DiSCo 2009

Distributional Semantics
beyond Concrete Concepts

CogSci 2009 Workshop
July 29th 2009
Amsterdam

Digital proceedings

Preface

In the last decade, corpus-based distributional models of semantic similarity and association have slipped into the mainstream of cognitive science and computational linguistics. On the basis of the contexts in which a word is used, they claim to capture certain aspects of word meaning and human semantic space organization. In computational linguistics, these models have been used to automatically retrieve synonyms (Lin, 1998) or to find the multiple senses of a word (Schütze, 1998), among other tasks. In cognitive science, they have been applied to the modelling of semantic priming (Burgess, Livesay, & Lund, 1998; Landauer & Dumais, 1997), semantic dyslexia (Buchanan, Burgess, & Lund, 1996), categorization and prototypicality (Louwerse, Hu, Cai, Ventura, & Jeuniaux, 2005), and many other phenomena. Yet, despite their claims to model human language behaviour, relatively little is known about the precise relationship between these distributional models and human semantic knowledge. While they offer a credible account of (thematic or general) similarity of unary predicates such as concrete nouns, the question remains if and how more complex knowledge can be modelled using distributional information. This workshop therefore wants to focus on new challenges to distributional approaches that lie beyond the traditional modelling of concrete concepts.

In our call for papers, we defined three specific challenges. A first challenge is the modelling of verb meaning. The results of a related workshop at the European Summer School in Logic, Language and Information (ESSLLI-2008), showed that verb clustering is a much more difficult task than noun clustering. It is an open question what precise information distributional models of verb semantics should take into account, and there is a lack of an uncontroversial Gold Standard.

A second challenge is the modelling of different aspects of semantics. Distributional models are typically used for the discovery of semantic relations like similarity (between plane and airplane) or association (between plane and airport). We may ask ourselves what other information, apart from these two types of relations can be collected from linguistic data.

A third challenge is the combination of different types of data. Distributional models of lexical semantics can only make a partial claim to the modelling of human semantic cognition. After all, when children learn the meaning of words, they probably make use of much more information than just the linguistic context that the words occur in. Andrews, Vigliocco, and Vinson (in press) argue that in order to arrive at realistic models of semantic cognition, the distributional approach therefore has to be complemented with experiential data that captures our experience with the physical world.

Central to these challenges is the relationship of distributional models to human semantic cognition. What are the main differences between the conceptualization of verbs and that of nouns? What aspects of human property spaces are found in large corpora, and how? And how does distributional learning relate to other types of learning? These are the main questions we wanted to address in our workshop.

We wanted to tackle these questions in particular by attracting researchers from both cognitive science and computational linguistics that have been concerned with distributional models. We believe that uniting these two perspectives will lead to a fruitful discussion about the present and future of distributional semantics. In this way, we would like to reconcile different directions in research on distributional semantics, and outline relevant paths for future research.

Looking at the workshop's papers (bundled together in these electronic proceedings) and our selection of invited speakers, we believe this workshop will succeed in doing so: Parisien and Stevenson, as well as Wagner *et al.* address the question of verb polysemy using probabilistic clustering models for sense discrimination that jointly consider syntactic and selectional preferences of the verb tokens in question; Shutova and Teufel show that it is possible to use distributional techniques to group metonymic interpretations of nouns in a context; and finally, Boussidan *et al.* use distributional clustering techniques to investigate the correlation between sub-morphemic elements (*phonaesthemes*, i.e., non-morphemic phoneme strings that are believed to influence meaning, and Proto-Indo-European roots) and the meaning of words.

Several of our invited contributions present research correlating distributional similarity with other experimental data to gain further insights on human representations: Baroni *et al.* compare distributional models with neurophysiological data from conceptual stimuli, Huettig uses distributional measures to predict subject behaviour in a visual world paradigm, and Vigliocco *et al.* show evidence that corpus-based collocation models and experiential features based on interaction with the outer world are complementary and can be used to yield a superior combined representation. Heylen investigates the differences between syntactic (subcategorization-based) and word-space (collocate-based) models for the semantics of verbs; Sahlgren presents an exploration on higher-order relationships in word space models.

Last but not least, we would like to thank our program committee – Marco Baroni (University of Trento), Yves Bestgen (Université catholique de Louvain), Simon Dennis (Ohio State University), Simon de Deyne (University of Leuven), Katrin Erk (University of Texas at Austin), Stefan Evert (University of Osnabrück), Dirk Geeraerts (University of Leuven), Peter Hastings (DePaul University), Falk Huettig (Max Planck Institute for Psycholinguistics), Alessandro Lenci (University of Pisa), Diana McCarthy (University of Sussex), Danielle McNamara (University of Memphis), Sebastian Padó (University of Stuttgart), Chris Parisien (University of Toronto) and Suzanne Stevenson (University of Toronto).

Wishing you a fruitful, interesting and swinging DiSCo workshop,

Yves Peirsman	Yannick Versley	Tim Van de Cruys
University of Leuven	University of Trento	University of Groningen

Contents

Part I

Invited Speakers: Abstracts and Papers

Integrating Experiential and Linguistic information in Semantic Representation

Gabriella Vigliocco and **Mark Andrews** and **David Vinson**
{g.vigliocco|m.andrews|d.vinson}@ucl.ac.uk
University College London

Abstract

We present an account of semantic representation that focuses on distinct types of information from which word meanings, across all domains of knowledge, can be learned. In particular, we argue that there are at least two major types of information from which we learn word meanings.

The first is what we call experiential information. This is data derived both from our sensory-motor interactions with the outside world, as well as from our experience of own inner states, particularly our emotions.

The second type of information is language-based. In particular, it is derived from the general linguistic context in which words appear. In our hypothesis semantic representations come about as a combination of these two types of information.

In order to assess this view, we implemented and tested against behavioural data, models of semantic representation based on experiential-only, linguistic-only and combined data showing that combining the two sources provides a better fit to the data.

On the use of distributional models of semantic space to investigate human cognition

Falk Huettig (Falk.Huettig@mpi.nl)
Max Planck Institute for Psycholinguistics, Post Box 310,
6500 AH Nijmegen, The Netherlands

Abstract

Huettig *et al.* (2006) demonstrated that corpus-based measures of word semantics predict language-mediated eye-movements in the visual world. These data, in conjunction with the evidence from other tasks, is strong evidence for the psychological validity of corpus-based semantic similarity measures. But can corpus-based distributional models be more than just good measures of semantic similarity? I briefly describe two research areas for which distributional models seem promising: word evolution and the influence of culture and language on semantic systems.

Keywords: cognition; eye movements; overt attention; semantic space

The psychological validity of high-dimensional models of semantic space has been assessed using a variety of methods such as semantic similarity ratings (e.g. Rastle, Davis, Marslen-Wilson, and Tyler, 2000), semantic interference effects in picture naming (Vigliocco, Vinson, Lewis, and Garrett, 2004), simulating the standardized synonym choice test taken by non-native speakers of English who apply for admission to US universities (Landauer and Dumais, 1997), semantic categorization tasks (Siakaluk, Buchanan, and Westbury, 2003), simulations of semantic and associative priming effects (Lund *et al.*, 1995; McDonald and Lowe, 1998) and dyslexia (Buchanan, Burgess, and Lund, 1996).

1 Distributional models predict language-mediated eye-movements

We (Huettig, Quinlan, McDonald, and Altmann, 2006) chose a different approach. We explored whether overt attention to a depicted object can be predicted from the degree of semantic/contextual similarity it shares with a spoken word as indexed by models of high-dimensional semantic space. To investigate this we used a visual-world eye-tracking paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy, 1995). In the visual world paradigm, participants are presented with an array of visual objects, usually while they listen to spoken utterances. This paradigm provides fine-grained eye movement measures of ongoing cognitive processing, in the form of fixations to different positions in the visual display over time. For instance, Huettig and Altmann (2005) investigated whether semantic properties of individual lexical items can direct eye movements towards objects in the visual field. Participants were presented with a visual display containing four pictures of common objects. During the course of a trial a spoken sentence was presented to the participant and the participant's eye movements were tracked as the sentence unfolded. We found that participants directed overt attention immediately towards a picture of an object such as a trumpet when a semantically-related but non-associated target word (e.g., 'piano'), acoustically unfolded. Importantly, the probability of fixating a semantic competitor correlated with a similarity measure derived from semantic feature norms (Cree and McRae, 2003). In the Huettig et al. (2006) study we examined the psychological validity of two corpus-based semantic distance measures: LSA (Latent Semantic Analysis) (Landauer and Dumais, 1997) and Contextual Similarity (McDonald, 2000). If such measures reflect the psychological nature of semantic representations of words, then such models should predict fixation behavior in the visual world paradigm.

1.1 Method

Participants 60 native British English speakers from the University of York student community participated.

Stimuli We selected 26 target-competitor pairs of words. On each trial in the experiment, participants were presented with a visual display containing line drawings of four spatially distinct objects whilst a spoken sentence was concurrently presented. Position of eye gaze was measured as the sentence unfolded. Each spoken sentence contained a critical word (such as 'toaster'), and performance was examined in a target condition and a competitor condition, which differed from the actual spoken word presented in the place of the critical word. In the target condition the spoken sentence contained the target word (e.g. '*First, the man disagreed somewhat, but then he noticed the toaster and appreciated that it was useful*'). In the competitor condition the target word was replaced by a semantic competitor word ('*corkscrew*') (e.g. '*First, the man disagreed somewhat, but then he noticed the corkscrew and appreciated that it was useful*'). The spoken sentences were identical in both conditions up to the point in the sentence when the critical word (e.g. '*toaster*' or '*corkscrew*') was heard.

The visual displays were composed of four quadrants and each contained a target object, e.g. a toaster, and three unrelated distractor objects with one object in each quadrant. The visual displays were identical in both conditions. The names of the pictures within a display each started with a different phoneme so that no phonological (cohort) competitors were present. In addition, the pictures were matched on picture naming agreement, image agreement, familiarity, visual complexity, word frequency of the corresponding name, and similarity of visual form.

We predicted that on hearing the critical words participants would shift their overt attention to the target object in both conditions.

Design The experiment was of a within-participant design, with each participant receiving a random order of 26 experimental and 26 filler trials. Half of the experimental trials were target trials and half were competitor trials. On the target trials one of the pictures was a depiction of the target spoken word. In contrast, for the 13 experimental items in the competitor condition the critical words did not match the target objects. For these trials the target picture (e.g. the toaster) was only semantically-related to the spoken competitor word (e.g. '*corkscrew*'). All of the fillers included a fully matching target object. Therefore across all trials in the experiment 75% of the 52 trials included a fully matching target object (e.g. hearing '*toaster*' and seeing a toaster in the display). Two counter-balanced groups were tested, in which the assignment of items to the target and competitor conditions was switched. Filler items were the same for both groups.

Procedure Participants were told that they should listen to the sentences carefully, that they could look at whatever they wanted to, but not to take their eyes off the screen throughout the experiment (see Huettig and McQueen, 2007, for discussion of this procedure).

2 Results

Gaze position was categorized by quadrant. The fixation proportions at the acoustic onset and offset of the critical words were of main interest. The acoustic *onset* of the critical word is of interest so as to assess whether there were any biases in attention before information from the critical word became available. In turn, the acoustic *offset* of the target word reflects the point at which the entire critical word has been heard by the participants.

We used difference scores to analyse the data. Difference scores were calculated by subtracting $p(\text{fix distractor})$ from $p(\text{fix targ})$. $p(\text{fix distractor})$ was averaged across the 3 distractor pictures before participant/item confidence intervals were calculated. At the acoustic *onset* of the critical words (e.g. '*toaster*' or '*corkscrew*') there were no reliable differences in looks to the corresponding target and distractor pictures. At the acoustic offset of the critical words however we found a reliable bias in overt attention to the target object in both conditions. As the critical word (i.e. '*toaster*' in the target condition and '*corkscrew*' in the competitor condition) acoustically unfolded, participants shifted their attention towards the target picture. Therefore participants showed a bias to attend to the named target in the target condition. Critically, participants also showed a bias to attend to the target object in the competitor condition, even though it was not the target but a semantically-related object (the semantic competitor) that was mentioned.

The primary goal of the study though was to explore the degree to which corpus-based measures predict fixation behavior. We used two semantic distance measures: (i) Contextual Similarity (McDonald, 2000), and (ii) LSA (Landauer and Dumais, 1997, semantic space: general reading up to first year college, 300 factors). If these corpus-based semantic distance measures reflect the nature of semantic representations of words, then such models should predict fixation behavior in the visual world paradigm with some degree of accuracy. We predicted that as the degree of similarity between target and competitor increases, the size of the competitor effect would increase. To examine this possibility, various correlational analyses were carried out. P(fix targ) at offset of the acoustic competitor word in the competitor condition correlated moderately with Contextual Similarity (Pearson correlation, $r = 0.58$, $p = 0.002$) and with LSA ($r = 0.42$, $p = 0.033$) but not with the visual similarity ratings ($r = 0.07$, $p > 0.1$). Importantly, there was no corresponding reliable correlation at the onset of the competitor words (Contextual Similarity: $r = 0.15$, $p > 0.1$; LSA: $r = 0.14$, $p > 0.1$; visual similarity: $r = 0.26$, $p > 0.1$) which shows that the cognitive processing of the competitor words resulted in the subsequent shifts in overt attention.

We also carried out various logistic regression analyses. Each participant's eye movement record for every trial was scored as to whether or not they had fixated the target picture at the offset time point. Only the competitor condition trials were used in the regression analysis. We computed separate regression equations for each participant and tested whether these regression coefficients differed reliably from zero as described by Lorch and Myers (1990). 13 of the 60 participants were removed from the analysis because fixations on the critical target objects occurred on less than four items for these participants (i.e. for these participants a competitor effect occurred on less than four items). Regression coefficients for p(fix targ) computed separately for the semantic measures for each subject differed reliably for Contextual Similarity (one-tailed, one-sample t-test, $t(1,46) = 3.534$, $p < .001$) and LSA (one-tailed, one-sample t-test, $t(1,46) = 2.917$, $p = .003$). When regression equations were computed simultaneously for Contextual Similarity, LSA, and visual similarity for each subject only Contextual Similarity (one-tailed, one-sample t-test, $t(1,46) = 1.826$, $p = .037$) remained reliable whereas LSA (one-tailed, one-sample t-test, $t(1,46) = -1.121$, $p > .1$) and visual similarity (one-tailed, one-sample t-test, $t(1,46) = 0.048$, $p > .1$) coefficients did not differ reliably from zero.

In sum, our study revealed that corpus-based measures of word semantics are good predictors of eye fixation behavior in the visual world. These data, in conjunction with the evidence from the other tasks mentioned above, is strong evidence for the psychological validity of corpus-based semantic similarity measures.

3 Can we use distributional models of semantic space to investigate more complex phenomena in human cognition?

The interesting question that arises is whether corpus-based distributional models are more than just good measures of semantic similarity. I will highlight two areas of research where the use of sophisticated distributional models of semantic space will be likely to be particularly fruitful in future. The first concerns the evolution of language, the second the influence of culture and language on semantic systems.

Using distributional models to investigate the evolution of language Pagel, Atkinson, and Meade (2007) have recently shown that word frequency in modern language use predicts their rate of replacement by other words over thousands of years of Indo-European language evolution. They found that high-frequency words evolve at slower rates than low-frequency words, i.e. word frequency influences the rate of lexical evolution. They estimated that word frequency accounts for about 50% of variation in the rates of replacement. Pagel (2009) has recently also proposed that 'strength or size' of connections in high-dimensional semantic space may influence the rate of word evolution. "For example, hasta is the Sanskrit word for hand, but among Latin speakers it became the word for spear. The sound 'hasta' may have been saved by the cognitive connection between hand and spear" (Pagel, 2009, p.411). Distributional models thus promise to be a valuable tool to understand differing rates of word evolution.

Using distributional models to investigate culture and/or language-dependent differences in semantic space The relationship between language and cognition remains a controversial issue. There is a strong tradition of researchers who subscribe to the view that languages directly encode cognitive categories (e.g., Fodor, 1975).

Li and Gleitman (2002) for example argue that “linguistic categories and structures are more or less straightforward mappings from a pre-existing conceptual space programmed into our biological nature”(p.266). There is however a different tradition of researchers who argue that different languages and cultures impose differing conceptual constraints on semantic systems, i.e. they give rise to semantic systems that ”carve the world at quite different joints” (e.g., Evans and Levinson, in press). Some languages for instance don’t have logical connectives such as ’or’ (Tzeltal) or ’if’ (Guugu Yimithirr) or don’t have words such as ’hand’, ’leg’, ’green’, or ’blue’ (Yélî Dnye; Evans and Levinson, in press).

Majid *et al.* (2007) found that individual semantic categories of cutting and breaking differed dramatically among different languages. Speakers of Yélî Dnye, a language spoken on an isolated island of Papua New Guinea, used only three verbs (Levinson, 2007) to describe more than 60 video clips but speakers of Tzeltal, a language spoken in Mexico, used more than 50 different verbs (Brown, 2007). There is now a wealth of data documenting the enormous variation in semantic distinctions among existing languages (see Evans and Levinson, in press; and Majid and Huettig, 2008; for further discussion of cross-linguistic influences on semantic cognition).

To be able to make better use of distributional models for research on the sources of language variation and the relationship between language and cognition two issues must be addressed. First, sophisticated distributional models of semantic space exist so far only for English and a few other mostly western languages. The consequence is that we lack sophisticated comparisons of semantic spaces across natural languages. There is a need for models in a greater variety of languages (including more distantly related non-western languages). Second, individual models are vulnerable to criticism that similarity measures, parameters, and training texts selected, etc. are not appropriate for one particular model. There is a need for more work on the validity of cross-language/cross-model comparisons.

References

Brown, P. (2007). “She had just cut/broken off her head”: Cutting and breaking verbs in Tzeltal. *Cognitive Linguistics*, 18, 19–30.

Buchanan, L., Burgess, C., and Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain and Cognition*, 32, 111–114.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6, 813-839.

Cree, G. S., and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132, 163-201.

Evans, N. and Levinson, S. (in press). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*.

Fodor, J. A. (1975) *The language of thought*. Thomas Y. Crowell.

Huettig, F. and Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, 96, B23-B32.

Huettig, F. and McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *Journal of Memory and Language*, 54, 460-482.

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121, 65-80.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Levinson, S. C. (2007) Cut and break verbs in Yélî Dnye, the Papuan language of Rossel Island. *Cognitive Linguistics*, 18, 207–18.

Li, P. W. and Gleitman, L. (2002) Turning the tables: language and spatial reasoning. *Cognition* 83, 265–294.

Lorch, R. F., and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 149-157.

Lund, K., Burgess, C., and Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660-665). Mahwah, NJ: Erlbaum.

Majid, A., Bowerman, M., van Staden, M., and Boster, J. S. (2007) The semantic categories of cutting and breaking: A crosslinguistic perspective. Cognitive Linguistics, 18, *133–52*.

Majid, A. and Huettig, F. (2008). A crosslinguistic perspective on semantic cognition [Commentary on Prcis of semantic cognition: A parallel distributed approach by Timothy T. Rogers and James L. McClelland]. *Behavioral and Brain Sciences*, 31, 720-721.

McDonald, S. A. (2000). *Environmental determinants of lexical processing effort*. Unpublished doctoral dissertation, University of Edinburgh, Scotland. Retrieved December 10, 2004, from http://www.inf.ed.ac.uk/publications /thesis/online/IP000007.pdf

McDonald, S. A., and Lowe, W. (1998). Modeling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 675-680). Mahwah, NJ: Erlbaum.

Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10, 405-415.

Pagel, M., Atkinson, Q. D., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449, 717-719.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., and Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, 15, 507-537.

Siakaluk, P. D., Buchanan, L., and Westbury, C. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory and Cognition*, 31, 100-113.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Vigliocco, G., Vinson, D. P., Lewis, W., and Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48, 422-488.

EEG Responds to Conceptual Stimuli and Corpus Semantics

Marco Baroni and **Brian Murphy** and **Massimo Poesio**
{marco.baroni|brian.murphy|massimo.poesio}@unitn.it
Università di Trento

Abstract

Corpus-based semantic models have proven effective in a number of empirical tasks (Sahlgren, 2006) and there is increasing interest in looking for non-trivial similarities between the knowledge extracted by such models and human semantic memory (e.g., Schulte im Walde, 2008).

A particularly direct way to test such potential correlations is by comparing model-generated representations and neural activation patterns in response to conceptual stimuli. Mitchell et al. (2008) have shown that corpus-extracted models of semantic knowledge can predict fMRI activation patterns. Following up on this groundbreaking study, we report experiments showing that the EEG signal can also be predicted using corpus-based features.

Moreover, we explore automated feature selection/reduction techniques (Mitchell and colleagues used manually picked lists), and we compare different corpus-based models. Our best results are currently obtained with a simple fixed-context-window model trained on newspaper text.

References

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meaning of nouns. *Science*, 320:1191–1195.

Sahlgren, M. (2006). Towards pertinent evaluation methodologies for word-space models. In *LREC 2006*.

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, 6(1):79–111.

Distinguishing the Similar: An analysis of the semantic distinctions captured by distributional models of verb meaning.

Kris Heylen
kris.heylen@arts.kuleuven.be
University of Leuven

As in lexical semantics in general, distributional methods have also proven a successful technique for the automatic modeling of verb meaning. However, much more than with other lexical categories, the research into verb semantics has been based on the idea that a verb's meaning is strongly linked to its syntactic behavior and more specifically, to its selectional preferences. This has led distributional methods of verb meaning to make use of two distinct types of syntactic contexts to automatically retrieve semantically similar verbs.

The first approach is in principle purely syntactical and looks at a verb's distribution over subcategorization frames, i.e. the possible combinations of syntactic verb arguments like subject, direct object, indirect object etc. This purely syntactic information can be extended with some high-level semantic information like the animacy of the verb arguments (see Schulte im Walde, 2006 for an overview). Whereas this first approach is specifically geared towards verbs and is inspired by the long linguistic research tradition on 'verb valency', the second approach is more generally applicable to all lexical categories and was mainly developed within computational linguistics. These so-called word space models use other words as context features with a specific implementation only using those context words that co-occur in a given dependency relation to the target word (see Padó and Lapata, 2007 for an overview). In this second approach, one specific context feature corresponds to one lexeme plus its syntactic relation to the target verb, whereas in the first approach, one context feature is a possible combination of syntactic arguments that a verb can govern. Whereas the first approach is mostly used to automatically induce Levin-style verb classes, the second approach is typically applied to retrieve semantic equivalents for specific verbs (but see Li and Brew, 2008 for a comparison of the two methods on the task of inducing Levin-style classes).

In this presentation we will look more closely at the kind of semantic information that is captured by these two distinct types of distributional methods for verb meaning. For a sample of 1000 highly frequent Dutch verbs we constructed the two basic models described above from an automatically parsed corpus of Dutch newspapers. In a first step, we used all of the verb-specific dependency relations covered by the parser and in a second step we reduced the number of different dependency relations to only include core arguments (excluding so-called complements). For the comparison of the models, we first looked at the overall correlation between the verb similarities calculated based on the different models and see that they, at least partially, capture comparable semantic distances. In second analysis, we zoomed in on a number of specific verbs and compared the semantic aspects captured by the different models. The models based on subcategorization frames showed a tendency to reflect a verb's aspectual properties whereas the models based on co-occurring lexemes demonstrate more strictly semantic and topical information.

References

Li, J. and Brew, C. (2008). Which are the best features for automatic verb classification. In *Proc. ACL 2008*.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32:159–194.

Chapter 5

Higher-order relations in word space

Magnus Sahlgren

Part II

Workshop Papers

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:

$$\begin{aligned} y_{ij}|\theta_{jz_i} &\sim \mathrm{Mult}(\theta_{jz_i}) \\ \theta_{jz_i}|G &\sim G \\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

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 θ_{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 θ, and α is a concentration parameter that affects how many clusters we expect to find.

In the above, G generates the parameters of the multinomial distribution (θ_{jz_i}) that in turn generates y_{ij}. Since G selects θ from across the set of clusters (*e.g.*, θ_{j1} or θ_{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 α and a *base distribution* G_0, gives a prior distribution on the number and size of the clusters as well as on the parameters θ to represent them. G_0 defines the prior distribution for θ. We set G_0 to $\mathrm{Dir}(\mathbf{1})$, 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

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

where n_k is the number of verb usages in cluster k and N is the total number of usages. Larger values of α 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 θ 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 θ), 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 θ parameters to random values drawn from a Dirichlet distribution. We iteratively update each z_i and θ_{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 θ_{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.

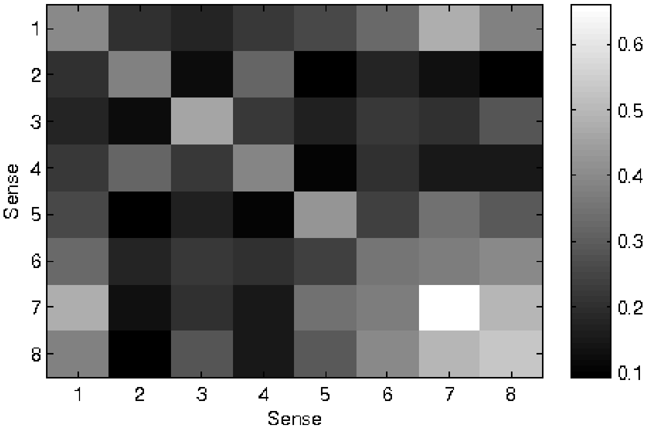

Figure 1: Likelihood of grouping usages from each pair of senses, averaged over all usages. Indices correspond to senses as in Table 3.

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.

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

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

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.*

Verb Sense Disambiguation using a Predicate-Argument-Clustering Model

Wiebke Wagner (wiebke.wagner@ims.uni-stuttgart.de)

Helmut Schmid (schmid@ims.uni-stuttgart.de)

Sabine Schulte im Walde (schulte@ims.uni-stuttgart.de)
Institute for Natural Language Processing, Azenbergstr. 12
D-70174 Stuttgart, Germany

Abstract

In this paper we present a verb sense disambiguation technique which is based on statistical clustering models which merge verbs with similar subcategorisation and selectional preferences into a cluster. The sense of a verb is disambiguated by (i) extracting the verb and its argument heads with a statistical parser from a given sentence, (ii) labeling the extracted verb-argument tuple with one or more clusters according to the clustering model, and (iii) assigning the verb to one of its possible senses based on this cluster information. Using only the cluster IDs as features, we obtained an accuracy of 57.06% which is close to the results of the best system in the Senseval-2 competition which used far more information. We also show that a generalization of the selectional preferences in terms of WordNet concepts leads to better performance due to a reduction of sparse data problems. **Keywords:** probabilistic verb clustering; verb sense disambiguation; selectional preferences.

Introduction

Word sense disambiguation has a long history (see (Agirre & Edmonds, 2006) for an overview) but still remains a core problem to many NLP applications such as message understanding, machine translation, and question answering. Especially the disambiguation of highly polysemous verbs with subtle meaning distinctions is difficult. The definition of sense inventories is also challenging, controversial, and not equally appropriate across NLP domains (Ide & Wilks, 2006).

High-performance Verb Sense Disambiguation (VSD) systems are trained on sense-tagged corpora and use a wide range of linguistic and non-linguistic features. The system described in (Chen & Palmer, 2009) e.g. employs a parser, a named entity tagger, and a pronoun resolver to extract syntactic features (voice, type of complements, complement heads), semantic features (WordNet synsets and hypernyms of complement heads), topical features (keywords occurring in the context), and local features (the two preceding and following words and their POS tags). A smoothed maximum entropy classifier disambiguates the sense based on these features. It achieved 64.6% accuracy on Senseval-2 data. Results on another data set (OntoNotes) with clearer sense distinctions came close to the inter-annotator agreement rate with 82.7%.

The high costs of manual semantic tagging motivated the development of semi-supervised methods. Stevenson and Joanis (2003) clustered verbs into Levin classes with an extensive feature space. Then they applied manual, semi-supervised and unsupervised approaches to automatic feature selection in order to reduce the 560 feature set to the relevant features. They reported a semi-supervised chosen set of features based on seed verbs as the most reliable choice. Lapata and Brew (2004) defined a simple probabilistic model with automatically identified verb frames, which generated preferences for Levin classes. This model was used for disambiguating polysemous verbs in terms of Levin classes. They showed that the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its arguments.

In this paper, we use a statistical clustering model which is trained on a large unlabelled corpus of verb argument tuples such as $\langle read, subj{:}obj, man, book \rangle$ which were extracted from a text corpus by means of a parser. The clusters provided by the model can be interpreted as 'sense labels'. However, these labels are unlikely to exactly match the senses of some independently defined sense inventory. Therefore the cluster labels must be mapped to these senses in order to use the clustering model for their disambiguation. The mapping is done by a statistical classifier which is trained on manually sense tagged text. The classifier computes the probability of each possible verb sense given the cluster labels.

The verb clustering model is based on the assumption that verbs which agree on their selectional preferences belong to a common semantic class. The two verbs *to sit* and *to lie* in Example 1 e.g. belong to a class of verbs which describe an entity placed on top of another entity.

(1) *The cat sits/lies on the sofa.*

Different readings of a verb usually differ in their argument preferences. Example 2 shows two readings of the verb *to roll* with different subcategorisation frames.

(2) *The thunder rolls. – Peter rolls the ton off the road.*

Example 3 demonstrates that also the class of arguments (weaponry vs. employee) can differentiate between verb meanings.

(3) *to fire a gun – to fire a manager*

These differences in subcategorisation and selectional preferences allow the clustering model to assign the readings of a verb to different clusters, which can then be used as evidence for verb sense disambiguation. We implemented a VSD system based on these ideas and evaluated it on Senseval-2 data[1].

[1] http://193.133.140.102/senseval2/, last visited June 2009

The Senseval-2 Data

The Senseval-2 shared task was a word sense disambiguation (WSD) competition for nouns, verbs and adjectives. In this paper, only the disambiguation of verbs is considered though. We tested our system on the English Lexical Sample task of the Senseval-2 data set, which contains 3565 verb instances in the training set and 1806 in the test set. This data comprises 29 different target verbs with 16.76 senses on average. This high polysemy rate is due to the fact that particle verb constructions such as *carry on* are subsumed under the base verb. Particle verbs are explicitly marked in the corpus. This facilitated disambiguation because it allowed the elimination of inappropriate readings. The Senseval-2 data are hand-tagged with one (sometimes two) WordNet sense keys of the pre-release WordNet version 1.7. The inter-tagger agreement (ITA) of the task was only 71.3% which can be taken as an upper bound for this task.

Description of the clustering models

We used two different statistical clustering models for verb-argument tuples which group the verbs based on their subcategorisation and selectional preferences. They are soft-clustering models and therefore able to assign a tuple to more than one cluster. The degree of membership in a given cluster is expressed by the conditional probability $p(cluster|tuple)$.

LSC

In the Latent Semantic Clustering (LSC) model (Rooth, Riezler, Prescher, Carrol, & Beil, 1999) the induction of clusters for a given verb-argument pair is based on the estimation of a probability distribution over tuples which consist of a cluster label, a verb and the argument heads. The LSC model is characterized by the following equation, where c is the cluster label, v is a verb and $a_1...a_n$ are the arguments, and $p(a|c,i)$ is the probability of the word a as the i-th argument in a tuple from cluster c.

$$p(c,v,a1,...,a_n) = p(c)p(v|c)\prod_{i=1}^{n} p(a_i|c,i) \qquad (4)$$

The cluster variable c is not observed in real data and therefore a ’hidden variable’. The LSC model assumes that the verb and the arguments are mutually independent given the cluster. In other words, it is sufficient to know that a verb belongs to some cluster c in order to predict its possible arguments. It follows that all verbs of a cluster must have similar argument preferences.

The model parameters are estimated with the EM algorithm which maximizes the likelihood of the training data consisting of verb-argument tuples without cluster information in an iterative process. After each iteration the model improves its parameters and increases the likelihood of the data. The independence assumptions mentioned above drive the clustering process because only models which approximately satisfy the independence assumptions will have a high training data likelihood. This technique is described more precisely in (Rooth et al., 1999) and (Schulte im Walde, Hying, Scheible, & Schmid, 2008). The number of clusters is predefined.

PAC

The PAC (predicate argument clustering) model (Schulte im Walde et al., 2008) is an extension of LSC. LSC considers only a fixed number of arguments from one particular subcategorisation frame, whereas PAC allows arbitrary subcategorisation frames. The tuple representation described above for LSC is augmented with a *frame argument*. The terms argument and subcategorisation frame here are used in a wider sense, since all subphrases that depend on the verb are considered as an argument phrase and belong to the subcategorisation frame; not only the obligatory ones. An example PAC tuple is given by ⟨begin, subj:obj:p:np, seller, discussion, with, buyer⟩. [2]

If f is a subcategorisation frame and n_f is the number of arguments in frame f, the PAC model is characterized by the following formula:

$$p(c,v,f,a_1,...,a_{n_f}) = p(c)p(v|c)p(f|c)\prod_{i=1}^{n_f} p(a_i|c,f,i) \qquad (5)$$

The tuple probability $p(c,read,subj{:}obj,man,book)$, for instance, is the product $p(c)p(read|c)p(subj{:}obj|c)$ $p(man|c,subj{:}obj,1)p(book|c,subj{:}obj,2)$.

Because the argument probability $p(man|c,subj{:}obj,1)$ is difficult to estimate due to sparse data problems, PAC generalizes the selectional preferences expressed in this probability distribution from words to concepts and replaces $p(man|c,subj{:}obj,1)$ by the product of a slot-specific concept probability such as $p(person|c,subj{:}obj,1)$ and a word probability such as $p(man|person)$ which is independent of the slot. The concepts are taken from a hierarchy such as WordNet. The selectional preferences of a given argument slot such as $\langle c,subj{:}obj,1\rangle$ are represented by a set of concepts which together constitute a *cut* through the WordNet hierarchy. In general, there might be more than one concept r in this set which dominates a given noun. Therefore it is necessary to sum over them:

$$p(a|c,f,i) = \sum_r p(r|c,f,i)p(a|r) \qquad (6)$$

The concept probabilities and word probabilities are not directly estimated. Instead they are derived from Markov models whose states correspond to WordNet concepts. The concept probability $p(person|c,subj{:}obj,1)$, for instance, is defined as the sum of the probabilities of all paths from *entity* to *person* in the Markov model for the slot $\langle c,subj{:}obj,1\rangle$, and the probability of a single path, in turn, is defined as the product of the state transition probabilities along that path. PAC models are trained on verb-argument tuples without cluster

[2] PP arguments contribute two elements to the frame, the preposition and the nominal head.

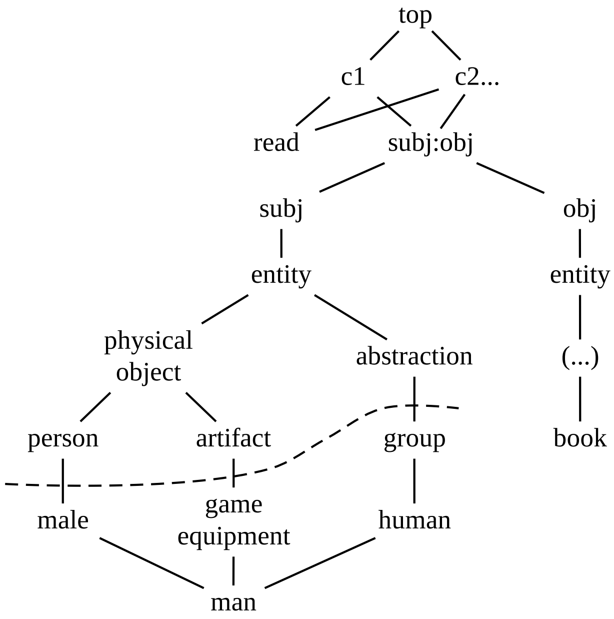

Figure 1: PAC tree of the tuple: <read, man, book>

information with a variant of the EM algorithm. Initially, the selectional preferences (SP) of the different slots only consist of the most general concept *entity*. During the training, the preferences become more specific, corresponding to a lower cut through the WordNet hierarchy. The specificity of the concepts is controlled by Minimum Description Length (MDL) pruning. In each iteration of the EM algorithm, the SP Markov models are first extended with all the hyponyms of the current terminal nodes. Then the E step and the M step of the EM training follow, and finally the resulting SP models are pruned back by eliminating all edges whose deletion decreases the total description length.

Description of the System

Our VSD system consists of two components, a clustering model (either LSC or PAC) and a classifier. It uses the verb, the subcategorisation frame, and the arguments as the only features for VSD. The clustering model is trained on the Reuters corpus[3] and learns to assign similar verbs (or actually verb readings) to the same cluster. The classifiers is trained on the Senseval-2 training corpus and learns which clusters correspond to which sense of a verb. It treats each verb separately.

Preprocessing of the Data

We parsed the Reuters corpus with the BitPar parser (Schmid, 2006) and extracted the verbs and their arguments. With the extracted tuples, we trained the verb clustering models.

The Senseval-2 corpus was also parsed with the BitPar parser but only the verbs to be disambiguated and their arguments were extracted. For each tuple, we calculated the cluster probabilities according to the verb clustering model.

$$p(c|tuple) = \frac{p(c,tuple)}{\sum_{c'} p(c',tuple)} \tag{7}$$

Cluster probabilities below a threshold of 0.1 were ignored.

Training of the Classifier

Next we used the Senseval-2 training set to train a classifier that estimates the probability of senses within a cluster. If c is a cluster and s is a sense, we first summed up the probabilities of c for any tuple that was labeled with s. This gave us the frequency of the joint occurrence of s and c.

$$f(s,c) = \sum_{tuple:sense(tuple)=s} p(c|tuple) \tag{8}$$

To get the probability of s given c, we calculated the relative frequency. The probabilities of all different senses within c therefore sum up to 1.

$$p(s|c) = \frac{f(s,c)}{\sum_s f(s,c)} \tag{9}$$

Sense Classification

The classifier assigns a sense to each tuple based on the verb, the cluster probabilities, and the sense probabilities. The most probable clusters of a tuple are obtained from the clustering model and the sense probabilities for these clusters were estimated in the training. The classifier multiplies the probability of each cluster with the probability of each sense of the cluster. The total probability of a sense for a given tuple is computed by summing over all clusters:

$$p(s|tuple) = \sum_c p(c|tuple)p(s|c) \tag{10}$$

To give an example: The cluster probabilities for the verb-argument tuple $\langle carry, subj{:}obj, man, suitcase\rangle$ might be c1=0.94, c2=0.05. The classifier would provide for c1: sense1=0.18 and sense2=0.81, whereas c2 would hold sense1=1 as a single sense. In this case, the most probable sense would be $p(s_2|tuple) = p(c_1|tuple)p(s_2|c_1) + p(c_2|tuple)p(s_2|c_2) = 0.94 * 0.81 + 0.05 * 0 = 0.76$.

In accordance with the Senseval scoring we counted each verb with an identical sense tag as a match (Kilgarriff, 2000). If no sense was found,[4] the most frequent sense (MFS) of the verb was assigned. If no MFS existed because the verb was not in the training data, we randomly chose one of the senses of the verb in WordNet1.7 and took 1 divided by the number of senses as the estimated correctness of this random decision.

[3] http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf, last visited June 2009

[4] This can be due to parsing errors or because the assigned clusters did not appear in the training data with that verb

Evaluation

The system was optimized on the training set of the English Lexical Sample task. All experiments that follow in this section are done on this data set with a tenfold cross-evaluation. We experimented with different settings of the model and of the preprocessing to find the best features.

We established a base system to explore the performance of our features. The base system uses a PAC clustering model with 50 clusters, and 100 training iterations. Additionally we compared the results to the MFS baseline which assigns all verbs to their most frequent sense.

If nouns from the verb-argument tuple were not in WordNet, we replaced them by a placeholder $\langle UNKNOWN \rangle$. Additionally we used the placeholder $\langle NONE \rangle$ when the parser failed to find the head of an argument (e.g. the subject in subject-less sentences).

For significance testing, we applied a Binomial test and considered only tuples that where classified correctly either in the base system or in the experiment system but not in both. We chose an significance threshold of 5%.

Experiments on the Data

Since the variable frame size and the conceptualisation of the arguments were an extension from LSC to PAC we aimed to discover to what extend the frames and arguments helped in the classification process. We tried to gradually increase the amount of information provided by the arguments. First we replaced the arguments in the Senseval2 and the Reuters tuples by the placeholder 'x' to use only information given from the frames. A tuple extracted from the sentence *"He began a battle"* is here represented as $\langle begin, subj{:}obj, x, x \rangle$.

In a second experiment we eliminated the generalization to concepts in PAC. This means that the probability $p(a_i|c, f, i)$ in Equation 5 is directly estimated from data and not decomposed according to Equation 6. A mapping of WordNet-unknown words is not required here. The above tuple would look as follows: $\langle begin, subj{:}obj, he, battle \rangle$

In a third experiment, we replaced pronouns that are likely to refer to humans such as *I*, *he*, *us* etc. to the WordNet concept 'person'. Other arguments which where not in WordNet were mapped to $\langle UNKNOWN \rangle$ as in the base system. Our example tuple turns into: $\langle begin, subj{:}obj, person, battle \rangle$.

Table 1 shows that the difference between no arguments at all and the base system amounts to only 2%. That means that the classification is mostly done by the subcategorisation frame. Selectional preferences improved performance just slightly. The data set where pronouns were mapped to 'person' shows the best results.

In the version without WordNet the arguments caused more damage than they helped. This was a problem of data sparseness. A given tuple with an argument *a* could only be assigned to a cluster if the model contained *a* in the same cluster, the same frame and the same slot. Because the corpus was not large enough it happened quite often that a tuple with a rare frame did not fit into any cluster. For comparison: in our 'no wordnet' data set 107 tuples out of 356 did not belong to any cluster. In the base system this happened only 21 times. This means, if we use detailed information about frames we have to generalize the nouns or we need much more data.

Table 1: Manipulating the Arguments

no arguments	53.40
no wordnet generalization	50.23
base system	55.68
pronouns to 'person'	56.88

Experiments on the Model

Number of Clusters In this experiment we trained clustering models with different numbers of clusters (see table 2)[5]. If the number of clusters was rather small, more senses were united in one cluster causing mis-classifications. An inspection of the classifier parameters of a model with 20 and 160 clusters[6] for the verb *to begin* showed that the average number of *begin*-senses in the 20 cluster model was 4.0 senses per cluster, where 13 clusters contained the verb *to begin*. The 160 cluster model had 72 clusters that contained this verb with an average number of *begin*-senses of 2.72. The total ambiguity rate of the verb *to begin* was 8.

Table 2: Variation in the Number of Clusters

c 20	54.72 (significance: 0.07)
c 40	55.90
c 50 (base system)	55.68
c 60	55.28
c 80	55.52 (significance: 0.05)
c 100	55.85
c 120	56.01
c 140	56.04
c 160	56.58 (significance: 0.07)
c 180	56.69 (significance: 0.05)
c 200	55.96

Although the results were not significant a tendency towards an improvement at higher cluster numbers was visible. It seems that the more clusters we defined the more consistent the clusters were and the better the sense classification turned out. If the number of clusters is too high, we would expect a data sparseness problem because the number of tuples per cluster decreases and the probability estimates become unreliable. Maybe this point is reached with 200 clusters.

Number of Iterations It was often observed that the performance of systems which are trained with the EM algorithm improves over a couple of iterations and then starts to decrease again. Our experiments on the number of iterations show that further training iterations did not make a significant difference after the 30th iteration (see table 3[7]). After 30

[5] Significance testing yielded values over 0.05%. Values that got close to the threshold are nominated.

[6] Only clusters with a probability over 0.01 were considered.

[7] Values marked with an asterisk are significant results compared to the base system.

iterations the results bounced up and down randomly. However, even after 100 iterations we did not reach a turning point where results got noticeably worse.

Table 3: Variation of the Number of Iterations

	c 20	c 50	c 100	c 180
i 10	51.05*	52.45*	53.38*	53.63*
i 20	54.25*	54.05*	55.06	55.06
i 30	54.50*	55.25	55.62	55.09
i 40	54.13*	55.82	55.59	56.24
i 50	54.19*	55.79	55.51	55.76
i 60	55.05*	55.68	55.42	56.01
i 70	54.38*	55.95	55.68	56.32
i 80	54.55*	55.65	55.93	56.60
i 90	54.41*	55.70	55.59	56.80*
i 100	54.72	55.68	55.85	56.69

Comparing LSC and PAC

Since the LSC model does not include the frame in its parameters and since the number of arguments must be fixed, we used a different tuple representation for LSC. We created a pseudo argument containing the frame and we chose only subject and object arguments (which are undefined if not contained in the frame): ⟨begin, subj:obj:p:np, it, visit⟩

If we applied LSC to a data set without arguments, the result was similar to the corresponding PAC result (see table 4). If we added arguments as described above, we got 50.65%. In this experiment the model was losing out because it was trained on a rather small data set[8] and had similar data sparseness problems as the PAC version without WordNet. If we used a larger training set[9], performance improved considerably (see the last row of table 4). The result shows that LSC suffers more from data sparseness than PAC which indicates that the argument generalization helps.

Table 4: Comparing LSC and PAC

	LSC	PAC
no arguments	53.07	53.40
arguments, small corpus	50.65	55.68 (base system)
arguments, large corpus	55.03	56.45

Results

The final evaluation was carried out on the test data of the English Lexical Sample task with the best combination of features according to the previous experiments. That was the data set where the pronouns were partially mapped to the WordNet concept 'person'. The model was trained on a large data set with 180 clusters and 90 iterations. Table 5 compares our results to the accuracy scores of other WSD systems on this task for verbs[10]. The performance of our system is close to that of the best system in the Senseval-2 evaluation (Seo, Lee, Rim, & Lee, 2001) but somewhat behind current state of the art (Chen & Palmer, 2009). However, it must be pointed out that we used very few features - only subcategorisation frames and arguments provided from the clustering model, and that our results are likely to improve after adding further features. Seo et al. (2001)[11] used no linguistic information at all, but took into account local contexts, topical contexts and bigram contexts. These features seem to be quite different from ours. Incorporating them in our system would probably improve the performance.

Table 5: Results on the evaluation data set

MFS	46.1
Seo/Lee	57.6
Dang/Palmer	59.6
Chen/Palmer	64.6
PAC	57.06

Error Analysis and Future Work

We had to deal with errors on different levels. Besides of parser errors – in the Senseval-2 training set 4.1% of the target verbs were not returned – we had the problem that the information in the tuples was often incomplete. Our Senseval-2 data set contained in 2669 out of 3565 tuples one or more placeholders corresponding to arguments missing in WordNet or to unrecognised objects. If we mapped pronouns that referred to humans to the concept 'person', still 2169 tuples contained a placeholder, but results got better. This indicates that future work should concentrate on data preprocessing with anaphora resolution and named entity tagging.

To avoid the bottleneck of manually annotated training data, we would like to turn our supervised system into an unsupervised system by taking the ID of the most probable cluster as the 'verb sense'. To get an intuition of how well our system covers the senses with the clusters we chose the most frequent clusters for the verb *to begin* in a 160-cluster model and looked up the most probable senses included in these clusters. In the following, clusters and senses are listed in descending order according to the frequency or probability respectively. The verb *to begin* has eight senses in the Senseval-2 data. The MFS begin%2:30:00:: was covered in several clusters (c110, c14, c21, c26, c128), which all selected for the frame $subj{:}s$[12] It was interesting to see, that the clusters listed above chose different arguments. c110 selected for a location as a subject, where as c14 selected for a process, c21 for a physical object – which seems to be a very general cluster – c26 for a person and c128 for an abstraction. This means that this model fractions the sense into finer

[8] The small data set contains only tuples with words existent in WordNet (2.4 million Tuple).

[9] In the large data set all tuples provided from the Reuters corpus were taken. Words not included in WordNet were replaced by a placeholder (4.9 million tuple).

[10] Listings of the English Lexical Sample results of verbs can be found in Dang and Palmer (2002)

[11] http://www.informatics.susx.ac.uk/research/groups/nlp/mccarthy/SEVALsystems.html#kunlp, last visited June 2009

[12] 's' is a sentence slot.

grained sense distinctions than WordNet does. The sense begin%2:42:04:: was included in c119 and c75 both holding the intransitive frame and again selecting for different argument concepts: 'process' and 'person'. The sense begin%2:30:01:: is modeled about as well as the described ones.

It was more difficult to model the sense begin%2:42:00:: which occurs only 24 times out of 508 *to begin*-instances. Besides its sparseness it is very similar to sense begin%2:42:04::. The WordNet description for the former is: 'have a beginning, of a temporal event' and for the latter: 'have a beginning, in a temporal, spatial, or evaluative sense'.

Sense begin%2:42:03:: shows that our system has problems if a sense occurs with different subcategorisation frames. This sense was only tagged correctly if it occurred with the frame *subj*:*p*:*np*. It must be pointed out though that we had only 17 instances of this sense in the Senseval-2 corpus. The remaining three senses were never chosen by the system because they occurred very rarely (seven times or less).

Since selectional preferences did not improve results as much as we expected, we had a closer look at the data. Table 6 gives some examples of Senseval-2 tuples, where the first column specifies the sense, the second the subject, and the last one the object of the highly ambiguous verb *to carry*. It shows that the nouns selected by the verb, group well on a higher abstraction level. These examples indicate that selectional preferences seem to be a reasonable feature even for highly ambiguous verbs like *to carry* which encourages to improve argument extraction.

Table 6: Selectional Preferences for *to carry*

carry	subject	object
42:01	Mr. Baker (person)	weapon (artifact)
42:01	he (person)	glass (artifact)
42:02	dept (abstract)	guarantee (abstract)
42:02	bill (abstract)	ban (abstract)
42:12	woman (person)	significance (abstract)
42:12	man (person)	stigma (abstract)
42:03	plane (artifact)	bomb (instrumentality)
42:03	she (= a ship) (artifact)	rigging (instrumentality)

Summary

We proposed a verb sense disambiguation method which labels English verbs with WordNet sense keys. The system consists of (i) a clustering model which is trained on unlabelled verb-argument tuples extracted from the Reuters corpus with a parser, and (ii) a classifier which is trained on the Senseval-2 data and assigns the most likely sense to a verb. The processing consists of three steps, (i) the extraction of the target verb and its arguments with a parser, (ii) the computation of cluster probabilities for the tuple with the clustering model, and (iii) the calculation of the most probable sense based on the cluster(s) assigned in the previous step.

We used two different clustering models (LSC and PAC) and found that PAC outperformed LSC and is not quite as sensitive to data sparseness. Experiments with the number of clusters indicate that a large number of clusters tends to be better. The number of senses per cluster was found to decline as the number of clusters increases.

Our experiments on different data sets showed that information about argument heads improved results by about 2%. However, many arguments were not properly extracted or could not be mapped onto WordNet senses. The improvement resulting from the replacement of personal pronouns with the word 'person' suggests that better argument extraction methods could further increase the performance.

References

Agirre, E., & Edmonds, P. (Eds.). (2006). *Word sense disambiguation: Algorithms and applications*. Springer-Verlag. (URL: http://www.wsdbook.org/)

Chen, J., & Palmer, M. (2009). Improving english verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Language Resources and Evaluation*, *34 (43/2)*, 181–208.

Dang, H., & Palmer, M. (2002). Combining contextual features for word sense disambiguation. In *Proceedings of the SIGLEX/SENSEVAL workshop on WSD: Recent success and future direction.* Philadelphia, USA.

Ide, N., & Wilks, Y. (2006). The simulation of verbal learning behavior. In E. Agirre & Ph.Edmonds (Eds.), *Word sense disambiguation: Algorithms and applications, chapter 3.* Springer.

Kilgarriff, A. (2000). Framework and results for english SENSEVAL. *Computers and Humanities*, *34 (1–2)*, 15–48.

Lapata, M., & Brew, C. (2004). Verb class disambiguation using informative priors. *Computer Linguistics*, *30*(2), 45–73.

Rooth, M., Riezler, S., Prescher, D., Carrol, G., & Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the ACL.* Maryland, MD.

Schmid, H. (2006). Trace prediction and recovery with unlexicalized PCFGs and slash features. In *Proceedings of COLING-ACL'06* (pp. 177–184). Sydney, Australia.

Schulte im Walde, S., Hying, C., Scheible, C., & Schmid, H. (2008). Combining EM training and the MDL principle for an automatic verb classification incorprating selectional preferences. In *Proceedings of the 46th Annual Meeting of the ACL.* Columbus, OH.

Seo, H., Lee, S., Rim, H., & Lee, H. (2001). Kunlp system using classification information model at senseval-2. In *proceedings of the second international workshop on evaluating word sense disambiguation systems (SENSEVAL-2).* Toulouse, F.

Stevenson, S., & Joanis, E. (2003). Semi-supervised verb class discovery using noisy features. In *Proceedings of CoNLL* (pp. 71–78).

Logical Metonymy: Discovering Classes of Meanings

Ekaterina Shutova and Simone Teufel
(Ekaterina.Shutova@cl.cam.ac.uk, Simone.Teufel@cl.cam.ac.uk)
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

Abstract

We address the problem of interpretation of *logical metonymy* using a statistical method. Previous approaches to logical metonymy produce interpretations in the form of verb senses, whereas our definition of the interpretation is a cluster of verb senses. Such a class-based computational model of logical metonymy is novel and more informative than the previous ones. It also complies with the linguistic theories, which we empirically validate. We propose feature sets not previously used for verb clustering. In addition to this we conduct an experiment to prove that our representation is intuitive to human subjects. The system clusters the senses with the F-measure of 0.64 as compared to the gold standard, given that the human agreement on the task is 0.76. **Keywords:** Logical metonymy; distributional semantics; sense clustering.

Introduction

Metonymy involves the use of a word or a phrase to stand for a related concept which is not explicitly mentioned. Here are some examples of metonymic phrases:

(1) The *pen* is mightier than the *sword*.

(2) He played *Bach*.

(3) He enjoyed *the book*. (Pustejovsky, 1991)

(4) After *three martinis* John was feeling well. (Godard & Jayez, 1993)

The metonymic adage in (1) is a classical example. Here the *pen* stands for the mass media and the *sword* for military power. In the following example *Bach* is used to refer to the composer's music. The sentences (3) and (4) represent a variation of this phenomenon called *logical metonymy*. Here both *the book* and *three martinis* have eventive interpretations, i.e. the noun phrases stand for the events of *reading the book* and *drinking three martinis* respectively.

Logical metonymy occurs in natural language texts relatively frequently. For example, according to the corpus study of Verspoor (1997) more than a third of the occurrences of the verb *finish* with a noun phrase in the British National Corpus (Burnard, 2007) are metonymic. Therefore, its automatic interpretation would be useful for many NLP applications that require semantic processing.

There have been a number of theoretical accounts of logical metonymy (Briscoe, Copestake, & Boguraev, 1990; Pustejovsky, 1991, 1995; Godard & Jayez, 1993; Pustejovsky & Bouillon, 1995; Lascarides & Copestake, 1995), corpus-based analyses (Verspoor, 1997), as well as data-driven attempts at its resolution (Lapata & Lascarides, 2003), (Shutova, 2009). The approach of Lapata and Lascarides (2003) generates a list of interpretations (ambiguous with respect to word sense) with their likelihood derived from a corpus. The likelihood of a particular interpretation is calculated using the following formula:

$$P(e,v,o) = \frac{f(v,e) \cdot f(o,e)}{N \cdot f(e)},$$

where e stands for the eventive interpretation of the metonymic phrase, v for the metonymic verb and o for its noun complement. $f(e)$, $f(v,e)$ and $f(o,e)$ are the respective corpus frequencies. $N = \sum_i f(e_i)$ is the total number of verbs in the corpus. The list of interpretations they report for the phrase *finish video* is shown in Table 1.

The approach of Shutova (2009) originates from that of Lapata and Lascarides (2003). It is different from the latter in that they take the interpretation of logical metonymy to be a particular word sense. Following Lapata and Lascarides (2003), their method derives metonymic interpretations using a non-disambiguated corpus, but subsequently maps them to WordNet (Fellbaum, 1998) senses.

They adopt the assumption that the sense frequency distribution is close to Zipfian. Based on this, WordNet sense numbering and the log-probabilities of the non-disambiguated verbs yielded by the model of Lapata and Lascarides (2003) they rank the synsets with respect to their likelihood as metonyic interpretations. They also use the information from WordNet glosses to refine the ranking. The top of the list of synsets Shutova (2009) produce for the metonymic phrase *finish video* and their log-likelihood are given in Table 2.

Taking such lists of sense-based interpretations as input, we extend this to clustering the senses based on their semantic similarity. It has been pointed out in the linguistics literature that the interpretations of metonymic phrases tend to form coherent semantic classes (Pustejovsky, 1991, 1995; Godard & Jayez, 1993). We aim to discover these semantic classes automatically. The challenge of our task is that we cluster particular senses as opposed to ambiguous verbs and, therefore, need to model the distributional information representing a single sense given a non-disambiguated corpus. We adopt the

Interpretations	Log-prob	Interpretations	Log-prob
film	-19.65	make	-21.95
edit	-20.37	programme	-22.08
shoot	-20.40	pack	-22.12
view	-21.19	use	-22.23
play	-21.29	watch	-22.36
stack	-21.75	produce	-22.37

Table 1: The Top of the List of Interpretations of Lapata and Lascarides (2003) for *finish video*

Rank	Synset and its Gloss	Log-L
1	(**watch-v-1**) - look attentively; “watch a basketball game”	-4.56
2	(**view-v-2 consider-v-8 look-at-v-2**) - look at carefully; study mentally; ”view a problem”	-4.66
3	(**watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6**) - see or watch; ”view a show on television”; ”see a movie”	-4.68
4	(**film-v-1 shoot-v-4 take-v-16**) - make a film or photograph of something; ”take a scene”; ”shoot a movie”	-4.91
5	(**edit-v-1 redact-v-2**) - prepare for publication or presentation by correcting, revising, or adapting; ”Edit a book on semantics”	-5.11
6	(**film-v-2**) - record in film; ”The coronation was filmed”	-5.74
7	(**screen-v-3 screen-out-v-1 sieve-v-1 sort-v-1**) - examine in order to test suitability; ”screen these samples”	-5.91
8	(**edit-v-3 cut-v-10 edit-out-v-1**) - cut and assemble the components of; ”edit film”; ”cut recording tape”	-6.20
	...	
207	(**give-v-18 dedicate-v-1 consecrate-v-2 commit-v-2 devote-v-1**)- give entirely to a specific person, activity, or cause;	-12.77

Table 2: Interpretations as Synsets (for *finish video*)

WordNet representation of a sense and cluster verb synsets.

A class-based representation of the interpretation of logical metonymy is novel and arguably more informative than the previous ones. It captures the conceptual structure behind logical metonymy as well as it allows to filter out some irrelevant senses (e.g. “(target-v-1 aim-v-5 place-v-7 direct-v-2 point-v-11) - intend (something) to move towards a certain goal” for *finish directing a video*).

Classes of Interpretations

Pustejovsky (1991) explains the interpretation of logical metonymy by means of lexical defaults associated with the noun complement in the metonymic phrase. He models these lexical defaults in the form of the *qualia structure* of the noun. The qualia structure of a noun specifies among others the following aspects of its meaning: (1) Telic Role (purpose and function of the object); (2) Agentive Role (how the object came into being). For example, the noun *book* would have *read* specified as its telic role and *write* as its agentive role in its qualia structure. Nevertheless, multiple telic and agentive roles can exist and be valid interpretations, as suggested by the data of Lapata and Lascarides and the data of Shutova (2009) (see Tables 1 and 2). Therefore, we propose that these lexical defaults should be represented in the form of classes of interpretations (e.g. {*read, browse, look through*} vs. {*write, compose, pen*}) rather than single word interpretations (e.g. *read* and *write*) as suggested by Pustejovsky (1991).

Godard and Jayez (1993) argue that the reconstructed event is in most cases a kind of a modification to the object referred to by the NP, more specifically, that the object usually *comes into being*, *is consumed*, or *undergoes a change of state*. This conveys an intuition that a sensible metonymic interpretation should fall under one of those three classes.

Comparing the interpretations obtained for the phrase *finish video* (Tables 1 and 2), one can clearly distinguish between the meanings pertaining to the creation of the video, e.g. *film, shoot, take*, and those denoting using the video, e.g. *watch, view, see*. However, the classes based on Pustejovsky’s telic and agentive roles do not explain the interpretation of logical metonymy for all cases. Neither does the class division proposed by Godard and Jayez (1993). For example, the most intuitive interpretation for the metonymic phrase *attempt peak* is *reach*, which does not fall under any of these classes. It is hard to exhaustively characterize all possible classes of interpretations. Therefore, we treat this as an unsupervised clustering problem rather than a classification task and choose a theory-neutral, data-driven approach to it. The objective of our experiment is to model the class division structure of metonymic interpretations and experimentally ascertain whether the obtained data conforms to it.

Clustering Verb Senses Automatically

In order to discover conceptual classes of interpretations and in order to be able to generalise over the obtained data, we need to cluster the synsets from our list to identify groups of synsets with related meanings. There has been a number of approaches to automatic verb clustering. The vast majority of them relies on syntactic information about verb subcategorisation (Korhonen, Krymolowski, & Marx, 2003; Joanis, Stevenson, & James, 2008) and the thematic roles assigned by the verb to its arguments (Merlo & Stevenson, 2001). Their work originates from the idea of Levin (1993) that the verbs exposing similar diathesis alternations form coherent semantic classes. Some approaches utilised selectional preferences (semantic classes of the nouns the verb selects for) along with subcategorisation frames to construct their feature sets (Lin, 1998; Schulte im Walde, 2006; Korhonen, Krymolowski, & Collier, 2008).

The Data

We used the method developed by Shutova (2009) to create the initial list of sense-based interpretations. The parameters of the model were estimated from the British National Corpus (BNC) (Burnard, 2007) that was parsed using the RASP parser of Briscoe, Carroll, and Watson (2006). We used the grammatical relations (GRs) output of RASP for BNC created by Andersen, Nioche, Briscoe, and Carroll (2008).

Feature Extraction

The goal is to cluster synsets with similar distributional semantics together. Our feature sets comprise the nouns co-occurring with the verbs in the synset in subject and object relations. The object relations were represented by the nouns co-occurring with the verb in the same syntactic frame as the noun in the metonymic phrase (e.g. `indirect object` with the preposition *in* for *live in the city*, `direct object` for *visit the city*). These nouns together with the co-occurrence frequencies were used as features for clustering. The subject and object relations were marked respectively. We use the following notation:

$$\begin{aligned}
&\mathbb{V}_1 = \{c_{11}, c_{12}, ..., c_{1N}\}\\
&\mathbb{V}_2 = \{c_{21}, c_{22}, ..., c_{2N}\}\\
&\dots\\
&\mathbb{V}_K = \{c_{K1}, c_{K2}, ..., c_{KN}\}
\end{aligned}$$

where K is the number of the verbs in the synset, $\mathbb{V}_1, ..., \mathbb{V}_K$ are the feature sets of each verb, N is the total number of features (ranges from 18517 to 20661 in our experiments) and c_{ij} are the corpus counts. The following feature sets were taken to represent the whole synset.

Feature set 1 - the union of the features of all the verbs of the synset: $\mathbb{F}_1 = \mathbb{V}_1 \cup \mathbb{V}_2 \cup \ldots \cup \mathbb{V}_K$. The counts are computed as follows:

$$\mathbb{F}_1 = \{\sum_{i=1}^{K} c_{i1}, \sum_{i=1}^{K} c_{i2}, ..., \sum_{i=1}^{K} c_{iN}\}$$

However, this featureset contains features describing irrelevant senses of the verbs. Such irrelevant features can be filtered out by taking an intersection of the nouns of all the verbs in the synset. This yields the following feature set:

Feature set 2 - the intersection of the feature sets of the verbs in the synset: $\mathbb{F}_2 = \mathbb{V}_1 \cap \mathbb{V}_2 \cap \ldots \cap \mathbb{V}_K$. The counts are computed as follows:

$$\mathbb{F}_2 = \{f_1, f_2, ..., f_N\}$$

$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \prod_{i=1}^{K} c_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

This would theoretically be a comprehensive representation. However, in practice the system is likely to run into the problem of data sparseness and some synsets end up with very limited feature vectors, or no feature vectors at all. The next feature set is trying to accommodate this problem.

Feature set 3 - union of features as in feature set 1, reweighted in favour of overlapping features.
$\mathbb{F}_3 = \mathbb{V}_1 \cup \mathbb{V}_2 \cup \ldots \cup \mathbb{V}_K \cup \beta * (\mathbb{V}_1 \cap \mathbb{V}_2 \cap \ldots \cap \mathbb{V}_K) =$
$= \mathbb{F}_1 \cup \beta * \mathbb{F}_2$
where β is the weighting coefficient.

We empirically set β to 5 in our experiments. The feature sets 4 and 5 are also motivated by the problem of sparse data. But the intersection of features is calculated pairwise, instead of an overall intersection.

Feature set 4 - pairwise intersections of the feature sets of the verbs in the synset.
$\mathbb{F}_4 = (\mathbb{V}_1 \cap \mathbb{V}_2) \cup \ldots \cup (\mathbb{V}_1 \cap \mathbb{V}_K) \cup (\mathbb{V}_2 \cap \mathbb{V}_3) \cup \ldots$
$\cup (\mathbb{V}_2 \cap \mathbb{V}_K) \cup \ldots \cup (\mathbb{V}_{K-2} \cap \mathbb{V}_{K-1}) \cup (\mathbb{V}_{K-1} \cap \mathbb{V}_K)$
The counts are computed as follows:

$$\mathbb{F}_4 = \{f_1, f_2, ..., f_N\}$$

$$f_j = \begin{cases} \sum_{i=1}^{K} c_{ij} & \text{if } \exists x, y | c_{xj} \cdot c_{yj} \neq 0, \\ & x, y \in [1..K], x \neq y; \\ 0 & \text{otherwise.} \end{cases}$$

Feature set 5 - the union of features as in feature set 1, reweighted in favour of overlapping features (pairwise overlap): $\mathbb{F}_5 = \mathbb{F}_1 \cup \beta * \mathbb{F}_4$, where β is the weighting coefficient.

Development Set	Test Set
enjoy book	enjoy story
finish video	finish project
start experiment	try vegetable
finish novel	begin theory
enjoy concert	start letter

Table 3: Metonymic Phrases in Development and Test Sets

Clustering Experiments

We ran the experiment with the following clustering configurations:

Clustering Algorithms: K-means, Repeated bisections, Agglomerative (single link, complete link, group average).

Similarity Measures: Cosine, Correlation coefficient.

Criterion Function: The goal is to maximize intra-cluster similarity and to minimize inter-cluster similarity. We use the function ε_2 (Zhao & Karypis, 2001) defined as follows:

$$\varepsilon_2 = \min \sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} sim(v,u)}{\sqrt{\sum_{v,u \in S_i} sim(v,u)}}$$

where S is the set of objects to cluster, S_i is the set of objects in cluster i, n_i is the number of objects in cluster i, k is the number of clusters and *sim* stands for the chosen similarity measure. As such, the numerator represents inter-cluster similarity and the denominator intra-cluster similarity.

Feature Matrix Scaling. We used the following scaling schemes: (1) IDF paradigm, whereby the counts of each column are scaled by the $\log_2$ of the total number of rows divided by the number of rows the feature appears in (this scaling scheme only uses the frequency information inside the matrix). The effect is to de-emphasize columns that appear in many rows and are, therefore, not very discriminative features. (2) We preprocessed the matrix by dividing initial counts for each noun by the total number of occurrences of this noun in the whole BNC. The objective was again to decrease the influence of generally frequent nouns that are also likely to be ambiguous features.

The Number of Clusters: We set the number of clusters (k) for each metonymic phrase manually according to the number observed in the gold standard.

We used the Cluto Toolkit (Karypis, 2002). Cluto has been applied in NLP mainly for document classification tasks, but also for a number of experiments on lexical semantics.

Evaluation

Our dataset consists of 10 metonymic phrases taken from the dataset of Lapata and Lascarides (2003). We split them into a development set (5 phrases) and a test set (5 phrases), as Table 3 shows.

The Gold Standard

The gold standard was created for the top 30 synsets from the lists of interpretations. This threshold allows to filter out a large number of incorrect interpretations. It was set experimentally: the top 30 synsets contain 70% of correct interpre-

Cluster 1: (film-v-1 shoot-v-4 take-v-16) (film-v-2) (produce-v-2 make-v-6 create-v-6) (direct-v-3) (work-at-v-1 work-on-v-1) (work-v-5 work-on-v-2 process-v-6) (make-v-3 create-v-1) (produce-v-1 bring-forth-v-3)

Cluster 2: (watch-v-3 view-v-3 see-v-7 catch-v-15 take-in-v-6) (watch-v-1) (view-v-2 consider-v-8 look-at-v-2) (analyze-v-1 analyse-v-1 study-v-1 examine-v-1 canvass-v-3 canvas-v-4) (use-v-1 utilize-v-1 utilise-v-1 apply-v-1 employ-v-1) (play-v-18 run-v-10)

Cluster 3: (edit-v-1 redact-v-2) (edit-v-3 cut-v-10 edit-out-v-1) (screen-v-3 screen-out-v-1 sieve-v-1 sort-v-1) (work-through-v-1 run-through-v-1 go-through-v-2)

Figure 1: Gold Standard for the Phrase *finish video*

tations (average recall over metonymic phrases from the development set). Our gold standard for each metonymic phrase consists of a number of clusters containing correct interpretations in the form of synsets and a cluster containing incorrect interpretations. The cluster containing incorrect interpretations is significantly larger than the others for the majority of metonymic phrases.

The gold standard was manually created by the authors. It is presented in Figure 1, exemplified for the metonymic phrase *finish video*. The glosses and the cluster with incorrect interpretations are omitted for the sake of brevity.

Evaluation Measures

We will call the gold standard partitions *classes* and the clustering solution suggested by the model a set of *clusters*. The following measures were used to evaluate clustering:

Purity (Zhao & Karypis, 2001) is calculated as follows

$$\text{Purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k \cap c_j|$$

where $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, ..., c_j\}$ is the set of classes, N is the number of objects to cluster. Purity evaluates only the homogeneity of the clusters. High purity is easy to achieve when the number of clusters is large. As such, it does not provide a measure for the trade off between the quality of clustering and the number of classes.

F-Measure was introduced by van Rijsbergen (1979) and adapted to the clustering task by Fung, Wang, and Ester (2003). It matches each class with the cluster that has the highest precision and recall. Using the same notation as above

$$F(\mathbb{C}, \Omega) = \sum_{j}\frac{|c_j|}{N}\max_{k}\{F(c_j, \omega_k)\}$$

$$F(c_j, \omega_k) = \frac{2 \cdot P(c_j, \omega_k) \cdot R(c_j, \omega_k)}{P(c_j, \omega_k) + R(c_j, \omega_k)}$$

$$R(c_j, \omega_k) = \frac{|\omega_k \cap c_j|}{|c_j|} \qquad P(c_j, \omega_k) = \frac{|\omega_k \cap c_j|}{|\omega_k|}$$

Recall represents a portion of objects of class c_j assigned to cluster ω_k and precision the portion of objects in cluster ω_k belonging to the class c_j.

Rand Index (Rand, 1971). An alternative way of looking at clustering is to consider it as a series of decisions for each pair of objects, whether these two objects belong to the same cluster or not. For N objects there will be $N(N-1)/2$ pairs. We then calculate the number of true positives (TP) (similar objects in the same cluster), true negatives (TN) (dissimilar objects in different clusters), false positives (FP) and false negatives (FN). Rand Index corresponds to accuracy: it measures the percentage of decisions that are correct considered pairwise.

$$RI = \frac{TP + TN}{TP + FP + TN + FN}$$

Variation of Information (Meilă, 2007) is an entropy-based measure defined as follows:

$$VI(\Omega, \mathbb{C}) = H(\Omega|\mathbb{C}) + H(\mathbb{C}|\Omega)$$

where $H(\mathbb{C}|\Omega)$ is the conditional entropy of the class distribution given the proposed clustering, $H(\Omega|\mathbb{C})$ is the opposite.

$$H(\Omega|\mathbb{C}) = -\sum_{j}\sum_{k}\frac{|\omega_k \cap c_j|}{N}\log\frac{|\omega_k \cap c_j|}{|\omega_k|}$$

$$H(\mathbb{C}|\Omega) = -\sum_{k}\sum_{j}\frac{|\omega_k \cap c_j|}{N}\log\frac{|\omega_k \cap c_j|}{|c_j|}$$

where $\Omega = \{\omega_1, \omega_2, ..., \omega_k\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, ..., c_j\}$ is the set of classes, N is the number of objects to cluster. We report the values of VI normalized by $\log N$, which brings them into the range $[0, 1]$.

It is easy to see that VI is symmetrical. This means that it accounts for both homogeneity (only similar objects within the cluster) and completeness (all similar objects are covered by the cluster). In the perfectly homogeneous case the value of $H(\mathbb{C}|\Omega)$ is 0, in the perfectly complete case the value of $H(\Omega|\mathbb{C})$ is 0. The values are maximal (and equal to $H(\mathbb{C})$ and $H(\Omega)$ respectively) when the clustering gives no new information and the class distribution within each cluster is the same as the overall class distribution.

Inter-Annotator Agreement

The subjectivity in annotator judgements is an inevitable obstacle for any semantic annotation. It is conventionally evaluated in terms of *inter-annotator agreement*, a measure of how similar the annotations produced by different annotators are. We define the ceiling for our task by the inter-annotator agreement.

Obtaining the Annotations We conduct an experiment with humans in order to show that they find our definition of interpretation of logical metonymy intuitive and that they cluster word sense-based interpretations similarly, i.e., they agree on the task.

We had 8 volunteer subjects altogether. All of them were native speakers of English and non-linguists. We divided them into 2 groups: 4 and 4. Subjects in each group annotated three metonymic phrases (Group 1: finish video, start

Algorithm	F. S.	Purity	RI	F-measure	VI
K-means	F1	0.6	0.52	0.54	0.45
No scaling	F2	0.61	0.58	0.61	0.45
Cosine	F3	0.57	0.5	0.54	0.47
	F4	**0.65**	**0.57**	**0.69**	**0.35**
	F5	0.6	0.54	0.57	0.44
RB	F1	0.61	0.51	0.58	0.43
No scaling	F2	0.62	0.57	0.63	0.44
Cosine	F3	0.63	0.52	0.61	0.40
	F4	**0.64**	**0.56**	**0.70**	**0.34**
	F5	0.61	0.52	0.59	0.42
Agglomerative	F1	0.61	0.47	0.76	0.33
No scaling	F2	0.61	0.57	0.70	0.44
Cosine	F3	0.61	0.47	0.64	0.35
Group average	F4	0.63	0.5	0.69	0.31
	F5	0.6	0.46	0.64	0.35

Table 4: Average Clustering Results (development set)

Algorithm	F. S.	Purity	RI	F-measure	VI
K-means	F1	0.7	0.54	0.58	0.35
No scaling	F2	0.67	0.48	0.57	0.36
Cosine	F3	0.7	0.54	0.58	0.35
	F4	**0.73**	**0.70**	**0.88**	**0.19**
	F5	0.7	0.54	0.58	0.35

Table 5: Best Clustering Results (for *enjoy concert*, development set)

Algorithm	F. S.	Purity	RI	F-measure	VI
Baseline		0.48	0.40	0.31	0.51
K-means	**F4**	**0.65**	**0.52**	**0.64**	**0.33**
RB	F4	0.63	0.48	0.60	0.37
Agreement		0.75	0.67	0.76	0.37

Table 6: Clustering Results on the Test Set

experiment, enjoy concert; Group 2: finish project, begin theory, start letter). They received written guidelines describing the task (2 pages). For each metonymic phrase the subjects were presented with top 30 synsets produced by the system and asked to (1) remove the synsets that do not have the right meaning in the context of the metonymic phrase and (2) cluster the remaining ones according to their semantic similarity (they were free to choose a number of clusters most intuitive to them).

Evaluating Agreement We calculate the agreement by comparing the annotations pairwise (each annotator with each other annotator and the gold standard) and access it in terms of the same clustering evaluation measures as the ones used to access the system performance.

In order to compare the groupings elicited from humans we added the cluster with the interpretations they excluded as incorrect to their clustering solutions. This was necessary, as the metrics described in section 4 require that all annotators' clusterings contain the same objects (all 30 interpretations).

After having evaluated the agreement pairwise for each metonymic phrase we calculated the average across the metonymic phrases and the pairs of annotators. We obtained the agreement of 0.75 (Purity), 0.67 (Rand index), 0.76 (F-measure), 0.37 (VI). It should be noted, however, that the granularity of clusters produced varies from annotator to annotator and the chosen measures (except for VI) penalize this.

Parameter Fitting

To select the best parameter setting we ran the experiments on the development set varying the parameters described in section 4 for feature sets 1 to 5. The system clustering solutions were evaluated for each metonymic phrase separately; the average values for the best clustering configurations for each algorithm and each feature set on the development set are given in Table 4. The best result was obtained for the phrase *enjoy concert* as shown in Table 5.

The performance of the system is similar across the algorithms. However, the agglomerative algorithm tends to produce single object clusters and one large cluster containing the rest, which is strongly dispreferred. For this reason, we test the system only using K-means and repeated bisections. The results obtained suggest that feature set 4 is the most informative, although for agglomerative clustering feature set 1 yields a surprisingly good result. We will use feature set 4 in our evaluation, as it proves to be useful for all three clustering algorithms.

Baseline

We compare the system clustering to that of a baseline built using a simple heuristic. The baseline assigns the synsets that contain the same verb string to the same cluster. The baseline clustering was evaluated using the measures described in section 4 and the results are presented below.

Results and Cluster Analysis

We present the results for the best system configuration on the test data in Table 6. The system outperforms the naive baseline, but does not reach the ceiling set by the inter-annotator agreement. K-means algorithm yields the best result of 0.65 (Purity), 0.52 (Rand index), 0.64 (F-measure) and 0.33 (VI).

Having a relatively small data set allows us to perform a qualitative analysis of the clusters. A particularity of our clustering task is that our goal is to eliminate incorrect interpretations as well as assign the correct ones to their classes based on semantic similarity. The cluster containing incorrect interpretations is often significantly larger than the other clusters. The overall trend is that the system selects correct interpretations and assigns them to smaller clusters, leaving the incorrect ones in one large cluster, as desired.

A common error of the system is that the synsets that contain different senses of the same verb often get clustered together. This is due to the fact that the features are extracted from a non-disambiguated corpus, which results in the following problems: (1) the verbs are ambiguous, therefore, the features, as extracted from the corpus, represent all the senses of the verb in one feature set. The task of dividing this feature set into subsets describing particular senses of the verb is very hard; (2) the features themselves (the nouns) are ambiguous (different senses of a noun can co-occur with different senses of a verb), which makes it very hard to distribute the counts realistically over verb senses.

It should be noted, however, that it is not always the case that synsets with overlapping verbs get clustered together (in 38% of all cases the same verb string is assigned to different clusters), which demonstrates the contribution of the presented feature sets. More importantly, synsets containing different verbs are often assigned to the same cluster, when the sense is related (mainly for feature sets 2 and 4), which is what we aimed for.

Future Work

Another possible solution to the problem of data sparsity would be to apply class-based smoothing to our feature vectors. In other words one can back-off to the broad classes of nouns and represent the features of a verb as its selectional preferences (the constraints that the verb places onto its arguments). To build a feature vector of a synset, we then need to find common preferences of its verbs. The class-based overlap can be applied both simultaneously and pairwise. Representing features in the form of semantic classes can also be viewed as a linguistically motivated way of dimensionality reduction of feature matrices. Essentially some of the dimensions (features belonging to the same class) will be merged and their counts will be added. Although this is potentially a promising experiment, we are aware of the fact that there is a risk of introducing additional errors into the system due to imperfect selectional preference acquisition.

Along with experimenting with the above feature sets we plan to apply a clustering algorithm that determines the number of clusters automatically. This can be achieved by using Bayesian non-parametric models (e.g. Dirichlet Process Mixture Models (Vlachos, Korhonen, & Ghahramani, 2009)).

In addition to this we plan to perform a more comprehensive evaluation. We will test the system on a larger data set using the described clustering evaluation techniques, as well as perform an extrinsic evaluation, i.e. evaluate our system in terms of how a different NLP application could benefit from the resolution of logical metonymy.

Conclusion

We presented a method for the automatic discovery of conceptual classes of metonymic interpretations. Such a class-based computational model of the interpretation is novel and significantly more informative than the previous ones. We showed that it is intuitive to human subjects and that it complies with the results of theoretical research on logical metonymy in linguistics.

In addition to this, we addressed the issue of modelling distributional semantics of single senses represented in the form of a WordNet synsets using a non-disambiguated corpus. The obtained results demonstrate the efficiency of our approach to synset clustering in the context of logical metonymy.

Acknowledgements

We would like to thank Anna Korhonen and our anonymous reviewers for their most helpful feedback on this work.

References

Andersen, O. E., Nioche, J., Briscoe, E., & Carroll, J. (2008). The BNC parsed with RASP4UIMA. In *Proceedings of LREC'08.* Marrakech, Morocco.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the rasp system. In *Proceedings of the COLING/ACL on interactive presentation sessions.*

Briscoe, E., Copestake, A., & Boguraev, B. (1990). Enjoy the paper: lexical semantics via lexicology. In *Proceedings of COLING-90.* Helsinki.

Burnard, L. (2007). *Reference guide for the british national corpus (xml edition).* Available from `http://www.natcorp.ox.ac.uk/XMLedition/URG/`

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database (isbn: 0-262-06197-x)* (First ed.). MIT Press.

Fung, B. C. M., Wang, K., & Ester, M. (2003). Large hierarchical document clustering using frequent itemsets. In *Proceedings of SIAM 2003.*

Godard, D., & Jayez, J. (1993). Towards a proper treatment of coercion phenomena. In *Sixth conference of the european chapter of the ACL* (p. 168-177). Utrecht.

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

Karypis, G. (2002). *Cluto: A clustering toolkit* (Tech. Rep.). University of Minnesota.

Korhonen, A., Krymolowski, Y., & Collier, N. (2008). The choice of features for classification of verbs in biomedical texts. In *Proceedings of COLING.* Manchester, UK.

Korhonen, A., Krymolowski, Y., & Marx, Z. (2003). Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL 2003.* Sapporo,Japan.

Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, *29(2).*

Lascarides, A., & Copestake, A. (1995). The pragmatics of word meaning. In *Journal of linguistics* (pp. 387–414).

Levin, B. (1993). *English verb classes and alternations.* Chicago: University of Chicago Press.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics.*

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5).

Merlo, P., & Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, *27*(3).

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, *17(4).*

Pustejovsky, J. (1995). *The generative lexicon.* Cambridge, MA: MIT Press.

Pustejovsky, J., & Bouillon, P. (1995). Logical polysemy and aspectual coercion. *Journal of Semantics*, *12*, 133–162.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336).

van Rijsbergen, C. J. (1979). *Information retrieval, 2nd edition.* London: Butterworths.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, *32(2).*

Shutova, E. (2009). Sense-based interpretation of logical metonymy using a statistical method. In *Proceedings of the ACL 2009 student workshop.* Singapore.

Verspoor, C. M. (1997). Conventionality-governed logical metonymy. In *Proceedings of the second international workshop on computational semantics.* Tilburg.

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

Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis* (Tech. Rep.). University of Minnesota.

Phonaesthemic and Etymological effects on the Distribution of Senses in Statistical Models of Semantics

Armelle Boussidan (armelle.boussidan@isc.cnrs.fr)
L2C2, Institut des Sciences Cognitives-CNRS,
Université Lyon II, Bron, France

Eyal Sagi (ermon@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Sabine Ploux (sploux@isc.cnrs.fr)
L2C2, Institut des Sciences Cognitives-CNRS,
Université Lyon I, Bron, France

Abstract

This paper uses methods based on corpus statistics and synonymy to explore the role language history and sound/form relationships play in conceptual organization through a case study relating the phonaestheme *gl-* to its prevalent Proto-Indo European root, **ghel*. The results of both methods point to a strong link between the phonaestheme and the historical root, suggesting that the lineage of a language plays an important role in the distribution of linguistic meaning. The implications of these findings are discussed.

Keywords: Corpus statistics, Synonymy, Historical Linguistics, Sound/form relationships.

Introduction

Recent years have seen a surge in the use of statistical models to describe the distribution and inter-relation of concepts at the cognitive level and meanings at the linguistic level.[1] These models have been applied to a wide range of tasks, from word-sense disambiguation (Levin et al., 2006) to the summarization of texts (Marcu, 2003) and the tracing of semantic change (Sagi, Kaufmann, & Clark, 2009). They have also been used to model a variety of cognitive phenomena, such as semantic priming (Burgess, Livesay, & Lund, 1998) and categorization (Louwerse, et al., 2005).

In this paper we will explore the role that language history and sound/form relationships might play in conceptual organization using two methods – one based on corpus statistics (Infomap, Schütze, 1996) and the other based on synonymy (Semantic Atlases, Ploux & Victorri, 1998). Importantly, the use of both corpus-based and lexicon-based statistics allows us to examine these phenomena at two different levels – lexical meaning and language in use. This examination will highlight that even though a language can undergo drastic changes over time, some aspects of the underlying cognitive organization remain stable.

Many models based on corpus statistics (e.g., LSA, Landauer & Dumais, 1997; Infomap, Schütze, 1996; Takayama, et al. 1999; HAL, Lund & Burgess, 1996) are built around the assumption that related words will tend to co-occur within a single context with higher frequency than unrelated words. As a result, this pattern of word co-occurrence can be considered an approximation of the underlying organization of concepts.

The relationship between words and concepts can also be described in terms of closest semantic equivalents, synonyms. (Wordnet, Fellbaum, 1998; Semantic Atlases, Ploux, 1997; Ploux & Victorri, 1998). The Semantic Atlas (SA) is a geometrical model of meaning based on fine grained units of meaning called 'cliques'. Each clique contains a series of terms all synonymous with each other.

While models that rely on measuring word co-occurrence might seem to be very different from those that are based on identifying clusters of synonyms in dictionaries, both approaches are distributional in nature and rely on very similar methods of investigation. Nevertheless, these approaches take somewhat different perspectives and examine different aspects of word distribution. Therefore, they may complete each other so as to reach a more complex and complete picture of how word meanings are anchored in language on the one hand, and how they relate to concepts on the other. Both synonymy and context participate in the architecture of meaning and in relating lexical items to a conceptual network.

We can use different types of data to enhance our understanding of language. For instance, following work by Firth (1930), Otis and Sagi (2008) demonstrate that the distribution of terms in a corpus is also related to the phonetic features of words known as *phonaesthemes*, sub-morphemic units that have a predictable effect on the meaning of a word as a whole. For instance, non-obsolete English words that begin with *gl-* are, more often than not, related to the visual modality (e.g., *gleam*, *glitter*, *glance*) whereas words that begin with *sn-* are usually related to the nose (e.g., *snore*, *sniff*, *snout*). More generally, it appears that some phonetic aspects of word form might be related to meaning and indicative of its conceptual underpinnings.

However, to properly utilize this new information it is important to understand how it relates to conceptual organization. For instance, phonetic similarity may be used as a cue for conceptual similarity. This suggests that phonaesthemes may be a specific case of a more general principle and that in contrast with the Saussurian tradition,

[1] As Jackendoff (1983: 95) notes, it is possible that "semantic structure is conceptual structure". However, for the purpose of this paper we will assume that these two levels of representation are distinct.

language might incorporate an abundance of non-trivial relations between word form or sound and word meaning.

Another factor that governs these similarities is the history of the language – For instance, reconstructions of Proto-Indo European, the ancestor of many of the languages spoken in Europe and western Asia, suggest that it was a root-based language and as such incorporated many meaningful morpho-phonological clusters. Some of these may have survived through the generations and formed the basis for phonaesthemes. In this case, the survival of these specific clusters might indicate that they are linked with important aspects of cognitive organization. As a result, identifying and cataloging these phonaesthemes might provide interesting insights into some of the basic dimensions underlying the organization of concepts. In this paper we examine this question by contrasting the influence of phonetic similarity and the historical roots of words in the case of the *gl-* phonaestheme and its prevalent Proto-Indo European root, **ghel*.

**ghel/gl-*: A case study

Indo European (IE) or Proto-Indo European (PIE) is a reconstructed common original language covering almost all languages spoken from Europe to India and dated around the fifth millennium BC. It gives birth to ten families of languages including the Germanic branch, of which English is a descendant. 19th century comparative linguists carried out PIE's reconstruction by observing similarities across languages and with the help of mutation rules. They determined a semantic common denominator for each root. As a consequence, root definitions are often vague, imprecise and all-encompassing. This calls for caution on the semantic plane: while the senses of PIE roots might seem more vague than those used in modern day English word definitions, this could be an effect of the reconstruction process rather than a real semantic difference.

In English, the vocabulary inherited from PIE appears to form the genuine core of the language even though it represents a small proportion of it compared to loan words. For example, Watkins (2000) reports that the 100 most frequent words in the Brown corpus are PIE based. PIE was an inflected language following the structure Root + Suffix + Ending. Some derivations were made on the basis of inflected words. The root is thus the most stable unit although roots can undergo extension and words can derive directly from these extensions. In PIE consonant alternation conveys semantic content whereas vowel change is apophonic, that is, it expresses morphological functions (Philps, 2008a). Although sound patterns and orthographic patterns follow laws of change which are quite regular, the semantic content attached to them often survives these changes and re-establishes a connection with the new sound forms and orthographic forms. This pattern seems to be central in language change processes.

Watkins (2000) identified **ghel* as a PIE root meaning "to shine" with derivatives referring to colors, bright materials, gold (probably yellow metal) and bile or gall[2]. It produces a series of words denoting colors (e.g., *yellow* from the extended root **-ghel-wo-*), words denoting gold (e.g., *gold* from the zero grade[3] form **ghl-to-*), words denoting bile and gall (*gall* from the o-grade form **ghol-no-*) and most interestingly a bag of Germanic words related to light and vision starting with *gl-* (e.g., *gleam, glass*).

Researchers identified the phonaestheme *gl-* as relating to the "phenomena of light", to "visual phenomena" (Bolinger, 1950, pp. 119 & 131) and to the concepts "light" and "shine" (Marchand, 1960, p. 327). However, while many English words that feature this phonaestheme seem to have a meaning that is obviously related to the visual modality (e.g., *glow*, *glare*, *glisten*), some other words (e.g., *glue*, *glucose*) appear to be unrelated. Therefore, it seems that phonaesthemes are not absolute – not all words that feature them fit the conceptual pattern of the phonaestheme. A phonaestheme is therefore more likely to be a statistical cue to some general conceptual features of meaning.

However some apparently unrelated items may be associated to the central meaning of the *gl-* phonaestheme via the process of antonymy ("fire, to be warm", balanced by "cold" in *glace*, and "light" balanced by "dark" in *gloom*) or other similar processes. Concepts related to the tongue and swallowing appear in words such as *glottis*, or *glutton* which might be explained by a conceptual mapping from mouth to eye in terms of their open-close characteristics as described in Philps (2008b). Similarly there are *gl-* words that do not have a meaning related to light (e.g., "to cut" from the **kel-* root, " sweetness" from **dlk-u-*, "clay" from **glei-*, and "cold" from **gel-*).

Otis and Sagi (2008) demonstrated that it is possible to statistically validate the internal consistency of meaning that is at the core of phonaesthemes – i.e., that the group of words which feature a specific phonaestheme are also closer in meaning than a similarly-sized group of words that do not share a phonaestheme. Furthermore, priming experiments conducted by Bergen (2004) suggest that cognitive processing of linguistic stimuli is affected by phonaesthemes and that these effects cannot be fully explained as the result of either semantic or phonetic similarity.

As a result, it appears that there are two possible factors that might explain the relationship between phonaesthemes and word meaning – the historical root of the words, and cognitive processes that relate phonetic and semantic similarity. Importantly, these hypotheses are not mutually exclusive. One way to compare them is to examine how much of the relatedness between sound and meaning that

[2] **ghel-*, to call, shout and **ghel-*, to cut, are homonymic roots which do not appear in the '*gl-*' set of words and therefore will not be investigated in this paper.

[3] There are three grades in Indo-European grammar: the full grade in -e-, the o-grade, and the zero-grade (without vowel). Here the zero grade form of **ghel-* (full grade) is **ghl-*, and its o-grade is **ghol-*.

identifies a phonaesthеme is attributable to the historical root and how much is attributable to phonetic similarity.

In other words, if the observed effect is due to the historical root **ghel* then it should extend equally to all words that resulted from that root, but not to words that resulted from other roots. Similarly, if the effect of phonaesthemes is primarily due to their phonetic similarity then the effect exhibited by the phonaestheme *gl-* should be restricted to words that begin with *gl-*, regardless of their PIE root, but should not extend to other words that originated from the **ghel* root. We will test this hypothesis using two different approaches. Firstly, we will employ the method developed by Otis and Sagi (2008). Because the cohesiveness of a word cluster is a measure of its inter-relatedness, we can use this measure to examine the relative role of the PIE root **ghel* and the phonaestheme *gl-* by comparing their relative cohesiveness. Specifically, we hypothesize that if the historical root **ghel* is the source of the phonaestheme *gl-* then the cluster of words belonging to the root should be more cohesive than the cluster of words that begin with *gl-*, and vice versa.

Secondly, we will examine clusters generated from the Semantic Atlases synonym database (Ploux & Victorri, 1998) and investigate whether *gl-* and non *gl-* sets have independent semantic status and sound/form within the **ghel* space and conversely for the **ghel* set within the *gl-* space.

Following our hypothesis, if the phonaestheme *gl-* has its roots in the PIE root **ghel*, then we would expect the average distance between words that come PIE root **ghel* and begin with *gl-* to be small compared to the average distance between words in other sets. In addition, we predict that the *gl-* set will be more cohesive within the **ghel* space than the whole, due to its phonetic unity, and that the **ghel* set will be more cohesive within the *gl-* space than the whole due to its historic unity.

Method

Materials

We identified PIE roots based on the work done by Watkins (2000). The lists of words starting with *gl-* were generated on the basis of the dictionary database for the SA and on the basis of the corpus for Infomap. A sample of words used in this study as well as their PIE roots (if known) can be found in Appendix A.

Using Infomap to measure cluster cohesiveness

The corpus

We used a corpus based on Project Gutenberg (http://www.gutenberg.org/). Specifically, we used the bulk of the English language literary works available through the project's website. This resulted in a corpus of 4034 separate documents consisting of over 290 million words. Infomap analyzed this corpus using default settings (a co-occurrence window of 15 words and using the 20,000 most frequent content words for the analysis) and its default stop list.

Computing Word Vectors

For our computational model we used Infomap (http://infomap-nlp.sourceforge.net/; Schütze, 1996), which represents words as vectors in a multi-dimensional space based on the frequency of word co-occurrence. In this space, vectors for words that frequently co-occur are grouped closer together than words that rarely co-occur. As a result, words which relate to the same topic, and can be assumed to have a strong semantic relation, tend to be grouped together. This relationship can then be measured by correlating the vectors representing those two words within the semantic space.[4] Importantly, as mentioned in Buckley, et al. (1996), the first factor identified by Infomap is somewhat problematic as it is monotonically related to the frequency of the term. Because of this we elected to omit it when computing word vector correlations.

For each occurrence of a target word type under investigation, we calculated a context vector by summing the vectors for the content words within the 15 words preceding and the 15 words following that occurrence. The vector for a word is then simply the normalized sum of the vectors representing the contexts in which the word occurs.

Measuring the cohesiveness of a word cluster

We measured the cohesiveness of a word cluster in a similar manner to that used by Otis and Sagi (2008). The cohesiveness of a cluster was defined as the average correlation of the vector pairs comprising the cluster – a higher correlation value represents a more cohesive cluster (*r* below). It is also possible to directly test whether the cohesiveness of a cluster is greater than that of another. For this purpose we used Monte-Carlo sampling to repeatedly choose 50 pairs of words from the hypothesized cluster and 50 pairs of words from a similarly size cluster chosen from the corpus as a whole. We used an independent sample t-test to test the hypothesis that the one of the clusters was more cohesive (had a higher average cosine) than the other. This procedure was repeated 100 times and we compared the overall frequency of statistically significant t-tests with the binomial distribution for α=.05. After applying a Bonferroni correction for performing 50 comparisons, the threshold for statistical significance of the binomial test was for 14 t-tests out of 100 to turn out as significant, with a frequency of 13 being marginally significant. Therefore, if the significance frequency (*#Sig* below) of a candidate cluster was 15 or higher, then one of the clusters was judged as being more cohesive than the other.

Synonym clustering

Clustering was conducted using the Semantic Atlas synonym database, which is composed of several dictionaries and thesauri enhanced with a process of symmetricality (available at http://dico.isc.cnrs.fr/). For each list of words, one comprised of all words that start with *gl-*, and one comprised of all words derived from the PIE **ghel*, a semantic space is built on the basis of all synonyms and near-synonyms of the words. For *gl-* this resulted in a

[4] This correlation is equivalent to calculating the cosine of the angle formed by the two vectors.

list of 2198 words, and for words derived from PIE this resulted in a list of 1130 words.

The set of cliques containing all these synonyms is calculated. Correspondence factor analysis is applied to the matrix composed of words in the columns and cliques in the lines to obtain the coordinates for each clique (Ploux & Ji 2003). To split the space into clusters, a hierarchical classification is obtained via the calculation of the Ward's distance of cliques' coordinates. A word belongs to a cluster if all the cliques that contain it belong to this cluster.

Results

Word Cluster Cohesiveness with Infomap

We first computed the cohesiveness of the cluster of all words that have been identified as descendents of **ghel* and that of all words that feature the *gl-* phonaestheme. We also computed the cohesiveness of the cluster formed by their intersection, that is, the cluster of words that start with *gl-* and are descended from the **ghel* root. The results of these computations, as well as the cohesiveness of related clusters are given in table 1. Interestingly, all of these clusters show a higher cohesiveness than would be expected by chance alone, as is evident by the fact that all of the *#Sig* measures are above the chance threshold of 15.

Table 1 - The cohesiveness of the **ghel* PIE root and the *gl-* phonaestheme clusters.

N – cluster size; *r* – cohesiveness;
#Sig – number of significant t-tests compared to baseline

Cluster	*N*	*r*	*#Sig*
****ghel* words**	38	.15	100
***gl-* phonaestheme**	88	.097	75
****ghel* words starting with *gl-***	25	.25	100
****ghel* words not starting with *gl-***	13	.046	22
Non-ghel* words starting with *gl-***	17	.15	95

In order to answer our research question, we also compared the clusters to one another. Overall, the results follow the pattern indicated by the relative cohesiveness of the clusters as seen in table 1. The *gl-* phonaestheme as a whole forms a less cohesive cluster than either part of it that is descended from words with a **ghel* PIE root (*#Sig*=28, p<.0001) or the part of it that is descended from words with PIE roots other than **ghel* (*#Sig*=28, p < .0001). However, that same cluster is more cohesive than the cluster comprised of words with a **ghel* PIE root that do not begin with *gl-* (*#Sig*=30, p<.0001). Finally, the cluster formed by words that begin with *gl-* and whose PIE root is **ghel* is stronger than any of the other clusters. More specifically, it is stronger than both the cluster formed by words with a **ghel* PIE root (*#Sig*=55, p<.0001) and that formed by words with a PIE root other than **ghel* (*#Sig*=45, p<.0001).

The most cohesive part of the *gl-* phonaestheme therefore seems to be formed by words with a **ghel* PIE root. Nevertheless, it appears that the set of words starting with *gl-* with other PIE roots also form a cohesive cluster of meaning, even if it is somewhat weaker. This suggests there is more to the phonaestheme than merely a historical root.

Interestingly, the weakest cluster identified in this analysis was formed by words with a PIE root of **ghel* that do not begin with *gl-*. One possible interpretation is that those words having gone through a variety of languages (eg., Greek, Sanskrit) have been subjected to many semantic and morpho-phonological changes creating a disparity in the set. However *gl-*words that relate to light and vision have mostly gone through Germanic, which may explain their high semantic and morpho-phonological cohesiveness.

Word Cluster Cohesiveness and Prototypicality with the SA

**ghel* clustering

Our analysis of the **ghel* data resulted in three main clusters (and a plethora of weak ones). For **ghel*'s main cluster we obtained 649 synonyms of which 609 were relevant[5]. This main cluster is further divided into three sub-clusters and included the central senses of **ghel*: The first sub-cluster (362 terms) relates to the visual modality and to shining. It also contains most *gl-* items (with the exception of terms related to *glide* in cluster 3 as well as *gladden* and *gloaming* in separate clusters). The second sub-cluster (149 terms) relates to melancholy and colors. The third sub-cluster (98 terms) relates to bile, gall and emotional states mapped onto them metaphorically. The last two sub-clusters are significantly separated from the first one.

gl- clustering

From the unstemmed total of 230 *gl-* words, 74 come from PIE **ghel* (32,17%) while in the stemmed list of 106 items 23 do (21,69%). The higher percentage of *gl-* words coming from the root **ghel* in the unstemmed list shows that these items are highly productive in terms of derivation and composition.

The strongest cluster of *gl-* was comprised of 1048 synonyms and was divided into three sub-clusters that form a total of 883 relevant synonyms. The strongest sub-cluster (678 terms) relates to the visual modality. The second sub-cluster (124 terms) relates to gloom and melancholy, and the third (81 terms) relates to the globular shape. Other significant clusters relate to the meanings “glide”, “glue” and “glove”. All other clusters are small and specialized.

Prototypicality

In the **ghel* space, one sub-cluster gathered most of *gl-* based words (38 out of 43) and the other two gather most of non-*gl*-based words. The meanings of light and vision are clearly correlated with the *gl-* phonaestheme, while non-*gl-* item clusters inherit the bulk of other semantic contents associated with **ghel*. The historic root clearly evolved into a *gl*-based conceptual network related to light and vision,

[5] 'Relevant' synonyms are in the cliques that only belong to one given cluster. Conversely some highly polysemous cliques belong to several clusters.

Table 2 - Prototypicality in the strongest clusters of the **ghel* space and the *gl-* space

***ghel**	*Sub-Cluster*	*In # of cliques*	*%*	**gl-**	*Sub-Cluster*	*In # of cliques*	*%*
glow	1	55	19%	**gleam**	1	59	10%
glitter	1	48	16%	**glow**	1	55	10%
glowing	1	48	16%	**shine**	1	51	9%
melancholy	2	52	53%	**gloomy**	2	85	64%
sad	2	25	25%	**dismal**	2	39	29%
yellow	2	17	17%	**dark**	2	37	28%
gall	3	58	81%	**globe**	3	20	50%
virulence	3	19	26%	**ball**	3	13	33%
bitterness	3	18	25%	**orb**	3	11	28%

while secondary meanings were distributed across non-*gl-* items.

Clusters classify words in decreasing order of importance: the ones that belong to a high number of cliques are considered to be more prototypical. Table 2 shows the 3 most prototypical items of **ghel* and *gl-*'s main clusters. The percentage denotes the number of cliques the item belongs to on the total of cliques composing the cluster.

In the *gl*-space, one sub-cluster gathers most **ghel*-based words (65 out of 82), while the two others gather a smaller number of then (8 in sub-cluster 2, and 9 in sub-cluster 3). Again the first sub-cluster is the largest and corresponds to the central meaning of the *gl-* phonaestheme, while the two others relate to antinomic and secondary meanings. The phonaestheme clearly divides into a major conceptual unit versus minor units mostly unrelated to the historic root.

Cohesiveness and semantic distances

We used independent samples t-tests to examine the semantic cohesiveness of **ghel* words within the *gl-* space and similarly for *gl-* words within the **ghel* space.

In the **ghel* space, the average semantic distance within the *gl-* cluster is lower than the average distance between the *gl-* and non-*gl-* clusters (M_{intra}=0.39, M_{inter}=1.65, $t(219)$=9.86, p<.0001). However, no significant difference was found between the non-*gl-* cluster and the overall **ghel*-set (M_{intra} =1.85, M_{inter}=1.66, $t(93)$=0.61, *n.s.*). Non-*gl-* items are therefore disparate and less cohesive than the *gl-* phonaestheme.

In the *gl-* space, words that have the same PIE root show higher cohesion than words that do not (M_{intra}=0.15, M_{inter}=1.81, $t(556)$=9.82, p<.0001). Words that are **ghel* based are more cohesive than the whole *gl-* space as the average distance between the **ghel* set and other PIE roots is lower than the internal average distance within the **ghel* set. (M_{intra}=0.13, M_{inter}=3.31, $t(187)$=2.36, p<0.05)

These results are congruent with the previous analysis, as the strongest cohesiveness is found in the set that is both *gl-* and **ghel* based.

General Discussion

In this paper we show that, in the case of *gl-/*ghel*, historical (here PIE) and morpho-phonological (here phonaesthemes) aspects are autonomous but highly correlated and that both have a tangible impact on word meaning. More specifically, we showed that phonaesthemic sets have a higher cohesiveness within historical sets and historical root sets have a higher cohesiveness within phonaesthemic sets.

These results suggest that the lineage of a language plays an important role in the distribution of linguistic meaning. In particular, the phonaestheme *gl-* seems to be based on the PIE root **ghel*. It therefore seems clear that, at least in some cases, historical information influences the distribution of word meaning in non-trivial ways. One reason for this could be that lexical items are linked to conceptual networks that are rooted in history. By incorporating historical and etymological information into statistical models such as word-space vectors or clique-based synonym sets we might improve their performance.

The conceptual networks visible for *gl-* words keep traces of older semantic content, notably the fact that verbs starting with *gl-* and related to light or vision can have two arguments, an animate one (as in *glance*) or an inanimate one (as in *glow*). This particular aspect relates vision to light emission and participates in creating a semantic unity contrary to modern beliefs that clearly separates emitting light from perceiving it (cf. Philps, 2008a). However, at this point it is unclear what the cognitive value of these semantic traces is and how it relates to the role of language as a means for decoding the world.

Interestingly, some words of obscure origin have high productivity although they cannot be traced back to PIE. One example of this is the word *globe* which seems related to the visual modality, though there is no historical evidence for such a connection. This gives rise to a new question – How do newly formed words find their place within an existing conceptual network? It may be that new additions to the vocabulary are likely to be patterned after existing words in a manner that makes them compatible with the rest of the set. New words which contain an existing phonaestheme are likely to fit its conceptual pattern as well.

In this paper we focused on examining the role that language history and sound/form relationships might play in conceptual organization in the case of **ghel/gl-*. Our results suggest that such analyses can provide important insights into the inter-relation of semantic concepts. In particular, it seems some aspects of meaning may be more stable than others. However, at this point it is not clear whether this stability is attributable to some fundamental characteristics of human cognition or to the broader social contexts in which language is used.

Moreover, using this information and integrating it with current distributional models is not a trivial task, and several different routes seem to present themselves. A possible route might involve defining a new, etymological, index that could be used to enrich current models of conceptual organization and semantic similarity. Finally, it seems that a better understanding of how languages change and evolve might lead to a better understanding of the interrelation between language, culture, and cognition.

References

Bergen, B. (2004). The Psychological Reality of Phonaesthemes. *Language*, 80(2), 291-311.

Bolinger, D. (1950) Rime, assonance, and morpheme analysis. *Word* (6), 117-136.

Buckley, C., Singhal, A., Mitra M., & Salton, G. (1996) New retrieval approaches using SMART:TREC4. *Proceedings of the Fourth Text Retrieval Conference (NIST Special Publication 500-236)*, 25-48.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25, 211–257.

Fellbaum, C. (1998) *WordNet, an electronic lexical database*. MIT Press, Cambridge, MA.

Firth, J. (1930) *Speech*. London: Oxford University Press.

Jackendoff, Ray, 1983. Semantics and Cognition, Cambridge, Massachusetts: MIT Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2), 211–240.

Levin, E., Sharifi, M., & Ball, J. (2006) Evaluation of utility of LSA for word sense discrimination. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City. 77-80.

Louwerse, M. M., Hu, X., Cai, Z., Ventura, M., & Jeuniaux, P. (2005). The embodiment of amodal symbolic knowledge representations. In I. Russell & Z. Markov (Eds.), *Proceedings of the 18th International Florida Artificial Intelligence Research Society*, pp. 542–547. Menlo Park, CA: AAAI Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28,* 203-208.

Otis, K. & Sagi E. (2008) Phonaesthemes: A corpora-based analysis. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society.*

Marchand H. (1960) *The categories and types of present-day English word-formation.* AL: University of Alabama Press.

Marcu, D (2003) Automatic Abstracting, In Drake, M. A., (ed), *Encyclopedia of Library and Information Science*, pp. 245-256.

Philps, D. (2008a) Sons et lumières: le marqueur sub-lexical <*gl*->. In G. Girard-Gillet (ed), *L'envers du décor, Études de linguistique anglaise*. Avignon, Publication des Presses de l'Université d'Avignon, pp. 24-43.

Philps, D. (2008b) From mouth to eye. in A. Smith, K. Smith & R. Ferreri (eds.), *The Evolution of Language*, Singapore: World Scientific Publishing, pp. 251-258.

Ploux, S. (1997) Modélisation et traitement informatique de la synonymie. *Linguisticae Investigationes*, 21(1):1-28.

Ploux, S. & Ji, H. (2003) A Model for Matching Semantic Maps between Languages (French/English, English/French), *Computational Linguistics*. 29(2):155-178.

Ploux, S., Victorri, B. (1998) Construction d’espaces sémantiques à l’aide de dictionnaires de synonymes. *TAL*, 39, n°1.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In Basili R., and Pennacchiotti M. (Eds.), *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*. Athens, Greece.

Schütze, H. (1996) *Ambiguity in language learning: computational and cognitive models*. CA: Stanford.

Takayama, Y., Flournoy, R., Kaufmann, S. & Peters, S. (1999). Information retrieval based on domain-specific word associations. In Cercone, N. and Naruedomkul K. (eds.), *Proceedings of the Pacific Association for Computational Linguistics (PACLING’99)*, Waterloo, Canada. 155-161.

Watkins Calvert. (2000). *The American Heritage Dictionary of Indo-European Roots.* Second Edition. Houghton Mifflin Harcourt Compagny.

Appendix A – Sample words used in this study

PIE Root	**Words**
*ghel	yellow, melancholy, gulden, guilder, gowan, gold, glow, gloss, gloat, gloam, glitter, glister, glisten, glissade, glint, glimpse, glimmer, glide, glib, gleg, gleeman, gleed, glee, glede, gleam, glaze, glass, glare, glance, glad, gill, gild, gall, felon, cholera, choler, chloroform
***Dḷk-u-**	glucose, glycerine
***gel-**^{2}	Glace
***gladh-**	glabrous
***glei-**	glue, gluten, glutinous
***glôgh-**	glossa, glottis
***gwelə-**^{2}	gland, glans
***kel-**^{1}	gladiator , gladiolus
***kelə-**^{2}	Glairy
***lep-**^{2}	Glove
Unknown root	glacier, glade, glam, glamour, glaucoma, glean, glebe, glen, gloaming, globe, gloom, gloriosa, glory, glout, glucinum, glum