# Modelling the acquisition of verb polysemy in children

**Christopher Parisien and Suzanne Stevenson**
Department of Computer Science, University of Toronto
Toronto, ON, Canada
[chris,suzanne]@cs.toronto.edu

## Abstract

We use token-level clustering methods to simulate children's acquisition of the senses of a polysemous verb. Using actual child-directed data, we show that simple syntactic features commonly used in distributional models are sufficient to reasonably distinguish verb senses. However, these features are inadequate to account for the order of acquisition of polysemy as observed in children, and we argue that future models will need to incorporate other types of information in order to better explain child behaviour.

**Keywords:** Verb semantics; polysemy; child language acquisition; Bayesian models; clustering.

## Introduction

The acquisition of verb polysemy has become an important target of study in cognitive linguistics and developmental psychology (*e.g.*, Nerlich, Todd, & Clarke, 2003; Theakston, Lieven, Pine, & Rowland, 2002). Some of the most highly frequent and earliest learned English verbs, like *put*, *make*, *get*, *go*, and *do*, are also among those with the largest number of senses (Clark, 1996). Children as young as two years of age freely understand and use many of these polysemous verbs, often with little apparent confusion (Theakston et al., 2002; Israel, in press). Computational models can help to elucidate the kinds of mechanisms capable of distinguishing the senses of massively polysemous verbs from very little input, as well as the linguistic features necessary to achieve this.

Information about verb senses has been said to correlate strongly with verb argument structure. Several computational models have been developed that make use of a verb's possible arguments to identify semantic structure and similarity to other verbs. Most of these models operate at a coarse-grained semantic level, clustering verb types into general classes of similar verbs (*e.g.*, Versley, 2008; Korhonen, Krymolowski, & Marx, 2003). On the other hand, computational models of child language acquisition have found success by clustering word *usages* (*e.g.*, Alishahi & Stevenson, 2008), that is, individual instances of verbs along with their contexts. In this paper, we argue that such usage-based models can be used to study children's acquisition of verb polysemy.

We analyze the English verb *get* as a case study. *Get* is a particularly interesting target since it is highly frequent, highly polysemous, and is one of the first verbs children learn (Clark, 1996). Table 1 outlines the major senses of *get*, with their frequencies estimated from a corpus of adult spoken language (Berez & Gries, 2009). Other sets of senses may be found in the literature, but this offers a good assessment of the breadth of meaning captured by the verb. Here, we conflate literal and metaphorical senses. For example, the metaphorical use *I got an idea* falls under the general sense *obtain*. Various infrequent senses are gathered under *other*.

| Sense | Freq. (%) | Example |
|---|---|---|
| obtain | 52.3 | I got a book. |
| cause obtain | 1.3 | I got you a book. |
| move | 16.5 | You should get on that bus. |
| cause move | 5.2 | It'll get you to Buffalo. |
| become | 15.0 | Jim got fired. |
| cause become | 2.5 | Suzie got Jim fired. |
| must | 6.3 | I've got to go home. |
| other | 0.8 | You get to eat cake! |

Table 1: Coarse-grained senses of *get*.

Children tend to learn more frequent verb senses earlier than less frequent senses (Theakston et al., 2002; Israel, in press). However, the order of acquisition does not completely follow the frequency ordering, and this shows that something other than the frequencies of these related polysemous senses contributes to the ease of acquisition. This is a challenge for distributional clustering models, where performance is generally improved with greater amounts of data.

In this paper, we use a hierarchical Bayesian clustering model to group individual usages of the verb *get*, drawn from a corpus of child-directed speech. We show good clustering results by using a set of simple, automatically extracted syntactic features. We argue that while these features are commonly used in distributional models of verb semantics, they are inadequate to explain order of acquisition behaviour in children.

## Related work

Several recent computational models have demonstrated the value in using argument structure information to learn about verb semantics. Versley (2008) and Schulte im Walde (2008) cluster verb types using various syntactic dependencies such as noun phrases, prepositional phrases, and adverbs. Joanis, Stevenson, and James (2008) achieve similar goals using syntactically shallow slot features – subject, direct and indirect object, for example. In each case, the simple argument structure patterns correlate with human judgements of semantic verb classes.

Few approaches explicitly address the problem of multiple senses of a single verb type. The work of Korhonen et al. (2003) uses a soft-clustering method that allows a verb to belong to multiple possible clusters, allowing a degree of polysemy in a verb's representation. Verbs are clustered by the distribution of their subcategorization frames. If two senses of a verb differ strongly in their subcategorization patterns, the verb will more likely be distributed across multiple clusters. Vlachos, Korhonen, and Ghahramani (2009) use similar subcategorization features in their approach, employing a Dirichlet process mixture model (DPMM) as the clustering algorithm to give the flexibility of learning an unspecified number of clusters. In this case a probabilistic soft-clustering is possible, although the authors do not examine this aspect of the model.

Each of these approaches is concerned with *type-level* clustering of verbs, that is, clustering verbs based on the distributional properties of all the verb's usages, taken together. The model may recognize that *run*, *skip* and *walk* are similar, and in the case of Korhonen et al. (2003), that *run* is also similar to *flow*, as in *the river runs east*. However, the verb itself is still represented as a single point in distributional space. A *token-level* method, on the other hand, clusters individual usages of verbs. This way, different senses can occupy distinct representations of the same verb. Evidence from psycholingusitics suggests that such a method may be necessary to fully explain polysemy (Theakston et al., 2002). Very few models address token-level verb clustering. Lapata and Brew (2004) use subcategorization patterns to perform token-level classification (not clustering) of verbs, thereby presupposing a defined set of verb classes. Alishahi and Stevenson (2008) cluster individual verb usages to simulate the acquisition of verb argument structure in children. Their method of clustering by using basic argument information is similar to our perspective, although the incremental algorithm is necessarily sensitive to the order of presentation of the input.

## Verb usage clustering

In this section, we describe our modelling framework for clustering verb usages into senses. We discuss the feature representations of individual verb usages, then describe our application of a DPMM, a Bayesian clustering framework well suited to models of human category learning.

### Verb features

Following from the type-level verb clustering approaches described above, we designed our feature space to capture some of the general argument structure distinctions between verb senses. We primarily use syntactic "slot" features, similar to those used by Joanis et al. (2008), to encode basic argument information about a verb usage. These are not subcategorization frames, but rather a set of individual features that record the presence or absence of syntactic positions – subject, direct and indirect object, for example – that potentially contain verb arguments. In any particular usage, a certain slot may be analyzed as an adjunct rather than a true argument. Such slot features are easier to extract than full subcategorization frames, and Joanis et al. (2008) show that in verb classification tasks, subcategorization frames offer no improvement over simple slot features.

| Symbol(s) | Feature values |
|---|---|
| SUBJ, CSUBJ, XSUBJ | Subjects |
| OBJ, OBJ2, IOBJ | Objects |
| COMP, XCOMP | Clausal complements |
| PRED, CPRED, XPRED | Nominal, adjectival or prepositional complements |
| LOC | Locatives |
| JCT, CJCT, XJCT | Adjuncts |
| PREP | Preposition (nominal value) |
| NSLOTS | Number of slots used |

Table 2: Slot features.

Table 2 presents the 17 features used in our representation. The first 15 are binary features denoting the presence or absence of a slot. Since our input data is extracted from the CHILDES database of child-directed speech and child language (MacWhinney, 2000), the labels correspond to the grammatical relations used by the CHILDES dependency parser (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007). When one of the other relations is a prepositional phrase, the nominal feature PREP denotes the preposition used.

### Dirichlet process mixture model

As stated earlier, the goal of our approach is to learn clusters of verb usages that approximate verb senses. To achieve this, we use a DPMM, a non-parametric Bayesian model that has gained significant attention in the machine learning community (Neal, 2000). A DPMM brings two main advantages over other clustering methods. Firstly, the modeller need not specify in advance the number of clusters necessary to represent the data. This is the "non-parametric" aspect of the model: as part of the learning process, the model itself determines an appropriate number of clusters, dependent on the data. Secondly, the DPMM has been shown to be a good model of human category learning behaviour (Sanborn, Griffiths, & Navarro, 2006). In addition to basic category-learning tasks, DPMMs and related models have successfully been applied to word segmentation (Goldwater, Griffiths, & Johnson, 2009) and type-level verb clustering (Vlachos et al., 2009).

A DPMM specifies a probability distribution over possible cluster arrangements of data. In contrast to typical algorithms that seek a single "best" clustering of data points, a DPMM gives a distribution over *all* possible clusterings. Given the observed verb usage data, we can estimate the parameters of that distribution to find the most likely clusterings.

We assume that each verb usage $\mathbf{y}_i$ belongs to a cluster, and that its features are drawn from a set of multinomial distributions (one per feature). Different clusters are associated with different feature distributions. Thus, one cluster may probabilistically represent a pattern of features such as SUBJ V OBJ, while another cluster may represent the pattern SUBJ V OBJ COMP. The number of clusters in turn depends on a Dirichlet Process (DP), a stochastic process which gives the model its non-parametric flexibility. The full model is:

$$y_{ij}|\theta_{jz_i} \sim \text{Mult}(\theta_{jz_i})$$
$$\theta_{jz_i}|G \sim G$$
$$G|\alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

The $\sim$ symbol should be read as "is distributed according to". In the above, $y_{ij}$ denotes feature $j$ of usage $i$. $z_i$ is the cluster chosen for usage $i$, and $\theta_{jz_i}$ are the multinomial parameters for feature $j$ in the probabilistic pattern represented by the cluster. $G$ and $G_0$ are probability distributions over the parameters $\theta$, and $\alpha$ is a concentration parameter that affects how many clusters we expect to find.

In the above, $G$ generates the parameters of the multinomial distribution ($\theta_{jz_i}$) that in turn generates $y_{ij}$. Since $G$ selects $\theta$ from across the set of clusters (*e.g.*, $\theta_{j1}$ or $\theta_{j2}$), it is in effect a mixing distribution that gives the probabilities of choosing each cluster.

The DP, being defined by both the concentration parameter $\alpha$ and a *base distribution* $G_0$, gives a prior distribution on the number and size of the clusters as well as on the parameters $\theta$ to represent them. $G_0$ defines the prior distribution for $\theta$. We set $G_0$ to $\text{Dir}(\mathbf{1})$, a noninformative Dirichlet prior. Also, the DP gives a prior probability on the entire partitioning of the data into clusters. It is derived from the following stochastic process: assume that all verb usages have been clustered *except* $\mathbf{y}_i$. Then the prior probability of a cluster $k$ is given by

$$P(k) = \begin{cases} \frac{n_k}{N-1+\alpha} & \text{if } n_k > 0 \text{ (existing cluster),} \\ \frac{\alpha}{N-1+\alpha} & \text{otherwise (new cluster),} \end{cases} \tag{1}$$

where $n_k$ is the number of verb usages in cluster $k$ and $N$ is the total number of usages. Larger values of $\alpha$ make it more likely that overall, more clusters will be used. In all our experiments, we set $\alpha = 1$, a moderate setting that compares with similar DPMM applications. This formulation has two interesting properties. Firstly, larger clusters tend to attract more usages. Secondly, as more data is processed, the probability of choosing a new cluster decreases.

The above model, as written, specifies a prior distribution over the complete set of possible parameters to the model (*i.e.*, all possible values for $\theta$ and $\mathbf{z}$). To find clusters of verb usages, we update this distribution using the observed data, thus obtaining a posterior distribution over parameters.

### Parameter estimation

Given the set of verb usage data, we estimate the posterior distributions over the model parameters using Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method (Neal, 2000). Essentially, to estimate a probability distribution, we draw a large number of samples from that distribution. The samples give an approximation of the distribution, and as the number of samples approaches infinity, the approximation becomes exact. With Gibbs sampling, we choose an initial random setting for the model parameters (*i.e.*, the cluster assignments $\mathbf{z}$ and the cluster parameters $\theta$), then iteratively adjust these settings according to the observed data.

In our experiments, we randomly set each $z_i$ to one of a small number of clusters (1, 2, or 3). For each cluster, we set the $\theta$ parameters to random values drawn from a Dirichlet distribution. We iteratively update each $z_i$ and $\theta_{jk}$ *individually* by drawing it from a posterior distribution conditioned on the data and all the *other* parameters in the model. In the case of a cluster assignment $z_i$, we do this by sampling a cluster for $\mathbf{y}_i$ given assignments for all the other usages, as if $\mathbf{y}_i$ were the last usage observed. We may choose a new cluster (as in Equation 1), thus potentially changing the total number of clusters. We repeatedly cycle through the model parameters, sampling each $\theta_{jk}$ and each $z_i$ many times. By averaging over a large number of these samples, the posterior approximation converges on the exact solution. In practice, we can achieve a good estimate in a few thousand samples, depending on the complexity of the data and the details of the algorithm.

## Experiments

In our experiments, we use child-directed speech data drawn from the CHILDES database of parent-child interactions (MacWhinney, 2000). We use four longitudinal corpora from the American English component of the database, corresponding to four children: Eve, Naomi, Nina, and Peter. Together, the data cover an age range from 1;2 (years;months) to 4;9. We extract each child-directed utterance of the verb *get*, then randomly split the utterances into development and test sets (1275 and 1276 utterances respectively), dividing each child's data equally. The corpora contain part-of-speech tags and syntactic dependencies, obtained using an automatic tagger and parser (MacWhinney, 2000; Sagae et al., 2007). As described above, we extract 17 slot features for each usage of *get*. Due to errors in the automatic part-of-speech tagging, parsing and feature extraction, the data contains some noise. Some utterances were dropped when parsing errors prevented extraction of the features, and others contain multiple instances of *get*. The final development set and test set contain 1272 and 1290 usages, respectively. For evaluation purposes, we manually annotate each of the usages with one of eight sense labels, corresponding to the eight senses in Table 1. We refer to this labelling as the gold standard.

We implement the DPMM in OpenBUGS, a general framework for performing MCMC simulations of hierarchical Bayesian models. We run five chains with different initial conditions: one chain is initialized with all usages in one cluster, two chains start with two clusters, and two with three clusters. Each chain is randomly initialized as described in the previous section. As per standard practice, we run each chain for 60,000 iterations, discarding the first 10,000 as burn-in. To reduce correlation in the samples, we keep only every 25th sample, giving 2,000 samples per chain, 10,000 in total.

Each sample contains one clustering of the verb usages. To evaluate the model's performance, we score each of the samples against the gold standard, then average the results over all samples. As a result, the reported scores give a weighted evaluation of the entire distribution of clusterings, not just the

| Sense | P (%) | R (%) | F (%) | Freq. (N) |
|---|---|---|---|---|
| 1. obtain | 61.3 | 53.1 | 56.9 | 576 |
| 2. cause obtain | 26.0 | 44.2 | 32.8 | 56 |
| 3. move | 62.4 | 50.7 | 56.0 | 196 |
| 4. cause move | 30.9 | 46.2 | 37.1 | 115 |
| 5. become | 59.7 | 58.2 | 59.0 | 253 |
| 6. cause become | 6.7 | 50.2 | 11.8 | 52 |
| 7. must | 2.9 | 75.3 | 5.6 | 19 |
| 8. other | 3.9 | 64.8 | 7.3 | 23 |

Table 3: Precision (P), recall (R) and F-measure (F) for each sense of *get*.

single "best" cluster. We evaluate each sample using the cluster F-measure (Larsen & Aone, 1999). Given one sample, for each sense $s$, we score each cluster $k$ as follows. Let $a$ be the number of usages in $k$ with sense $s$. Let $b$ be the total number of usages in the cluster, and let $c$ be the total number of usages with sense $s$, over all clusters. Then precision (P), recall (R), and F-measure (F) are given by:

$$P = \frac{a}{b}, \quad R = \frac{a}{c}, \quad F = \frac{2PR}{P+R}. \tag{2}$$

We record P, R, and F for the cluster with the best F-measure for that sense, then report averages over all 10,000 samples.

## Results

Table 3 presents the results of clustering using the DPMM on the test set usages of *get*. The model uses on average 5.2 clusters. The more frequent senses, *obtain*, *move*, and *become*, achieve the best performance. The less frequent causative senses show worse clustering behaviour, although the recall scores indicate that the model recognizes some internal similarity among the usages. In these cases, low precision scores suggest that the features of the causative senses are quite similar to those of other senses.

We examine this possibility in Figure 1, which shows the likelihood of grouping together verb usages from different senses. We calculate the likelihood of each usage of a given gold standard sense being placed in the same cluster as each other usage of the gold standard senses, taken over all 10,000 samples and averaged over usages within each sense. A perfect clustering would give a diagonal matrix. High values along the diagonal roughly translate to high recall, and low values on the off-diagonal indicate high precision. The figure shows that *cause obtain*, *cause move* and *cause become* are frequently grouped together (column 2, rows 2, 4 and 6). One possibility is that the model distinguishes causative meanings from non-causatives based on the larger number of arguments in causative forms, but lacks features that would effectively distinguish the various causative meanings from each other.

A common observation in child language acquisition studies is that the more frequent senses of a verb tend to be the earliest senses children produce (Theakston et al., 2002; Israel, in press). This role of frequency is unsurprising from a machine learning perspective, since we expect more data to make learning easier. Indeed, we see this effect in the results above: the more frequent senses tend to be easier to learn.

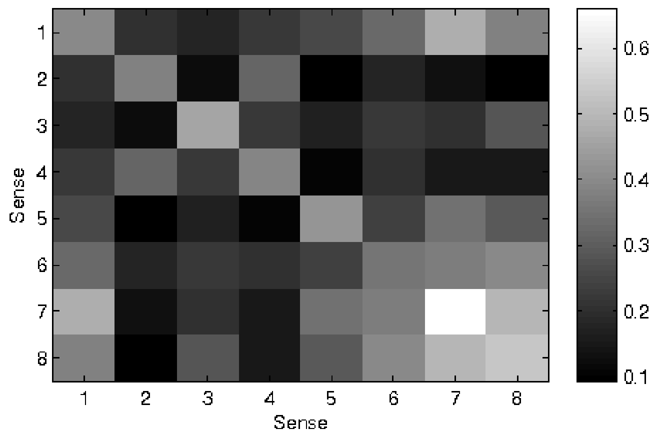


Figure 1: Likelihood of grouping usages from each pair of senses, averaged over all usages. Indices correspond to senses as in Table 3.

On the other hand, the role of frequency in acquisition is not a hard-and-fast rule. There are notable exceptions that can shed light on distributional semantic methods. Israel (in press) studied the order of acquisition of various senses of *get*, using the same transcripts as in our own study. Using the same sense categories as ours (excluding our category *must*), Israel compared the frequencies of senses in child-directed speech with the order in which the children first produce these senses. He notes that, in most cases, what a child hears most frequently, he or she learns quickly. The most common exception is *cause obtain*: despite comprising only 2-3% of the input, children often produce it before far more frequent senses like *become* or *cause move*.

This effect does not appear in our own results. We simulate the learning of verb senses over time by running the model on different-sized subsets of data, randomly sampled from the test set. Table 4 shows F-measures of each of the senses, for 400- and 800-usage subsets as well as the full test set. To replicate Israel's observations, we should expect to see high scores for *cause obtain* from small amounts of data, that is, earlier than when the scores improve for more frequent senses like *become* or *cause move*. We do not see this effect. Rather, *cause obtain* shows relatively poor performance for all three dataset sizes. It appears then that while slot features give promising clustering behaviour, they do not lend themselves to the kind of order of acquisition effects we observe in child behaviour.

Israel (in press), as well as Gries (2006), have suggested that the acquisition of polysemous verb senses may depend on complex inferential mechanisms on the part of the child. For example, the *become* sense of *get* may be a metaphorical extension of the *move* sense, for which children must observe a metaphorical connection between states and locations.

| Sense | N=400 | N=800 | N=1290 |
|---|---|---|---|
| obtain | 55.1 | 53.2 | 56.9 |
| cause obtain | 22.1 | 22.4 | 32.8 |
| move | 34.7 | 43.9 | 56.0 |
| cause move | 29.1 | 35.3 | 37.1 |
| become | 42.4 | 49.0 | 59.0 |
| cause become | 6.7 | 11.3 | 11.8 |
| must | 4.1 | 3.9 | 5.6 |
| other | 4.4 | 5.1 | 7.3 |
| Number of clusters | 2.8 | 3.6 | 5.2 |

Table 4: F-measures for varied amounts of data, simulating order of acquisition.

As an explanation for the early acquisition of *cause obtain*, a child could extend *obtain* by adding a causal agent, a connection which children appear to make quite early (Fisher, 2002). Our model does not make explicit inferences like these, which may explain why our results do not exhibit the same order of acquisition as in children. However, it may be that the behaviour we see in our model is due to the simplicity of our features, or the noise inherent in using automatically extracted data. Children may attend to some other aspect of the input not captured in our fairly simple feature set, something that helps them to acquire certain senses at an early age from comparatively little input. To investigate this, in the next section we apply our model to a richer set of hand-annotated features drawn from a corpus of adult spoken language.

## Richer syntactic features

Berez and Gries (2009) analyzed 600 adult-language instances of *get*, sampled from the British component of the International Corpus of English, ICE-GB. The authors annotated the data with 47 fine-grained senses, which we regroup into the 8 coarse-grained labels of Table 3. Each usage has been tagged with 13 features commonly used in verb clustering, drawn from the manual annotations of ICE-GB. These features cover a broad range of phenomena, including verb transitivity, verb form, grammatical relations such as the presence of auxiliary verbs, and clausal features including dependency types and the transitivity of dependent clauses.[1]

By encoding verb arguments and certain semantic relationships among them, transitivity patterns capture more information than subcategorization frames or slot features alone. For example, in the "copula" pattern used in this data, an adjectival or prepositional complement describes a property of the subject, as in, *I got rid of the car*. This semantic property distinguishes the copula from the syntactically similar intransitive pattern. Since these features are hand-annotated, we expect the data to contain fewer extraction errors and less noise than our own automatically extracted data. We cluster the verb usages using the DPMM and present the results in Table 5, scored as in the above experiments.

[1]See Berez and Gries (2009) for the full list of features.

| Sense | P (%) | R (%) | F (%) | Freq. (N) |
|---|---|---|---|---|
| obtain | 68.4 | 56.7 | 62.0 | 314 |
| cause obtain | 2.0 | 49.9 | 3.8 | 8 |
| move | 29.9 | 44.5 | 35.7 | 99 |
| cause move | 13.4 | 61.3 | 22.0 | 31 |
| become | 59.0 | 23.7 | 33.8 | 90 |
| cause become | 4.4 | 44.1 | 7.9 | 15 |
| must | 75.8 | 99.9 | 86.2 | 38 |
| other | 1.2 | 46.8 | 2.3 | 5 |

Table 5: Precision (P), recall (R) and F-measure (F) from clustering the data of Berez and Gries (2009).

Overall, these results show a similar pattern to the experiments on CHILDES data. The more frequent senses, *obtain*, *move*, and *become*, perform reasonably well, while the less frequent causative senses perform poorly. The exception is *must*, with a remarkably high F-measure of 86.2%. This sense is nearly always used in a form similar to *I've got to X*, with highly consistent auxiliary use, verb form and clausal form, all missing from our simple slot representation.

Even with a richer, manually annotated data set, the clustering results do not exhibit Israel's key observation that the *cause obtain* sense can be learned earlier than its frequency might predict. These results suggest that in order to accurately model this pattern in acquisition, we would need either a different type of information, or a different approach to learning. The model's excellent performance on the *must* sense shows that given suitable features, a DPMM is capable of learning an infrequent sense very well. Accordingly, our focus will be on determining the appropriate features.

Detailed semantic distinctions may be difficult to capture automatically, particularly given the assumption of a child's limited linguistic development. One option would be to include argument fillers in addition to syntactic slot features. Such an approach may offer additional developmental plausibility: children may associate verb senses with specific lexical items before they are able to access more general argument types. However, selectional preferences have been shown to be largely ineffective for type-level verb clustering (Joanis et al., 2008), although they may offer some benefit at the token level of our approach. Results from sentence processing experiments show that the semantic category of a subject can bias an adult reader's interpretation of a verb sense, which in turn predicts argument structure (Hare, Elman, Tabaczynski, & McRae, 2009). We may be able to incorporate this effect by using a word space model for NP arguments (Baroni, Lenci, & Onnis, 2007), or perhaps a simple animacy feature (Joanis et al., 2008).

## Conclusions and future directions

In this paper, we use token-level clustering methods to simulate children's acquisition of the senses of a polysemous verb. With the English verb *get* as a case study, we use a Bayesian framework to cluster usages of *get* drawn from a corpus of child-directed speech. We show that simple, automatically extracted syntactic slot features give reasonably accurate clustering results on the senses of *get*. However, these features are insufficient to account for the order of acquisition of polysemy as observed in children. Children do not show a consistent correlation between frequency and age of acquisition. We show that even with a more detailed, manually-annotated feature set, clustering results in the model do not reflect child behaviour. This suggests that for a token-level clustering method to accurately model this pattern in child language acquisition, it would need either a different kind of information or a substantially different learning mechanism.

One other possible explanation for children's apparent ease in learning certain infrequent verb senses is that children may generalize meaning from other similar verbs. For example, children may recognize that the ditransitive use of *get*, as in *I got you a sandwich*, is similar to that of other benefactive verbs like *buy*, *catch*, or *find*. This class of verbs is systematically used in both causative and non-causative forms, and children may recognize this regularity and use it to their advantage. Children are known to generalize verb argument structure and its associated semantic knowledge across many different verbs, and computational simulations suggest that this is an important factor in children's ability to learn verbs with such ease (Alishahi & Stevenson, 2008). Accordingly, our ongoing work investigates the ways that developing argument structure knowledge affects the acquisition of polysemy across a range of early verbs.

## Acknowledgments

We are deeply grateful to Andrea Berez and Stefan Gries for sharing their annotated corpus data. We thank the anonymous reviewers for their helpful comments, and we gratefully acknowledge the financial support of NSERC of Canada and the University of Toronto.

## References

Alishahi, A., & Stevenson, S. (2008). A probabilistic model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789-834.

Baroni, M., Lenci, A., & Onnis, L. (2007). ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proc. of ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition.*

Berez, A. L., & Gries, S. Th. (2009). In defense of corpus-based methods: a behavioral profile analysis of polysemous get in English. In *Proc. of 24th Northwest Linguistics Conference* (Vol. 27, p. 157-66).

Clark, E. V. (1996). Early verbs, event-types, and inflections. In C. E. Johnson & J. H. V. Gilbert (Eds.), *Children's language* (Vol. 9, p. 61-73). Lawrence Erlbaum.

Fisher, C. (2002). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, *5*(1), 55-64.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Gries, S. Th. (2006). Corpus-based methods and cognitive semantics: the many meanings of to run. In S. Th. Gries & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis* (p. 57-99). New York: Mouton de Gruyter.

Hare, M., Elman, J. L., Tabaczynski, T., & McRae, K. (2009). The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension. *Cognitive Science*, *33*(4), 610-628.

Israel, M. (in press). How children get constructions. In M. Fried & J.-O. Ostman (Eds.), *Pragmatics in construction grammar and frame semantics.* John Benjamins.

Joanis, E., Stevenson, S., & James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, *14*(3), 337-367.

Korhonen, A., Krymolowski, Y., & Marx, Z. (2003). Clustering polysemic subcategorization frame distributions semantically. In *Proc. of ACL2003* (p. 64-71).

Lapata, M., & Brew, C. (2004). Verb class disambiguation using informative priors. *Comp. Ling.*, *30*(1), 45-73.

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. *KDD '99*, 16–22.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, p. 281-197).

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed., Vol. 2). Lawrence Erlbaum.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*(2), 249-265.

Nerlich, B., Todd, Z., & Clarke, D. D. (2003). Emerging patterns and evolving polysemies: the acquisition of get between four and ten years. In B. Nerlich, Z. Todd, V. Herman, & D. D. Clarke (Eds.), *Polysemy: Flexible patterns of meaning in mind and language.* Mouton de Gruyter.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proc. ACL-2007 Wkshp on Cognitive Aspects of Computational Language Acquisition.*

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proc. of the 28th annual conference of the Cognitive Science Society*.

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, *6*, 79-111.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2002). Going, going, gone: the acquisition of the verb 'go'. *Journal of Child Language*, *29*, 783-811.

Versley, Y. (2008). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In *Proc. ESSLLI wkshp. on distributional lexical semantics*.

Vlachos, A., Korhonen, A., & Ghahramani, Z. (2009). Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proc. of EACL workshop on geometrical models of natural language semantics.*