Event Categorization beyond Verb Senses

Aron Marvel
University at Buffalo
609 Baldy Hall
Buffalo, NY 14260, USA
aronmarv@buffalo.edu

Jean-Pierre Koenig
University at Buffalo
609 Baldy Hall
Buffalo, NY 14260, USA
jpkoenig@buffalo.edu

Abstract

Verb senses are often assumed to distinguish among different conceptual event categories. However, senses misrepresent the number of event categories expressed both within and across languages and event categories may be "named" by more than a word, i.e. a multi-word expression. Determining the nature and number of event categories in an event description requires an understanding of the parameters relevant for categorization. We propose a set of parameters for use in creating a Gold Standard of event categories and apply them to a corpus sample of 2000 sentences across 10 verbs. In doing so, we find an asymmetry between subjects and direct objects in their contributions to distinguishing event categories. We then explore methods of automating event categorization to approximate our Gold Standard through the use of hierarchical clustering and Latent Semantic Analysis (Deerwester et al., 1990).

1 Introduction

A word form is associated with one or more senses, each of which may denote a distinct conceptual category. This association is many-to-many; one word may have many senses, while different words may also share the same sense. Additionally, just as two different words may denote the same concept, so may a sequence of words. Consider the sentences in (1).

(1) a. The officer entered the building.
b. The officer went into the building.

How many concepts do these sentences contain? Probably *officer* and *building* each count as one and so does *enter*. But it is difficult to justify labeling *enter* as a single concept while treating *go* and *into* as separate. *Enter* and *go into* seem to denote the same concept, the first by means of a single word and the second through a multi-word expression (MWE). The mapping between concept and lexicalization becomes a real problem for AI reasoning systems. These systems often translate natural language input into a lingua franca, such as the HPSG representation used by SNePS (Shapiro & Rappaport, 1992), and there is no clear way for them to know when they have encountered a MWE that represents a single concept.

While the sentences in (1) indicate that a single conceptual category may span syntactic boundaries and involve different verbs, it is also possible for distinct conceptual categories to be denoted using a single verb sense as in (2).

(2) a. The senator raised a glass in celebration.
b. The crane raised the car out of the water.

Both (2a) and (2b) employ the same sense of *raised*[1] but denote very different categories of events. In prototypical contexts, (2a) describes a toast, while (2b) describes the extraction of a large object. The events described in (2) differ in a number of ways, among them duration, complexity,

[1] To determine whether two uses of a word instantiate the same sense, we use the American Heritage Dictionary (AHD), which features several notable linguists among its contributors and consultants.

Proceedings of NAACL-HLT 2015, pages 77–86,
Denver, Colorado, May 31 – June 5, 2015.

available inferences, and the types of agents involved. Further, several inferences one can draw from (2a) arise non-compositionally, i.e. cannot be inferred from just the meaning of the parts and the sentence's syntactic structure.

What is crucial for our purposes in (2) is that the two distinct event categories described by the sentences are differentiated by information outside of the verb sense. Recognition of this fact prompts the question of what kinds of information beyond verb sense are relevant for differentiating event categories, as well as how to distinguish between MWEs that denote distinct event categories and those that do not. In this paper, we explore these problems and develop a new method of automatically categorizing event descriptions.

The paper is structured as follows. Section 2 briefly discusses the limitations of lexical approaches to event categorization and outlines an alternative approach that takes into account clausal constituents beyond the verb. In Section 3, we propose a set of six general parameters by which categories of events may be distinguished beyond the verb sense. Those parameters are applied to a categorization task in Section 4 using a sample of corpus sentences for 10 different verbs. In Section 5, we describe an attempt to automate the sorting task using relatedness measures from Latent Semantic Analysis in combination with hierarchical clustering.

2 Event categories as MWEs

Lexical approaches to event categorization, i.e. those that only rely upon the verb, encounter significant problems stemming from the arbitrariness of lexicalization both within and across languages. Within a language, the same conceptual event category may be expressed by a verb or a verb plus non-verbal expressions as in (1). Confining event categorization to the verb may additionally miss important differences between event categories as in (2). Additionally, languages differ both in the sizes of their verbal lexicons and in the number of senses assigned to each verb. The average adult English speaker knows approximately 4,000 verbs (Koenig et al. 2003), each of which has on average three (COBUILD, pc, 2006) or four (WordNet Statistics, 2015) senses. Under the assumption that verb senses approximate event categories, this results in a total of 12,000 - 16,000 distinct event categories. Speakers of a language such as Wagiman, a northern Australian language, have an inventory of only about 500 verbal expressions, 90% of which have only a single recorded sense (Wilson, 1999). The upshot of only using verb senses to distinguish event categories would be the claim that speakers of Wagiman are capable of (linguistically) distinguishing only 4% of the event categories distinguished by speakers of English – an implausible statistic.

Wagiman speakers achieve parity with speakers of other languages by combining verbal expressions to create what Wilson calls 'complex predicates': the English word *watch* translates to a combination of two words in Wagiman: the word *nanda*, meaning 'to see', from a closed class of basic verbs, and the word *letta*, meaning 'to look', from an open class of verbal expressions called coverbs. Wagiman verb-coverb combinations provide an example of a multi-word expression in one language serving the purpose of a single-word expression in another. This phenomenon has received due attention within the MWE literature (see, e.g., Sag et al., 2001; Villavicencio, 2007), though most often as it relates to idiom translation. It has also received attention in the typological literature, e.g. in discussion of 'verb-framed' vs. 'satellite-framed' languages, the former of which express motion path as part of verb meaning and the latter of which express it verb-externally through 'satellite' phrases (Talmy, 1985a).

In addition to the above motivations, the problems we investigate are related to a large body of research devoted to selectional preferences, including efforts from both psycholinguistics (e.g. McRae et al. 1998, 2005) and computational modeling (e.g. Erk & Padó 2008, Lenci 2011). These efforts are primarily concerned with measuring the sensitivity of people and NLP systems to distributional properties of verbs, though some, such as FrameNet (Baker et al. 1998) and Corpus Pattern Analysis (Hanks 2004), do flesh out the boundaries within these distributions more fully. Our aim here is to explore the parameters that underpin divisions within these distributions. Many of the parameters involve non-compositional meaning components and thus benefit from an understanding of event descriptions as MWEs. We propose here that all events, not just idiomatic and institutionalized phrases, may be categorized at the level of multi-word expressions.

Consider again the two uses of *raised* in (2). In comparing the two different categories of raising events, we may look beyond the verb and investigate the contribution of other parts of the clause: How is the kind of raising involved in raising a glass distinct from the kind involved in raising a car? How is the kind of raising a senator does predictably different from the kind of raising a crane does? We propose partial answers to these questions and motivate them with examples in the next section.

3 Parameters for event categorization

What follows is an attempt to extract from both previous research and common sense a set of general parameters by which event categories may be distinguished beyond the level of the verb sense. We should be clear at the outset that the following parameters are not to be taken as complete, but rather as a subset of dimensions of experience that are available for event categorization. Examples are drawn from our corpus sample and, importantly, share the same verb sense as per the AHD.

3.1 Complexity

Event complexity often refers to the number of sub-events represented in the semantics of a verb (see e.g. Dowty 1979). However, such accounts ignore the contribution of event participants in influencing event categorization. Consider the two uses of *sell* in (3).

(3) a. He refused to sell any of his antiques.
b. The support staff sells their expertise to the community beyond the school.

Though the sense of *sell* remains constant between (3a) and (3b), selling done by a support staff to a community is likely to include a larger total number of sub-events than selling done by or to an individual. This sub-event information, though, is only available to language users when they combine verb information with information they glean from disparate parts of the clause.

In addition to the number of sub-events, complexity includes the relations among sub-events and the participants within them. Sensitivity to this kind of complexity has been found as a general trait in infants and adults. Infants have a harder time processing complex relations like containment than they do processing simpler relations such as interposition (Baillargeon & Wang, 2002). Additionally, recent research suggests that adult speakers are sensitive to event complexity in their willingness to violate iconicity expectations during narrative discourse (Dery & Koenig, in press). We therefore consider both the quantitative and qualitative aspects of complexity to be relevant for event categorization.

3.2 Time scale

The parameter of time scale includes binary distinctions such as events that are permanent rather than temporary or bounded rather than unbounded, but also includes differences in duration along a continuum, e.g. events that occur in the space of one second in comparison to events that happen over the course of several months, years or millennia. An example of a difference along a continuum is found in (4).

(4) a. Royal Bank of Scotland bought Bank Worcester at the end of 1990.
b. I stopped at a bar just long enough to buy two cheese rolls.

While buying a couple of cheese rolls as in (4b) takes only a moment, the consolidation of two banks as in (4a) generally does not.

The linguistics literature on event structure is rife with binary time scale distinctions. Events are often discussed in terms of whether or not they are telic, bounded (Verkuyl, 1972), culminating (Moens & Steedman, 1988), or delimited (Tenny, 1987). In addition to the latter theoretical support, experimental evidence for sensitivity to binary time scale distinctions may be found within both the acquisition literature — e.g. children's marking those distinctions even when their languages do not (Clark, 2001 & 2003) — and studies of adult narrative discourse, where situations with inherent endpoints bias narrators towards different types of continuations (Dery & Koenig, in press). In establishing Gold Standard categories for our data, we consider both the binary and continuous dimensions of temporal distinctions described above.

3.3 Agent type

We use the term 'agent' here in a broad sense; while characteristics such as animacy and volition are prototypical, they are not required. Agents are distinguishable from one another according to such properties as whether they are individuals or groups, animate or inanimate, physical or abstract, etc., and the type of agent exerts an influence on event interpretation and categorization. Example (5) presents two sentences that involve distinct types of agents.

(5) a. A Genoese fleet rescued the city.
b. Archaeologists rescue information about the past before it is destroyed.

From differences in agent type it is possible to predict that the rescuing events described are different categories of rescuing. (5a) describes a large concerted operation involving many individuals, machinery, national resources, extensive planning and so on, while (5b) involves none of these things.

Evidence for the parameter of agent type also comes from reading time experiments and experiments using event-related potentials (ERPs) in which participants show sensitivity to the combination of agent and verb when processing event patients (Bicknell et al. 2010).

3.4 Sociocultural salience

A factor that, to our knowledge, has been entirely missed or ignored in the literature on event categorization – perhaps because it is so difficult to quantify – is social or cultural salience. Yet it is uncontestable that some objects, characteristics, or events are set apart from others because of their importance within the practices of a community. (6a) differs from (6b) because the event category described, book-borrowing, has become institutionalized to the extent that we have public buildings devoted solely to facilitating that practice.

(6) a. The room is for pupils to borrow books.
b. Can you borrow an iron for me?

To our knowledge, the borrowing of irons has yet to achieve such lofty status on the public agenda. The salience of any particular category of event will vary across populations of language users, as well as across languages, to the extent that language and cultural practices co-vary.

3.5 Inferences

As additional information combines with that of the verb, more inferences become available, and many of these inferences may be relevant to event categorization. Consider the examples in (7).

(7) a. She adjusted the scarf to cover the bruises forming on her neck.
b. The children covered their eyes and turned away as the needle went in.

In (7a), the agent presumably desires to hide a bruise from the sight of others, while in (7b), the inference is not that the children are trying to prevent others from seeing their eyes; rather, they are trying to keep themselves from witnessing something unpleasant. Such inferences are often unavailable compositionally. World knowledge associated with the description conveyed by the verb *and* its arguments must be added to the compositional meaning before such inferences can be drawn.

3.6 Specific motion sequence

Certain events are characterized by a sequence of motions that set them apart from events that can be performed in any number of ways. These events may often be described as actions performed according to a recognizable motor program put into action by the event participant(s). Though distinctions along this dimension are admittedly rarer than those made via many of the other parameters, they do exist, as the examples in (8) show.

(8) a. Charlery pulled the ball behind Halsall.
b. The General shouted at his men to pull the barricade down.

The category of event described in (8a) requires a specific motion in which the leg is moved forward over the ball, the toe is brought down into contact with the top of the ball, and the leg and ball are pulled back together; pulling down a barricade as in (8b), however, may be accomplished through a variety of unspecified means.

4 Experiment 1: Manual categorization

While the above parameters for distinguishing event categories may sound plausible, there is no guarantee that their application will result in a division of event descriptions that is equally plausible. In order to make such a determination, each of the authors categorized the same large set of event descriptions by hand. The results of this process were then used as a Gold Standard for subsequent automation of the categorization task. For the purposes of this exploratory study, we elected to limit our investigation to variation within the head noun included in subjects and direct objects for a given verb, while recognizing that information from other portions of the clause may play a role in event categorization. The methods we employ are easily extendible to include other constituents such as prepositional phrases.

4.1 Materials and procedure

Through the use of the software package Tgrep2 (Rohde, 2005), a full list of sentences containing the following 10 verbs was obtained from the British National Corpus (BNC): *bake*, *borrow*, *buy*, *cover*, *deliver*, *frighten*, *immerse*, *pull*, *rescue*, and *sell*. The total sample comprised approximately 43,000 sentences. The sentences in the sample were then randomized and a list of the first 100 sentences with unique subjects was compiled for each verb. Items with pronominal subjects were excluded because without access to an anaphoric or deictic referent, pronouns contribute relatively little information beyond that contributed by the verb. Items with subjects that were proper names, which similarly contribute little or no information useful for categorization, were also excluded. Lastly, sentences with ambiguous or incorrect parses were removed from the sample by hand. Sentences in the sample were then randomized once more and another list for each verb was compiled containing the first 100 sentences with unique direct objects. The product of this process was 20 lists — two for each of ten verbs — totaling 1602 sentences.[2]

Because pronouns constitute a much larger proportion of subjects than direct objects, our decision to exclude pronouns may artificially inflate the contribution of subjects (vs. direct objects) to the diversity of event categories, though this is primarily an issue only with small sample sizes. In total, pronouns constituted 49.64% of subjects and 19.51% of direct objects for the verbs included in our sample and proper names constituted 12.23% of subjects and 3.18% of direct objects.

Each of the authors independently categorized each list of sample sentences. The event categories discovered were discussed until consensus was reached.[3] The resulting event categories were then compared against verb senses obtained from the American Heritage Dictionary (AHD) in order to determine the efficacy of verb senses in capturing the event category distinctions we found. Dictionary senses that were not found in any of our sample sentences were ignored.

4.2 Results

The AHD provides an average of 3.8 senses per verb in our list.[4] Categorization by application of our parameters provided an average of 16.5 event categories per verb. Of these categories, 62% came from the direct object sentence lists, suggesting that there is an asymmetry between subjects and direct objects in distinguishing among event categories (p = .009, n = 165 categories). A comparison of AHD senses to event categories is shown in Table 1.[5]

4.3 Discussion

Several regularities arose during the categorization process. The direct object lists almost always contributed larger numbers of categories than the subject lists. In some respects, this finding is not unexpected. Agents generally play a minor role in characterizing events. Intuitively, "A man raised a finger" could be paraphrased as a finger-raising event, but not as a man-raising event (as opposed to a woman-raising event). We also found that

[2] Not all lists were 100 items in length, simply because some verbs had fewer than 100 valid BNC results after filtering; while we do not explicitly address these cases here, the proper n value for each list was used in all analyses.

[3] Because stable categories were not yet available (the task being to create them), inter-rater agreement was not measured. It is worth noting, however, that our categorizations overlapped to a surprisingly high degree.

[4] The total average number of senses per verb, including those senses not found in our sample sentences, was 5.7 for our 10 verbs.

[5] Event category counts are summed for each verb from subject and direct object lists.

Verb	AHD senses	Categories
bake	2	10
borrow	2	18
buy	3	18
cover	8	30
deliver	7	17
frighten	2	14
immerse	3	8
pull	6	24
rescue	1	13
sell	4	13
Average	**3.8**	**16.5**

Table 1. Comparison of AHD senses to event categories discovered by application of the parameters discussed in Section 3.

some of our proposed parameters were more frequently applicable than others. Unsurprisingly, agent type played a major role in distinguishing event categories within the subject sentence list. It most often followed from differences in plurality (*an uncle borrowed* vs. *the crew borrowed*), animacy (*an uncle borrowed* vs. *an atom borrowed*) and abstractness (*the crew borrowed* vs. *the agenda borrowed*). Complexity, sociocultural salience and inferences also played a large part, while time scale and specific motion sequence tended to take a back seat in both subject and direct object lists.

One further finding not directly evident from the reported results concerns the verb *frighten*. This verb belongs to a relatively small class of psych verbs known as 'object-experiencer' verbs, where one sees a reversal of what otherwise occurs in subject and direct object positions – e.g., a verb like *watch* may occur in *Anne watched the storm*, but *frighten* may only occur in the reverse pattern *the storm frightened Anne*. This reversal was found in our corpus data. The general asymmetry in the number of pronominal subjects and direct objects we observed did not apply to *frighten*, and proper names were found in direct object position more than twice as often for *frighten* as they were for other verbs. If it is world knowledge about what the verb and its arguments describe that is informing event categorization, one would expect that, when the kinds of items typically found in direct object position are instead found in subject position and vice versa, the asymmetry in the relative importance of subjects and objects in distinguishing event categories is also reversed. This is exactly what we found: 64% of the *frighten* categories were distinguished by the combination of verb and subject. The results for *frighten* suggest that the asymmetry between subjects and objects is not due to grammatical function, lending support to our claim that the parameters outlined in Section 3 are independent of a language's morphosyntax.

One final finding of our first experiment is worth noting and bears directly on the design of Experiment 2. In general, the more semantically similar to one another any pair of a verb's subjects or direct objects were, the more likely the events described by the combination of those items with the verb were to be put in the same event category. For example, the events described by *covered their hands* and *covered their feet* are more likely to be in the same category than either is to be in a category with *covered their city*, simply because *hands* and *feet* are more semantically similar to one another than either is to *city*. We adopt this finding as an assumption for automating event categorization in Experiment 2.

5 Experiment 2: LSA categorization

Categorizing even a relatively short list for only ten verbs turned out to be quite difficult and time-consuming. It is therefore desirable to find a dynamic and automatic way to categorize any event description as it is encountered. Below we describe a first try at such automation, using Latent Semantic Analysis (LSA) and hierarchical clustering to approximate our Gold Standard categories.

LSA is a method for evaluating semantic similarity from corpora containing collections of independent documents. It requires the creation of large, sparse matrices which track each word's frequency of co-occurrence with each other word within each document. The matrix is reduced to a target number of only the most salient dimensions, usually between 50 and 400, and within the resulting semantic space it is possible to locate each word as a vector (see Deerwester et al., 1990 for a detailed description). The upshot of this process is that those words which occur together in the same documents most often (and whose frequent companions also occur together most often) are considered highly related and will usually occur near each other in the semantic space. LSA predictions matched scores of non-native college applicants in

TOEFL tests of word similarity (Landauer et al., 1998).

5.1 Materials and Procedure

Through the application of latent semantic analysis to a 400-dimensional semantic space created from the British National Corpus, pair-wise relatedness value were calculated for each subject list and each direct object list. The result was approximately $\sum_{i=1}^{100} i$ = 5050 relatedness values for each list. The `hclust` command in R (R Core Team, 2015) was then used to construct an average-linkage dendrogram from the half matrix containing each list of relatedness values. Though the full 100-item dendrograms cannot fit a page in a readable form, a slice from the *borrow* direct object list is included here as Figure 1 for illustrative purposes.

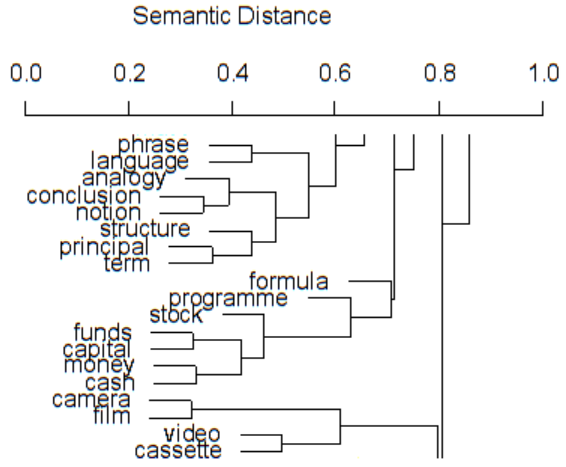

Figure 1. A section of the dendrogram created by using LSA semantic distance values to group direct objects of the verb *borrow*.

At a glance, three distinct categories are visible in Figure 1: a category of language-related items, a category of currency-related items, and a category containing videography-related items. Dendrograms are built from the bottom up (from left to right in Figure 1) by combining the most closely related branches at each step, eventually fusing the final two clusters into one unified tree. Using R's `cutree` command, this process can be reversed by counting splits from the top down until the number of categories identified in the Gold Standard categorization is reached – for the *borrow* direct objects list, that number was 13. Each list's full dendrogram was deconstructed in this way.

Precision and recall were obtained as they are for V-measures (Rosenberg and Hirschberg, 2007). For each subject or direct object in its respective list, we found the set **H** of all other words that had been assigned to the same category by LSA. The cardinality of this set represents the total number of *hypothesized* items in that word's category. We then found the set **A** of all words that had been assigned to the same category in the Gold Standard. The cardinality of this set represents the total number of *actual* items in that word's category. Thirdly, we found the intersection of the latter two sets **H**∩**A**. The cardinality of this set represents the total number of items correctly categorized by the automated categorization.

Precision ($p = |\mathbf{H}\cap\mathbf{A}|/|\mathbf{H}|$) and recall ($r = |\mathbf{H}\cap\mathbf{A}|/|\mathbf{A}|$) values were then calculated for each item and combined for an *F*-score that is their harmonic mean ($F = 2pr/(p + r)$). Finally, 100 random categorizations were performed for each list as a measure of comparison.

5.2 Results

The average LSA and randomized *F*-scores for each list type are reported in Table 2.[6]

List	p LSA	r LSA	F LSA	F rand	Ratio
Subj	40%	80%	.53	.39	1.38
DO	35%	66%	.46	.32	1.46
Overall	**38%**	**73%**	**.50**	**.35**	**1.42**

Table 2. *F*-scores for LSA categories, compared to *F*-scores for randomized categories. Ratios represent how much better than chance LSA categorization performed.

The LSA automated categorization resulted in an average of 42% more accurate categorization than that obtained by random categorization.

[6] Average *F*-scores are weighted by list length, i.e. those lists significantly shorter than 100 items – specifically lists for *bake* and *immerse* in our sample – were given proportionally less weight in calculating overall averages.

5.3 Discussion

The combination of high recall and low precision suggests that the automated categorization tends to lump a large portion of each list into only a few categories, populating the remaining categories with only a small number of outliers. Our categorization when creating the Gold Standard, in contrast, tended to distribute items more evenly over event categories. The imbalance turns out to be a consequence of the particular clustering method used in creating the LSA categories – in this case, the average linkage method. Some methods (e.g. the Ward method) instead favor increased precision over recall. In our tests, recall-biased methods invariably resulted in better F-measures.[7]

Looking at the differences in category members within Gold Standard and LSA results may provide insight into both where LSA fails and where alternative parameters may have escaped our notice. The *bake* object list yields several such exemplars. In creating our Gold Standard, we categorized baking events according to such criteria as whether or not the baked item requires preparation (e.g. making and rolling dough, etc.), which adds to the complexity of the baking event, and whether the item undergoes a transformation in the baking process (e.g., dough becomes bread, but a potato remains a potato). The LSA categorization, in contrast, appeared to reflect ethnic/cultural cuisine categories rather than processes undergone by the materials involved: the cluster containing *soufflé*, *aubergine*, *fillet* and *flan* was separated from that containing *potato, pie* and *cake*. This makes sense when one considers that the relatedness measures used by the LSA are obtained from co-occurrence of words within documents – and recipes, from which many of the baking event clauses were extracted, are often found in documents that focus on a specific kind of cuisine. It is worth stressing that this difference in categorization is not simply an indication of the limitations of LSA. Rather, it brings to light an important dimension of categorization that was not considered in our Gold Standard; baking events may quite plausibly be divided into French baking, American baking, etc. It is possible that in this instance we simply missed differences in sociocultural salience (the fourth parameter in Section 3) that stem from the role that baking plays in cultural nutrition.

We also found reflexes of the asymmetry between subjects and objects within LSA relatedness measures. Average relatedness among direct objects for a given verb was significantly higher than relatedness among subjects for seven of the eight verbs listed in the results. The one verb for which this did not hold was *frighten*, where we expected and saw a reversal in number of categories discovered when sorting by hand. When *frighten* is excluded, inter-object relatedness is on average 35% higher than inter-subject relatedness. In other words, the direct objects for a verb tend to be more closely related to one another than the subjects of that verb are. The exact nature of the relationship between this asymmetry in relatedness scores and the asymmetry in contribution to category formation remains to be determined.

6 Conclusion

Preliminary categorizations suggest that language users are capable of much finer-grained event categorization than that provided at the level of verb senses (at a ratio of over 4:1) and that these event categories are associated with multi-word expressions which include the verb plus direct object/subject head. Using the methods described in this paper, it is possible to automate this finer-grained level of event categorization to some degree. With respect to both of these findings, there is an asymmetry between English subjects and direct objects in their contribution to categorization – the combination of direct objects and verbs accounts for a greater share of category distinctions than the combination of subjects with verbs. This asymmetry is purely conceptual, independent of any theoretical assumptions regarding order of syntactic composition, and is reflected in LSA relatedness measures.

We are at the time of writing conducting experiments with naïve speakers to norm our Gold Standard categorization and assess the independent contribution of different parameters in event categorization. The contribution of information other than the subject and direct object also deserves to be explored in more detail and the analysis should be expanded both to languages beyond English.

[7] Methods tested in order of improvement over random categorization were average linkage (42%), single linkage (41%), McQuitty (33%), complete linkage (18%) and Ward (7%). Note that single linkage prefers 'lumping' to a greater degree than average linkage, but results in slightly less improvement.

Additionally, LSA is only one source of relatedness measures among many; it competes with various WordNet algorithms, mutual information measures, and newer predictive measures (see e.g. Baroni et al. 2014). Though one might expect a high correlation among these measures, it turns out that very often the correlation is surprisingly low, and thus one could conceivably obtain very different categories depending on the method used to measure semantic similarity (Maki et al., 2004). It may be that some methods result in relatedness scores that better approximate human categorization than others, and these alternatives deserve exploration.

References

Baillargeon, R. & Wang, S. (2002). Event categorization in infancy. *Trends in Cognitive Sciences, 6*, 85-93.

Baker, C., Fillmore, C. & Lowe, J. (1998). The Berkeley FrameNet project. In *Proceedings of COLING-ACL 1998*, pp. 86-90.

Baroni, M., Dinu, G. & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238-247.

Bicknell et al. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63*, 489-505.

Clark, E. V. (2001). Emergent categories in first language acquisition. In M. Bowerman & S.C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 379-405). Cambridge: Cambridge University Press.

Clark, E.V. (2003). *First language acquisition*. New York: Cambridge University Press.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391-407.

Dery, J. & Koenig, J.P. (in press). A Narrative-Expectation-Based Approach to Temporal Update in Discourse Comprehension. *Discourse Processes*, 00: 1-26.

Dowty, D. (1979). *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Erk, K. & Padó, S. (2008). A structured vector space model forword meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 897-906.

Evans, G. (1980). Pronouns. *Linguistic Inquiry*, 11: 337-362.

Hanks, P. (2004). Corpus Pattern Analysis. In *Proceedings of the 2004 Conference of the European Association for Lexicography (EURALEX)*, pp. 87-97.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2&3).

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd ACL Workshop on Cognitive Modeling and Computational Linguistics*, pp. 58-66.

Maki, W., McKinley, L., Thompson, A. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers, 36* (3): 421-431.

McRae, K., Spivey-Knowlton, M. & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*: 283–312.

McRae, K., Hare, M., Elman, J. & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition, 33*(7): 1174–1184.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/

Rohde, Douglas (2005). Tgrep2 [Computer software]. Department of Brain and Cognitive Science, Massachusetts Institute of Technology. Retrieved March 19, 2014. Available from http://tedlab.mit.edu/~dr/Tgrep2/

Rosenberg, A. & Hirschberg, J. (2007). Vmeasure: 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), pp. 410–420.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2001). Multiword expressions: A pain in the neck for NLP. In *Proceedings from 3rd International Conference on Intelligent Text Processing and Computational Linguistics*: CICLing.

Shapiro, S. C., & Rapaport, W. J. 1992. The sneps family. *Computers & Mathematics with Applications*, 23, 243-275.

Talmy, L. (1985a). Force Dynamics in language and thought. In *Papers from the Regional Meetings, Chicago Linguistic Society*, 21, 293–337.

Talmy, L. (1985b). Lexicalization patterns: Semantic structure in lexical form. In T. Shopen (Ed.). *Language typology and syntactic description: Vol. 3. Grammatical categories and the lexicon*, pp. 57–149. Cambridge: Cambridge University Press.

Tenny, C. (1987). *Grammaticalizing aspect and affectedness*. Doctoral dissertation, MIT.

Verkuyl, H.J. (1972). *On the compositional nature of the aspects*. Dordrecht: Reidel.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), pp. 1034–1043.

Wilson, S., and Center for the Study of Language and Information (U.S.). (1999). *Coverbs and Complex Predicates in Wagiman*. Stanford, Calif: CSLI Publications.

WordNet Statistics. Princeton University. 2015. < https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>