Mintz points out that a much larger percentage of tokens (48%–61%) belong to the same types as those clustered, suggesting that these tokens could be added to the same clusters. However, this does nothing to cluster the large number of word types that never appear in a frequent frame. Moreover, it ignores the problem of syntactic ambiguity: first, because it is not clear what to do if a word type is initially assigned to multiple clusters, and second, because it assumes that all remaining tokens should belong to the same cluster, which may not reflect any true ambiguity.

**Hierarchical Clustering**

Researchers in both cognitive science and computational linguistics have proposed algorithms for syntactic category induction based on clustering context vectors (Redington et al., 1998; Clark, 2000; Schütze, 1995). We implemented the algorithm described by Redington et al. (1998), which has probably had the most impact in language acquisition. It treats the \(n\) most frequent word types in the corpus as the target words, and the \(m\) most frequent types are used as context words (where \(m < n\)). For each target word, a context vector \(\vec{v} = v_1 \ldots v_m\) is created, with \(v_i\) equal to the number of times the \(i\)th context word co-occurs with the target word. Specific context positions (e.g., one word to the left of the target, two words to the right) are accounted for by collecting separate vectors for each position and concatenating them. The similarity between vectors is computed using the Spearman rank correlation, and a tree structure is created by iteratively clustering the most similar words (or previously created clusters). By “cutting” the tree at different heights, different numbers of clusters can be produced. The best results of Redington et al. (1998) are with \(n = 1000, m = 150\), and two positions on either side of each target word as context. We use the same parameters here.

An important property of this hierarchical clustering (HC) model is the fact that the context vector for each word type combines the context counts for all tokens of that word. Therefore, every token of a particular word is assigned to the same syntactic category, regardless of the specific context in which it appears.

Although HC does not cluster every word in the corpus, its coverage is far more complete than that of FF. Even in a very large corpus, Zipf’s law ensures that the 1000 most frequent words account for most of the corpus. Despite its broad coverage, however, HC only performs well on words with high frequency, unlike children, who can learn words and their usage, i.e., their syntactic categories) on the basis of very few observations (Woodward et al., 1994). In contrast, FF may categorize some words that occur only once, provided they occur inside frequent frames.

**Bayesian Hidden Markov Model**

Unlike the previous algorithmic approaches, the third approach is based on a probabilistic model. We consider the Bayesian HMM (BHMM) proposed by Goldwater & Griffiths (2007) as our third model because it contrasts with the previous two on several levels: in addition to being based on a probabilistic model, it categorizes every word in the corpus, and it can deal with ambiguity, i.e., it may assign different tokens of the same word type to different clusters.

As a variant of the standard HMM, this model assumes that the corpus is probabilistically generated as a sequence of cluster labels (tags), each of which in turn generates the observed word. The model considers different possible sequences of tags, searching for a sequence that can explain the observed words well, while also being linguistically plausible. In this case, plausibility is enforced using Bayesian priors to capture the intuition that the HMM transition and output distributions are sparse, i.e., that each tag is followed by relatively few other tags with high probability, and outputs relatively few words with high probability. In contrast to the other models, neighboring words affect the BHMM’s decision about a word’s category only indirectly, through their category labels.

Our implementation of the BHMM uses Gibbs sampling to identify a sequence of tags that has high probability under the model. In this implementation, the only free parameter of the model is the number of clusters used. In our experiments, we ran the Gibbs sampler for 2000 iterations.

**Model Implementation and Experiments**

For all our experiments, we used the Manchester corpus (Theakston et al., 2001), which is annotated with syntactic
categories and is part of the CHILDES database (MacWhinney, 2000). The Manchester corpus consists of transcribed recordings of 12 children interacting with adults, and covers an age range of 1 year 8 months to 3 years. Our models are trained only on child-directed speech (CDS), so we removed all child utterances, as well as any utterances containing unintelligible words; additionally, we split contractions (e.g., aren’t) into separate words and added beginning-of-sentence and end-of-sentence markers (which were included in the frames used to create S-clusters, but not in the frames used to train the FF model). This left approximately 1.5M words and 360,000 child-directed utterances.

The original set of syntactic categories used for the Manchester corpus contains detailed morphosyntactic information, e.g., playing is annotated part|play-PROG. After stripping out the morphological information, the category inventory contained 53 categories. We also created a collapsed inventory consisting of 12 categories (see Table 1).

For each of the models described in the previous section, we varied the number of clusters \( K \) over three conditions: 12 (as in the collapsed category inventory), 53 (as in the original inventory) and 80 (to create more fine-grained clusters). One of the advantages of the substitutable precision-recall measures is that they do not depend on the gold standard for the “true” number of clusters; thus there is no penalty for a clustering that does not have the same number of clusters as the gold standard.

We also compared each \( K \) condition against a random baseline. For each cluster in the gold standard, we created a cluster with the same number of word types, selected at random from the full vocabulary (the \( K = 80 \) and \( K = 53 \) conditions shared the same random baseline). This results in a random clustering with the same cluster size distribution as the gold standard, and all tokens of each type in the same cluster.

Our goal is to show that substitutable precision and recall yield informative evaluation results without requiring a gold standard. We therefore evaluated each category acquisition model not only with substitutable precision and recall, \( SP \) and \( SR \), but also with the measures introduced earlier: matched accuracy \( MA \) and pairwise precision \( PP \) and recall \( PR \).

A problem arises, however, when we try to compare the three clustering models; they each categorize a different subset of the data. The BHMM model assigns categories to every token, the FF model assigns categories to only those word types which appear within frequent frames, and the HC model only categorizes the 1000 most frequent word types. We resolve this problem in two ways. In the merge condition, we combine all words that are not clustered by the model into one large cluster. In the split condition, we assign each unclustered word to its own cluster. The difference in performance between these two conditions thus indicates the effect of the unclustered words.

Additionally, it is necessary to assign each token in the text to a category, if the model does not categorize tokens (as BHMM does). For HC, each token of a given type is assigned to the type’s category. In the FF model, a word type can belong to multiple categories, making it unclear which category a particular token should be assigned to. We assigned tokens found in frequent frames to the category defined by that frame; other instances of ambiguous types were assigned to a given cluster with probability \( p(c_i|w_j) = \frac{|c_i|}{\sum_{i \in c_j} |c_i|} \), that is, according to the size of the clusters that include the ambiguous word type.

**Results**

The results are given in Table 3. We first discuss them in light of our proposed evaluation measures, \( SP \) and \( SR \), and then go on to compare the performance of the different models.

**Comparison of Evaluation Measures**

Figure 1 shows the similar performance of the two precision-recall measures. Results for \( SP \) and \( SR \) are significantly cor-

---

### Table 1:Collapsed gold-standard categories

| ADJ | Advjectives, e.g., funny, pink |
| ADV | Adverbs, e.g., today, just, normally |
| OTH | Miscellaneous, e.g., yes, well, hurray |
| CONJ | Conjunctions, e.g., and, or |
| DET | Determiners, e.g., a, those, six |
| INF | Infinitival to |
| N | Nouns and Pronouns, e.g., you, ducky |
| NEG | Negations, e.g., not |
| PART | Particles, e.g., raining, hidden |
| PREP | Prepositions, e.g., on, to, after |
| QN | Quantifiers, e.g., many, all, some |
| V | Verbs, e.g., swim, do, is |

---

### Table 2: Spearman’s rho correlations between the rankings given by a pair of evaluation metrics (\( N = 12 \), computed using the merge condition results; **: \( p < 0.01 \); *: \( p < 0.05 \); all correlations not included in the table are non-significant.

<table>
<thead>
<tr>
<th>Model</th>
<th>( K )</th>
<th>( PP )</th>
<th>( PR )</th>
<th>( MA )</th>
<th>( SP )</th>
<th>( SR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>12</td>
<td>0.205</td>
<td>0.324</td>
<td>0.796</td>
<td>0.000065</td>
<td>0.254458</td>
</tr>
<tr>
<td>Random</td>
<td>53</td>
<td>0.096</td>
<td>0.254</td>
<td>0.720</td>
<td>0.000092</td>
<td>0.173907</td>
</tr>
<tr>
<td>BHMM</td>
<td>12</td>
<td>0.570</td>
<td>0.263</td>
<td>0.721</td>
<td>0.000221</td>
<td>0.308508</td>
</tr>
<tr>
<td>BHMM</td>
<td>53</td>
<td>0.624</td>
<td>0.175</td>
<td>0.747</td>
<td>0.000347</td>
<td>0.109927</td>
</tr>
<tr>
<td>BHMM</td>
<td>80</td>
<td>0.657</td>
<td>0.128</td>
<td>0.775</td>
<td>0.000330</td>
<td>0.084811</td>
</tr>
<tr>
<td>HC</td>
<td>12</td>
<td>0.201</td>
<td>0.864</td>
<td>0.361</td>
<td>0.000046</td>
<td>0.375467</td>
</tr>
<tr>
<td>HC</td>
<td>53</td>
<td>0.330</td>
<td>0.654</td>
<td>0.523</td>
<td>0.000117</td>
<td>0.202372</td>
</tr>
<tr>
<td>HC</td>
<td>80</td>
<td>0.484</td>
<td>0.512</td>
<td>0.639</td>
<td>0.000159</td>
<td>0.183736</td>
</tr>
<tr>
<td>FF</td>
<td>12</td>
<td>0.220</td>
<td>0.244</td>
<td>0.448</td>
<td>0.000027</td>
<td>0.217124</td>
</tr>
<tr>
<td>FF</td>
<td>53</td>
<td>0.219</td>
<td>0.079</td>
<td>0.392</td>
<td>0.000039</td>
<td>0.120499</td>
</tr>
<tr>
<td>FF</td>
<td>80</td>
<td>0.224</td>
<td>0.053</td>
<td>0.423</td>
<td>0.000043</td>
<td>0.096760</td>
</tr>
</tbody>
</table>

---

### Table 3: Results for the merge condition. The best score for each evaluation type and number-of-clusters condition is highlighted.

The Manchester corpus contains detailed morphosyntactic information, e.g., playing is annotated part|play-PROG. After stripping out the morphological information, the category inventory contained 53 categories. We also created a collapsed inventory consisting of 12 categories (see Table 1).

For each of the models described in the previous section, we varied the number of clusters \( K \) over three conditions: 12 (as in the collapsed category inventory), 53 (as in the original inventory) and 80 (to create more fine-grained clusters). One of the advantages of the substitutable precision-recall measures is that they do not depend on the gold standard for the “true” number of clusters; thus there is no penalty for a clustering that does not have the same number of clusters as the gold standard.

We also compared each \( K \) condition against a random baseline. For each cluster in the gold standard, we created a cluster with the same number of word types, selected at random from the full vocabulary (the \( K = 80 \) and \( K = 53 \) conditions shared the same random baseline). This results in a random clustering with the same cluster size distribution as the gold standard, and all tokens of each type in the same cluster.

Our goal is to show that substitutable precision and recall yield informative evaluation results without requiring a gold standard. We therefore evaluated each category acquisition model not only with substitutable precision and recall, \( SP \) and \( SR \), but also with the measures introduced earlier: matched accuracy \( MA \) and pairwise precision \( PP \) and recall \( PR \).

A problem arises, however, when we try to compare the three clustering models; they each categorize a different subset of the data. The BHMM model assigns categories to every token, the FF model assigns categories to only those word types which appear within frequent frames, and the HC model only categorizes the 1000 most frequent word types. We resolve this problem in two ways. In the merge condition, we combine all words that are not clustered by the model into one large cluster. In the split condition, we assign each unclustered word to its own cluster. The difference in performance between these two conditions thus indicates the effect of the unclustered words.

Additionally, it is necessary to assign each token in the text to a category, if the model does not categorize tokens (as BHMM does). For HC, each token of a given type is assigned to the type’s category. In the FF model, a word type can belong to multiple categories, making it unclear which category a particular token should be assigned to. We assigned tokens found in frequent frames to the category defined by that frame; other instances of ambiguous types were assigned to a given cluster with probability \( p(c_i|w_j) = \frac{|c_i|}{\sum_{i \in c_j} |c_i|} \), that is, according to the size of the clusters that include the ambiguous word type.

**Results**

The results are given in Table 3. We first discuss them in light of our proposed evaluation measures, \( SP \) and \( SR \), and then go on to compare the performance of the different models.

**Comparison of Evaluation Measures**

Figure 1 shows the similar performance of the two precision-recall measures. Results for \( SP \) and \( SR \) are significantly cor-
Recall and Precision

Pairwise Precision and Recall

Substitutable Precision and Recall

Figure 1: Pairwise precision and recall on the left, substitutable precision and recall on the right. The size of the points indicates the number of clusters: small points are for 12 clusters, medium points for 53, large points for 80. HC and FF results are for the merge condition.

Figure 2: Ranked cluster sizes, measured in types: the x-axis represents the clusters, which are ordered according to size; the y-axis gives the size of the clusters on a log scale (GS: gold standard; BHMM: Bayesian HMM; HC: hierarchical clustering, FF: frequent frames). Note the random baseline clusters have the same type-size-distribution as GS.

Model Performance

The different model types find very different word clusterings, as Fig. 2 helps to illustrate. HC creates clusters with highly skewed sizes (most extremely so in the 12 cluster condition, in which 969 of the 1000 clustered word types are put into one cluster). The cluster size distribution of FF models is much flatter, indicative of FF’s propensity to create highly ambiguous clusterings, in which each word type belongs to many clusters. The BHMM clusterings also have higher levels of lexical ambiguity than the gold standard, resulting in more larger clusters overall, both in terms of types and tokens. Both BHMM and FF tend towards more ambiguity with more clusters. It should be noted as well that the token distributions are highly similar to the type distributions.

Effect of Unclustered Words

Both the HC and FF do not cluster all word types found in the training data. The HC model clusters only the most fre-
The paper deals with the problem of evaluating computational models of human syntactic category acquisition. We started from the observation that children’s syntactic categories change during language development, which means that an evaluation against a fixed gold-standard (typically based on adult linguistic intuitions) is not adequate. As an alternative, we proposed substitutable precision and recall, a measure based on the idea that words which share the same category occur in similar syntactic environments. We included 90% of the word types (and thus many with the same gold standard category).

**Conclusions and Future Work**

This paper also presented the first systematic comparison of three standard acquisition models from the literature: Redington et al.’s (1998) hierarchical clustering model, which performed well on recall-oriented measures, Goldwater & Griffith’s (2007) Bayesian HMM, which performed well on precision-oriented measures, and Mintz’s (2003) frequent frame model which showed surprisingly poor performance.

Finally, we also demonstrated that evaluation results strongly depend on how unclustered words are evaluated.

In future work, we will explore the external validity of substitutable precision and recall. While it is important to show that it correlates with existing evaluation measures, we also need to test it against experimental data (e.g., substitutability judgments). Additionally, we plan to apply it to longitudinal acquisition corpora to evaluate models which follow the time course of category development.

**References**


