Evaluating Hierarchies of Verb Argument Structure with Hierarchical Clustering

Jesse Mu and **Joshua K. Hartshorne**
Department of Psychology
Boston College
Chestnut Hill, MA
{muj,hartshoj}@bc.edu

Timothy J. O'Donnell
Department of Linguistics
McGill University
Montreal, Canada
timothy.odonnell@mcgill.ca

Abstract

Verbs can only be used with a few specific arrangements of their arguments (*syntactic frames*). Most theorists note that verbs can be organized into a hierarchy of *verb classes* based on the frames they admit. Here we show that such a hierarchy is objectively well-supported by the patterns of verbs and frames in English, since a systematic hierarchical clustering algorithm converges on the same structure as the handcrafted taxonomy of VerbNet, a broad-coverage verb lexicon. We also show that the hierarchies capture meaningful psychological dimensions of generalization by predicting novel verb coercions by human participants. We discuss limitations of a simple hierarchical representation and suggest similar approaches for identifying the representations underpinning verb argument structure.

1 Introduction

Why can Sally *like to read* but not **appreciate to read*? Key to the grammar of sentences are verbs and the arguments with which they appear. How children learn the constraints that govern the ways verbs and arguments combine is a central question in language acquisition.

Theorists have long noted that verbs can be organized into *classes* based on their syntactic constructions and the events they express (see Levin and Rappaport Hovav, 2005 for review). Verb classes are included in most theories of argument structure acquisition, whether as first class objects (Perfors et al., 2010) or mere epiphenomena of other claims about the structure of form-meaning mappings (Pinker, 1989; Goldberg, 1995).

Most theories also propose further structure between classes. One common assumption is that verb argument structure can be at least partially described by a hierarchy: Each verb belongs to a class, which itself may belong to a number of broader *superclasses*.

While many theories predict more complex structure (e.g. cross-cutting categories; Levin and Rappaport Hovav, 2005), providing (psycho)linguistic evidence for a simple hierarchy of verbs is an important starting point for investigating more complex theories. VerbNet (Kipper et al., 2008), the largest English verb argument structure resource,[1] organizes verbs and classes into a shallow hierarchy, but its structure has been handcrafted incrementally over time (starting with seminal work by Levin, 1993). On the other hand, recently-developed, state-of-the-art machine learning methods offer a unique alternative approach to constructing such a hierarchy.

In this paper, we first conduct a broad-coverage analysis of how verbs might be hierarchically arranged by comparing VerbNet's handcrafted hierarchy to structure systematically inferred by a Bayesian hierarchical clustering algorithm. We find that the two arrive at similar structure, thus substantiating both methods (i.e. intuition vs. clustering) and the common hierarchy they find.

Second, we investigate the psychological validity of this representation: if classes capture meaningful dimensions of generalization, one would intuit that a verb in a class should behave more similarly to verbs in nearby classes than distant classes according to some measure of "distance". Indeed, this kind of assumption plays an important role

[1] VerbNet combines many important expert-crafted verb resources into a single database, and thus is used in diverse NLP applications including semantic parsing (Giuglea and Moschitti, 2006), natural language inference (Palmer et al., 2009), and information extraction (Maynard et al., 2009).

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 997–1002
Copenhagen, Denmark, September 7–11, 2017.

in theoretical (Suttle and Goldberg, 2011; Pinker, 1989) and empirical (Ambridge et al., 2011) work. We thus ask human participants to rate the compatibility of a wide range of existing verbs in attested and unattested syntactic frames. We find that such coercions are indeed predicted by a hierarchical taxonomy of verbs.

2 Related work

There is a substantial literature from both the NLP and psycholinguistics communities on unsupervised learning of verb classes from corpora and other resources (e.g. Reichart and Korhonen, 2013; Vlachos et al., 2009; Sun et al., 2008; Joanis and Stevenson, 2003) and computational cognitive models of argument structure acquisition (e.g. Barak et al., 2016; Ambridge and Blything, 2015; Barak et al., 2014; Parisien and Stevenson, 2010; Perfors et al., 2010), respectively.

Our work differs in several ways. First, we do not consider the basic problem of learning verb classes from semantic or syntactic primitives (cf. Sun et al., 2008) or verb usages extracted from corpora; instead, we examine what higher-level structure is implied by the gold-standard catalog of already-clustered verbs and syntactic frames in VerbNet. Second, we do not attempt to model incremental learning (cf. Parisien and Stevenson, 2010) or instantiate a specific theory (cf. Ambridge and Blything, 2015). Rather, we conduct an at-scale investigation of verb argument structure through cluster analysis.

3 Discovering structure via clustering

VerbNet suggests a shallow and disconnected hierarchy of verbs, with lower-level *subclasses* of verbs that take the exact same frames, broader *standard classes*, and at the top, 101 unrelated *superclasses* (Figure 1a). There is a broad assumption of weaker relations between members of higher-level classes than lower-level classes.

We compared this to the hierarchy obtained from Bayesian Hierarchical Clustering (BHC; Heller and Ghahramani, 2005) implemented in R by Savage et al. (2009), a state-of-the-art agglomerative clustering method that can be seen as a bottom-up approximation to a Dirichlet Process Mixture Model. Unlike traditional hierarchical clustering algorithms, BHC uses Bayesian hypothesis testing to merge subtrees: at each proposed merge, BHC evaluates the probability p that

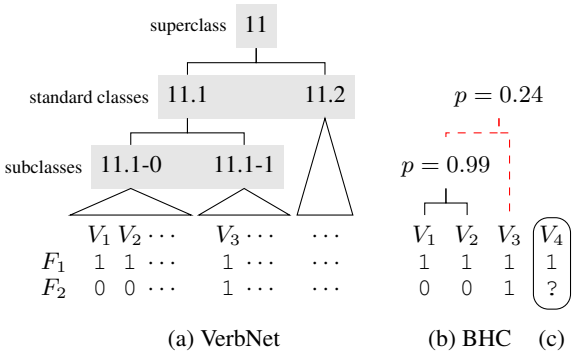

Figure 1: (a) Simplified VerbNet hierarchy, depicting a superclass, standard classes, subclasses, and toy verbs V_i and frames F_i. (b) We train BHC on the frame data $\mathcal{D}$. Dotted lines are merges BHC prefers not to make ($p < 0.5$). To obtain a flat clustering, the tree is cut at nodes where $p < 0.5$ and each subtree is a cluster. (c) Using BHC to evaluate $P(V_4 \text{ admits } F_2 \mid V_4 \text{ admits } F_1, \mathcal{D})$.

the data are generated from a single probabilistic model, rather than two or more different models consistent with the subtrees.[2] Crucially, nodes with probability $p < 0.5$ are merges that BHC prefers *not* to make; the tree can be cut at these nodes to obtain a flat clustering (Figure 1b), which can then be compared to VerbNet.

3.1 Data

As input to BHC, we used VerbNet's comprehensive set of verb-frame combinations. VerbNet v3.2[3] can be represented as a 6334 verb $\times$ n frame binary matrix, with 1s in cells with attested verb-frame pairs (Figure 1a). Thus, each verb is represented as a binary vector of frames.

The number of frames n depends on what semantic and syntactic annotations are considered to be part of the frame. VerbNet includes 3 kinds of annotations: selectional restrictions on arguments, thematic roles, and prepositional literals (Figure 2). For this paper, we included selectional restrictions and thematic roles, resulting in 1613 frames. These annotations made it easiest to produce experimental stimuli in Section 4, although our analysis produced similar results across the other possible frame encodings (see Appendix A).

[2] In our case, the assumed generative model is a set of independent Beta-Bernoulli models predicting the probability of occurrence of each frame, with the priors found by optimizing the marginal likelihood of the overall model. For full details of the algorithm, see Heller and Ghahramani (2005) and Savage et al. (2009).

[3] `verbs.colorado.edu/verb-index`

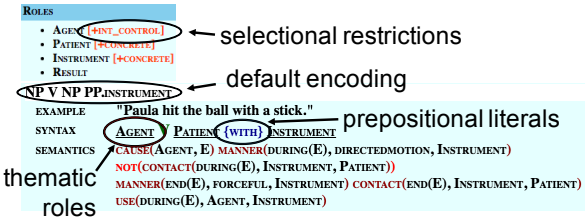

Figure 2: Information associated with a VerbNet frame entry.

3.2 Evaluation

Here we evaluated the extent to which BHC converged on VerbNet's structure at low (sub and standard classes) and high levels (superclasses).

Comparing flat clusterings with H and C First, we obtained the flat clustering from BHC (Figure 1b) and asked how it compared to VerbNet. Here, we used homogeneity (H) and completeness (C), entropy-based measures of clustering similarity analogous to precision and recall in binary classification (Rosenberg and Hirschberg, 2007). Treating VerbNet classes as ground truth, $H = 1$ indicates that every BHC cluster contains only members of a single VerbNet class. $C = 1$ indicates that members of a VerbNet class are always assigned to the same BHC cluster. The worst case for both is 0.

H and C have different meanings depending on what we consider to be VerbNet's flat ground truth classes. We consider ground truth classes across the levels of VerbNet granularity: low-level subclasses ($H_{\text{sub}}, C_{\text{sub}}$), standard classes ($H_{\text{standard}}, C_{\text{standard}}$), and superclasses ($H_{\text{super}}, C_{\text{super}}$) (Table 1).

Table 1: Homogeneity and completeness. Random baselines are mean statistics across 1000 clusterings made by uniformly sampling a BHC cluster for each verb. C_{sub} is trivially 1, since members of VerbNet subclasses have identical features and were always grouped into the same class by BHC.

Statistic	Granularity	BHC	Random
H	H_{super}	.88	.31
	H_{standard}	.88	.19
	H_{sub}	.83	.34
C	C_{super}	.72	.31
	C_{standard}	.99	.14
	C_{sub}	1	.37

The important comparison is with superclasses, for which both H and C were high. This indicates that BHC clusters rarely included verbs from multiple VerbNet superclasses ($H_{\text{super}} = .88$) and rarely split verbs from the same VerbNet superclass into different BHC clusters ($C_{\text{super}} = .72$).

Tanglegram While H and C focus on the size and membership of two clustering solutions, *tanglegrams* (Huson and Scornavacca, 2012) allow a more general visualization and comparison of two hierarchies. Using the heuristic of Scornavacca et al. (2011),[4] we drew the *optimal tanglegram* of VerbNet and BHC, where the two trees are drawn such that lines connect common leaves and the number of intersections made by these lines is minimized. We computed the *entanglement* of the tanglegram by normalizing the number of intersections to the 0–1 interval by dividing by the worst case; this is a holistic measure of the similarity of the hierarchies (Galili, 2015).

The tanglegram (Figure 3) shows that qualitatively, much of VerbNet's structure aligns well between the trees. We observed an entanglement of 0.20, compared to a random baseline of 0.66.

3.3 Discussion

The high H and C (Table 1) and low entanglement (Figure 3) suggest that both VerbNet's handcrafted hierarchical taxonomy and the one systematically created by BHC converge on similar results. Interestingly, both methods result in a fairly shallow hierarchy with many unrelated subtrees. This suggests that while small clusters of verbs are highly related, the principles governing verb argument structure are relatively narrow and do not generalize across more than a small subset of verbs. Alternatively, it could suggest that a hierarchical taxonomy is too simple to fully capture argument structure patterns.

4 Human coercion judgments

We next evaluated the hierarchies for their ability to account for human generalization. Researchers often test generalization along a specific dimension through extension to novel verbs ("wug tests"; Ambridge et al., 2013; Pinker, 1989). While this works well for studies of specific phenomena, it is difficult to deploy in a large study like ours, where we do not have hypotheses

[4] dendroscope.org

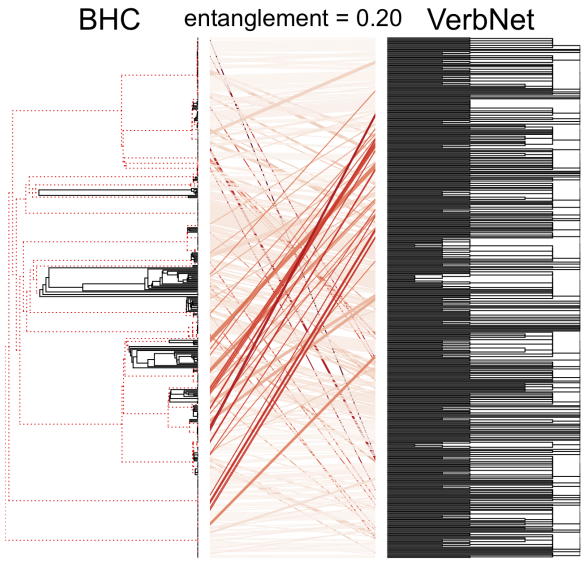

Figure 3: Tanglegram. We prune leaves (verbs) from BHC and VerbNet so that each leaf here represents a VerbNet subclass. Dotted lines are merges BHC prefers not to make. Connectors are colored by shades of red indicating worse alignment, as measured by vertical distance traveled. The vast majority of lines are light-colored, indicating strong alignment.

about what drives generalization language-wide. Thus, we assessed generalization through a coercion task, asking whether speakers are more likely to extend a known verb to a unattested frame if the frame is attested for verbs in a closely-related class. This matches a common theoretical claim that verbs are attracted to the frames of similar verbs, with the notion of similarity varying by theory (Ambridge et al., 2011; Suttle and Goldberg, 2011).

4.1 Predicting verb-frame coercion

VerbNet makes straightforward coarse predictions. For any syntactic frame, we grouped verbs into 3 categories: *Exact*, if the verb can take the frame; *Sibling*, if one of the verb's super or subclasses can take the frame; and *None* otherwise. Conversely, as a Bayesian probabilistic model, BHC defines a predictive distribution on new data. We were interested in whether this precision resulted in better fit, so we also tested BHC: for any verb and frame, we can evaluate the posterior probability that the verb admits the frame of interest while conditioning on the verb's other frames (Figure 1c; see Appendix B for details).

Table 2: 2 sampled frames and their corresponding sentence templates, each with 3 example verbs and predicted compatibilities. To form the stimuli, each verb is placed into the sentence template, e.g. *Beyond the place arose the thing*.

Frame	Sentence	Verb	VN[1]	BHC[2]
BEYOND NP.LOCATION V NP.THEME	Beyond the place V-ed the thing	arise	E	6.1
		stretch	S	−7.6
		assume	N	−9.1
NP.THEME V THROUGH NP.LOCATION	He V-ed through the place	hum	E	6.1
		motor	S	−6.9
		regard	N	−7.3

[1] VerbNet. E = Exact; S = Sibling; N = None
[2] log odds $P(\text{verb takes frame} \mid \text{verb's other frames}, \mathcal{D})$

4.2 Materials and methods

We sampled 10 frames and 10 verbs for each frame, resulting in 100 verb-frame pairs. To control for possible verb frequency effects (Braine and Brooks, 1995), we ensured there was no significant correlation between the predicted compatibility of a verb-frame pair and the Brown corpus (Kučera and Francis, 1967) frequency of the verb ($r = 0.13, p = 0.17$). We then converted verb-frame pairs into sentence stimuli, which required that we choose nouns to represent NPs in frames. We chose the most generic noun compatible with the thematic role restriction, if present. For example, for NP.AGENT, we used a generic name, and for NP.LOCATION, we used *place*. Example stimuli are located in Table 2.

We recruited 50 native English speakers from Mechanical Turk. For each sentence, participants judged the grammaticality of the sentence on a Likert scale, from 1 ("not at all") to 5 ("perfect").

4.3 Results and discussion

First, we noticed that all verbs in some frames received consistently low coercion judgments (< 3). For example, while the verb *fly* and the frame THERE V NP.THEME FOR NP.LOCATION is attested (Exact), *There flew a thing for the place* received a mean judgment of 2.4. We translated judgments so that the mean judgments across verbs for each frame was average (3), to examine the *relative* effects of coercing verbs into frames.

Figure 4a shows that VerbNet's 3 categories predict differences in the mean coercion ratings of verb-frame pairs ($F = 43.46,\ p < 0.001$). Notably, there was a significant difference between the means of the unattested categories (Sibling vs.

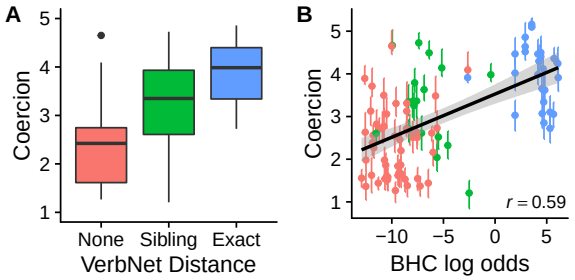

Figure 4: (a) distribution of mean coercion judgments for verb-frame pairs in the 3 VerbNet categories. (b) correlation between the same mean coercion judgments and BHC posterior predictive predictions, colored by VerbNet category. Error bars are bootstrapped 95% confidence intervals.

None; $t = 3.55$, $p < 0.01$). While there was a high correlation between the judgments and BHC predictions (Figure 4b; $r = 0.59$), BHC's hierarchy did not significantly improve fit to the data.

These results provide additional psychological evidence for the effects associated with VerbNet's coarse distinctions: for unattested verb-frame pairs, participants tend to assign a higher compatibility rating when the verb has sibling VerbNet classes that can take the frame. However, the range of compatibility judgments is highly variable across all three categories, and BHC's finer-grained predictions fail to account for much of this variability. Given the similarity of BHC to VerbNet's hierarchy, this result is unsurprising.

5 General discussion

We presented converging evidence that a shallow hierarchy of verbs (1) is well supported by the distribution of verbs and syntactic frames in language, since VerbNet's hand-crafted hierarchy and a systematic unsupervised learner (BHC) reach similar results; and (2) captures important features of verb argument structure by predicting human generalization intuitions in a coercion task.

Of course, it is clear from the variability of our coercion data that a simple hierarchy is not a sufficiently sophisticated representation of argument structure to fully explain language-wide coercion. However, our novel computational framework (unsupervised learning on VerbNet data) opens up many potentially fruitful avenues for providing language-wide evidence for argument structure hypotheses. The lack of broad-coverage predictions is often a limitation of work in this area (see Section 2).

Sophisticated machine learning models that make the assumptions proposed by richer theories of argument structure and can operate at VerbNet scale are only recently coming into fruition. For example, since some theories argue for a cross-categorization of verbs and argument structures (Levin and Rappaport Hovav, 2005), using models that find such a (possibly hierarchical) cross-categorization (e.g. Mansinghka et al., 2016; Li and Shafto, 2011) is a particularly interesting avenue for further exploration.

Acknowledgments

We thank anonymous reviewers for helpful comments and suggestions. JM is supported by an undergraduate research fellowship from the Boston College Psychology Department and a Barry M. Goldwater scholarship.

References

Ben Ambridge and Ryan P Blything. 2015. A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of Child Language*, 43(6):1245–1276.

Ben Ambridge, Julian M Pine, and Caroline F Rowland. 2011. Children use verb semantics to retreat from overgeneralization errors: A novel verb grammaticality judgment study. *Cognitive Linguistics*, 22(2).

Ben Ambridge, Julian M Pine, Caroline F Rowland, Franklin Chang, and Amy Bidgood. 2013. The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):47–62.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2014. Learning verb classes in an incremental model. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–45.

Libby Barak, Adele E. Goldberg, and Suzanne Stevenson. 2016. Comparing computational cognitive models of generalization in a language acquisition task. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 96–106.

Martin D S Braine and Patricia J Brooks. 1995. Verb argument structure and the problem of avoiding an overgeneral grammar. In Michael Tomasello and William E C Merriman, editors, *Beyond names for things: Young children's acquisition of verbs*, pages 352–376. Erlbaum, Hillsdale, NJ.

Tal Galili. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936.

Adele E Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Cambridge, MA.

Katherine A Heller and Zoubin Ghahramani. 2005. Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 297–304.

Daniel H Huson and Celine Scornavacca. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, 61(6):1061–1067.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–170.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University Of Chicago Press, Chicago, IL.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument realization: Research surveys in linguistics*. Cambridge University Press.

Dazhuo Li and Patrick Shafto. 2011. Bayesian hierarchical cross-clustering. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 443–451.

Vikash Mansinghka, Patrick Shafto, Eric Jonas, Cap Petschulat, Max Gasner, and Joshua B. Tenenbaum. 2016. CrossCat: A fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *Journal of Machine Learning Research*, 17(138):1–49.

Diana Maynard, Adam Funk, and Wim Peters. 2009. Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proceedings of the 2009 International Conference on Ontology Patterns*, pages 39–52.

Martha Palmer, Jena D Hwang, Susan Windisch Brown, Karin Kipper Schuler, and Arrick Lanfranchi. 2009. Leveraging lexical resources for the detection of event relations. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 81–87.

Christopher Parisien and Suzanne Stevenson. 2010. Learning verb alternations in a usage-based Bayesian model. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 2674–2679.

Amy Perfors, Joshua B Tenenbaum, and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(3):607–642.

Steven Pinker. 1989. *Learnability and Cognition*. The MIT Press, Cambridge, MA.

Roi Reichart and Anna Korhonen. 2013. Improved lexical acquisition through DPP-based verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 862–872.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Richard S Savage, Katherine Heller, Yang Xu, Zoubin Ghahramani, William M Truman, Murray Grant, Katherine J Denby, and David L Wild. 2009. R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics*, 10(1):1.

Celine Scornavacca, Franziska Zickmann, and Daniel H Huson. 2011. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*, 27(13):i248–i256.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Automatic classification of english verbs using rich syntactic features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Laura Suttle and Adele E Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics*, 49(6):1237–1269.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet Process Mixture Models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82.