DANTE: a New Resource for Research at the Syntax-Semantics Interface

Diana McCarthy
Lexical Computing Ltd,
Brighton, UK
diana@dianamccarthy.co.uk

Abstract

Since Levin's seminal work (Levin, 1993) there has been a rising interest in computational linguistics research which aims to examine the relationship between the syntax and semantics of verbs. A substantial portion of the work comprises efforts to discover semantic classes from syntactic behaviour and also from selectional preferences. There is also some work on directly examining related phenomena, such as detecting subcategorisation frames and diathesis alternations. Work in this area is typically corpus based, although many manually constructed resources have also been used as start points and for evaluation. In this paper, we present an English lexical database (being finalised at the time of writing, and to be released late 2010) which we believe will be a major catalyst for work of this nature, both as a starting point for automatic methods and as a gold standard for evaluation

1 Introduction

There has been a growing interest in computational linguistics in the semantics-syntactic interface, particularly as regards verbs. A trigger for this was Levin's work (Levin, 1993) on verbs which, following her predecessors e.g (Fillmore, 1967), demonstrated that given that a verb's meaning is related to its syntactic behaviour, we can group verbs into semantic classes by virtue of their shared syntactic behaviour. A key issue in any research on this relationship is identifying what the key syntactic behaviour and semantic components are since there are a great many possibilities and it is a non trivial task to identify the appropriate features. Diathesis alternations are different surface realisations of a verbs arguments. Levin's work demonstrated that diathesis alternations are extremely useful in classifying verbs.

Levin's alternation inventory, whilst the first of its kind and providing a broader and more thorough manual analysis than anything that had been been available before, was restricted to a subset of subcategorisation frames (SCFs) involving NPs and PPs, i.e excluding sentential complements. The resource was produced manually and not from corpus examples. Baker and Ruppenhofer (2002) point out that many examples of syntactic behaviour Levin provides, are not attested in the corpus data (the BNC (Leech, 1992)) that they used for the FrameNet project. Furthermore, actual use of alternations for verb classification would give rise to a finer granularity than is present in Levin's classification; many of Levin's classes are semantically motivated, rather than being totally determined by the alternation behaviour. Despite these limitations, the book has triggered a large amount of research in computational linguistics in automatically identifying the links between syntactic behaviour and verb meaning.

Prior to the work on automatic classification, there was research on automatic acquisition of verbal information from corpora that would in turn be exploited for subsequent work on classification. Acquisition of SCFs (Brent, 1991; Manning, 1993) was conducted with a view to improving results in parsing (Carroll et al., 1998). Selectional preference acquisition (Resnik, 1993) was performed

to help with structural and lexical ambiguity resolution (Li and Abe, 1998; Resnik, 1997; McCarthy and Carroll, 2003). Levin's work spurred further research using automatically acquired lexical information for diathesis alternation identification (McCarthy, 2000; McCarthy and Korhonen, 1998; Lapata, 1999) and for verb classification (Schulte im Walde, 2006; Sun and Korhonen, 2009; Stevenson and Merlo, 1999; Merlo and Stevenson, 2001).

In this paper we will give a very brief overview of the lexical acquisition work in this direction [1], and a summary of some of the key existing lexical resources that can be used as input to the work or for evaluation purposes. We then describe DANTE (Atkins et al., 2010) a recently released lexical database produced by a team of lexicographers scrutinising a 1.7 billion word corpus of English. The database includes over 6,300 headword verbs with just under 3000 phrasal verbs with just under 300,000 examples of the various features of these verb and phrasal verb entries. [2] We expand on the potential of this resource for lexical research and we end by highlighting the possibilities for integration of DANTE with existing lexical resources to further its potential yet still.

While there is interesting related work in other languages (Schulte im Walde and Brew, 2002) the bulk of the resources and lexical acquisition work in this area has been with regard to English. DANTE presented here is also an English resource. For this reason, this paper will focus on how DANTE relates to English resources. Fully automatic methods that simply use such resources for evaluation are in many cases applicable to languages other than English.

2 Background: automatic acquisition of verbal subcategorisation, selectional preferences, diathesis alternations and semantic class

We will highlight some key contributions, but unfortunately have not been able to include all due to lack of space.

[1] Related topics of semantic role labelling, word sense induction and word sense disambiguation are outside the scope of this paper.

[2] There is likewise a wealth of information and examples for other PoS, but we do not go into those details here.

2.1 Automatic Acquisition of SCF and Selectional Preferences

There have been many works on automatic acquisition of SCFs. The earliest is due to Brent (1991) who proposed a system capable of recognising five frames, using information from unambiguous cases, for example using pronouns for detecting noun phrases. Following this pioneering work there has been increasing attention paid to a more comprehensive classification, and coverage of more data using statistical techniques to filter parser errors. Briscoe and Carroll (1997) developed a system distinguishing 161 SCFs and, because it is not restricted to unambiguous input, can output relative frequencies of these frames for a given verb. Korhonen (2002) made various refinements of the system, including use of Levin style verb classes to improve statistical filtering to distinguish genuine frames from parser noise. Preiss et al. (2007) extended this approach to adjective and nominal frames.

Alongside the acquisition of SCFs, work has been conducted on selectional preference acquisition using data in the argument heads of these frames (McCarthy, 2000), or directly on parser output (Resnik, 1993; Li and Abe, 1998). Erk (2007) used example sentences from FrameNet as input to selectional preference acquisition. Early work used WordNet to provide classes for generalisation of the preferences (Resnik, 1993; Li and Abe, 1998; Clark and Weir, 2002), but more recently there has been work using distributional similarity for generalisation (Erk, 2007; McCarthy et al., 2007)

2.2 Automatic Identification of Verbal Participation in Diathesis Alternations

Resnik (1993) demonstrated a link between selectional preference strength and participation in alternations where the direct object can be omitted. e.g. *The boy ate the popcorn.* $\leftrightarrow$ *The boy ate*. Lapata (1999) identified participation in the dative and benefactive alternations using a shallow parser and various linguistic and semantic cues, which are specified manually for these two alternations. Another approach is to use cues for syntactic frames, coupled with the overlap of lexical fillers between the alternating slots. McCarthy and Korhonen (1998) carried out preliminary experiments which were extended by McCarthy (2000) on detecting 'role switching al-

ternations'. Role switching alternations are defined as those where an argument appears in different slots in different frames, examples are the causative and conative alternations. McCarthy and Korhonen (1998) and McCarthy (2000) used WordNet as a means of generalising the lexical fillers to semantic classes and used Li and Abe (1998)'s selectional preference models to find semantic classes with an appropriate level of generalisation. Tsang and Stevenson (2010) extended this work by a graphical method which compares the probability of the lexical items at the alternating slots in the WordNet hypernym structure as a whole rather than at a set of individual classes cutting across that structure. Using this method they demonstrated an improvement on (McCarthy, 2000), particularly with regard to low frequency verbs.

2.3 Automatic Identification of Verb Classes

In this subsection, we describe approaches which classify verbs according to evidence often also used for diathesis alternation detection, however alternation participation is not overtly detected in these methods. Merlo and Stevenson (2001) detected three major classes of optionally intransitive verbs (unergative unaccusative and object drop) verbs based on argument structure using corpus evidence of transitivity, causativity and animacy of the arguments as well as other surface features such as passivisation. Schulte im Walde (2006) demonstrated that SCF can be used for clustering German verbs. She also experimented with selectional preferences using GermaNet (Kunze and Lemnitzer, 2002) but without finding a significant improvement over syntactic information alone. More recently, (Sun and Korhonen, 2009) demonstrated that unsupervised clustering of the argument heads themselves can be used as selectional preference features which in turn improved the clustering of the verbs when used alongside SCFs in contrast to the SCFs features alone.

3 Lexical Resources Available for Research

The focus here is on verbal information. Note that DANTE and FrameNet also provide a wealth of information on other PoS.

Levin's classification A classification of 3100 verbs into 193 classes based on verbal participation in 80 diathesis alternations, involving mainly NP and PP constituents. This classification was produced manually and examples were obtained from introspection rather than corpus evidence.

VerbNet (Kipper-Schuler, 2005) (Now extended VerbVet) A verbal lexicon comprising 3769 lemmas with 5257 senses organised in hierarchical WordNet classes but supplemented with valuable syntactic information as well as thematic roles and selectional preferences

Propbank (Palmer et al., 2005) A one million word corpus which supplements the Penn Tree Bank (Marcus et al., 1993) and has been annotated with predicate-argument information. The semantic role labels assigned to arguments have meanings that are specific to each verb. This resource is particularly useful for research in semantic role labelling (Màrquez et al., 2008). Although the corpus is currently limited to Wall Street Journal News text, there is work underway to annotate further corpus data.

Valex (Korhonen et al., 2006) This is an automatically produced SCF lexicon of 6397 verbs using the system of Korhonen (2002) on a corpus of 900 million words. A portion of the output has been evaluated but the lexicon is automatically produced and each individual corpus occurrence has not been validated.

FrameNet (Ruppenhofer et al., 2010) is a lexicon produced from analysed texts that places lexical units (senses) in semantic frames, for example **removing** or **emptying** which classify verbs (and nouns and adjectives) according to the semantic frames that they participate in. Examples are provided from the BNC and an American newswire corpus. The database currently includes 135,000 corpus sentences for over 10,000 lexical entries (nouns, verbs and adjectives) in approximately 800 frames.

WordNet (Fellbaum, 1998) A list of 11529 verbs [3] (including multiword expressions marked as verbs) with synonyms and semantic relations marked. Although there is some information on derived forms and some domain tags, the resource is focused on senses

[3] Here we refer to the latest version of WordNet: version 3.0

and semantic relationships e.g. troponymy and entailment, and does not include syntactic, grammatical and collocational behaviour.

In the following section we describe DANTE, a new lexical database built from inspection of 1.7 billion word corpus.

4 DANTE

DANTE (Database of ANalysed Texts of English) [4] was produced during the first stage of production of a New English Irish Dictionary, and is funded by Foras na Gaeilge, the official body for the (Gaelic) Irish language. DANTE is a target-language-neutral monolingual analysis of the source language listing all the phenomena that might possibly have an unexpected translation. DANTE is a collection of lexical entries with information and examples on every variety of lexical information that the lexicographers have deemed potentially relevant for a thorough and accurate description of English. DANTE relates to the Corpus Pattern Analysis approach of Hanks (Forthcoming) in that a major focus is the prototypical syntagmatic patterns of words in use.

The project team combined expertise in corpora, computational linguistics and lexicography, and from the very outset the project has been solidly corpus-based The corpus used comprised 1.7 billion words from the UKWaC (Ferraresi et al., 2008), some contemporary American newspaper text and Irish English data from the NCI (Kilgarriff et al., 2006). This data was then part-of-speech tagged with TreeTagger [5] and loaded into the Sketch Engine corpus query system (Kilgarriff et al., 2004).

The distinctive feature of the Sketch Engine is 'word sketches': one-page, corpus-driven summaries of a word's grammatical and collocational behaviour. The corpus is parsed using a simple tag sequence grammar and a table of collocations is extracted for each grammatical relation. For DANTE, the set of grammatical relations was defined to give an exact match to the grammatical patterns that the lexicographers were to record. The word sketch for the word would, in so far as the PoS-tagging, parsing, and statistics worked correctly, identify precisely the grammatical patterns and collocations that the lexicographer needed to note in the dictionary. Figure 1 shows a smallish portion of the word sketch for the verb *blend*. The interface allows for seamless switching between specific collocations in the word sketch and a concordance containing those collocations. This switching from the word sketch to concordance is extremely useful for finding examples of significant phenomena. A key feature of DANTE, is that all lexical information is supplemented with example sentences from the corpus. The examples were not edited making them ideal for building and evaluating robust computational linguistics systems which can cope with real language. In order to help the lexicographers find good examples for the phenomena under scrutiny an automatic program (GDEX) that is part of the sketch engine suite of tools was used for sorting the examples so that the 'best' (according to a set of heuristics) are shown to the lexicographer first (Kilgarriff et al., 2008).

4.1 Lexical Information within DANTE

For a full description of the contents of DANTE, refer to the web site [6] and (Atkins et al., 2010). Here we provide a summary of information pertinent to automatic lexical acquisition of verbs. [7] Note that all the subsequent categories of information are associated with word senses.

senses Lexicographers break headwords into senses based on corpus evidence and provide examples of each, along with brief definitions. The definitions are designed to differentiate one sense from another within the same entry for a given lemma and are not as polished as they would be in a conventional dictionary. The focus in DANTE is on comprehensive corpus citations as examples of all lexical information. Extensive exemplification of senses are potentially more useful to computational approaches compared to definitions which are produced for human readers.

subcategorisation frames There are 42 frames in total for verbs, with additional specification of preposition (see figure 2). These are based on the work of Charles Fillmore and are described in (Atkins et al., 2003).

[4] DANTE is described at www.danteweb.com where you can also find a interface for querying the database.

[5] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://www.webdante.net/

[7] In this paper we provide our own labels for information provided in DANTE.

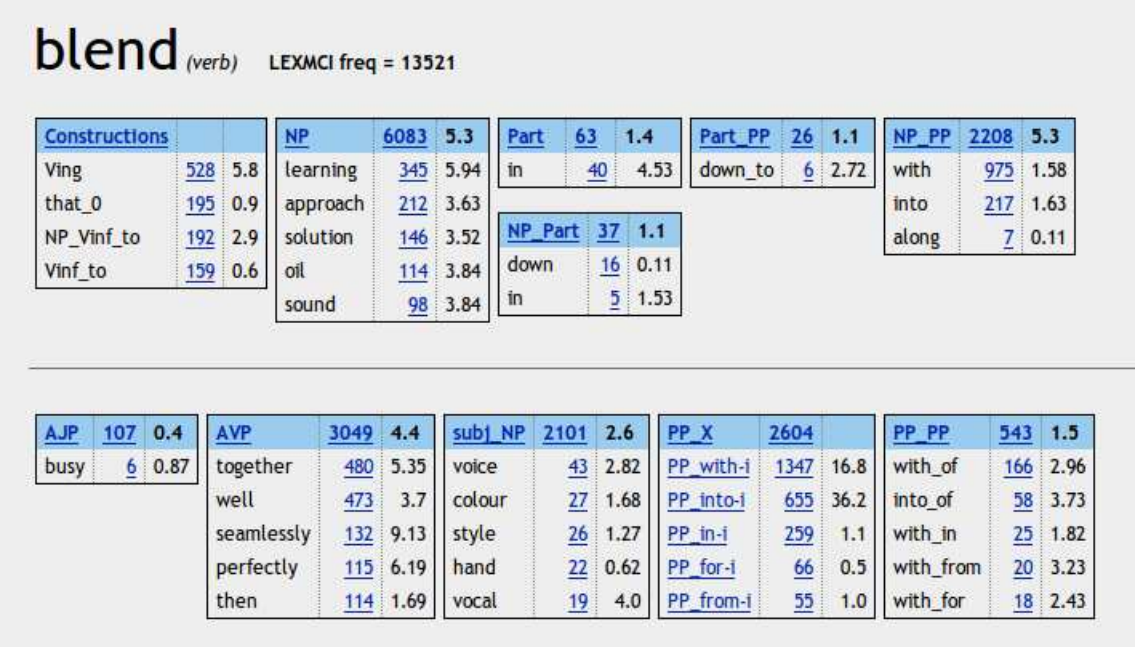

Figure 1: A portion of the word sketch for *blend*.

inherent grammar e.g. *rain* impersonal

multiword expressions including idioms, support verbs, phrasal verbs, compounds, chunks

collocations e.g. *fire* (discharge a weapon) NP collocations *shot, round, gun* ...

corpus patterns tendencies e.g. plural noun as object

usage markers include:

- evaluative e.g. *meddle* (pejorative)
- regional variety e.g. *nick* (British) as in *you're nicked*
- domain e.g. *multiply* (maths)

4.2 DANTE as a Resource for Research at the Syntax-Semantics Interface

DANTE is being released without charge for research purposes. For computational linguistics, and perhaps also other linguistics research it is the combination of syntactic, semantic and usage information alongside numerous examples that makes DANTE stand out in contrast to previously available resources. While some existing resources do have corpus examples (Ruppenhofer et al., 2010; Palmer et al., 2005), DANTE provides a far greater number (300,000 for verb and phrasal verb entries alone) and from a far larger and more and varied source (in contrast to previous resources with examples from the BNC (FrameNet) or the Wall Street Journal (Propbank)) with manual verification of the data (in contrast to automatically produced resources such as valex (Korhonen et al., 2006)). This makes it a perfect resource for systems which experiment with data exhibiting specific phenomena e.g. particular SCFs for diathesis alternation detection contrasting argument fillers at different slots. For example, the PP slots in the two NP_PP_X frames with prepositions *with* and *into* as exemplified in figure 2. While it is of course possible to use automatic resources as a start point (McCarthy, 2000) use of DANTE would enable researchers to isolate PoS, parser error and other sources of noise that are difficult to avoid (Korhonen et al., 2000) when using fully automatic methods.

In addition to the 300,000 verbal manually verified corpus examples [8] it is possible to obtain further examples direct from the 1.7 billion word corpus using the SCF and collocation information. Indeed, this information is already being used in a preliminary word sense disambiguation project. [9] Computational linguistic approaches for selectional preference and diathesis alternation acquisition could use the data to gather argument heads in specific slots of SCFs. Since all the data is assigned to word senses, and the word senses

[8] There are 622,000 examples over all PoS.

[9] See http://www.webdante.com/disambiguation_project.html.

blend: (PoS: v)
`meaning:` combine
`SCF:` NP
`corpus pattern:` with plural noun as object
`example:` *I have very little idea of how to* ***blend*** *colour.*
`corpus pattern:` blend sth and sth
`example:` *High Points : The attempt to* ***blend*** *melodrama comedy and horror is a worthy if failed effort.*
`SCF:` NP_PP_X with
`example:` *Kazakhstan was interested in* ***blending*** *palm oil with its own cotton seed and sunflower seed oils for industrial application , officials said.*
...
`SCF:` NP_PP_X into
example: *I* ***blend*** *different colours into the background of my paintings to evoke sections of light .*

Figure 2: A portion of the entry for *blend*. The portion has been simplified and shortened for presentation here, with only a couple of examples and features shown. Further examples are provided at http://www.webdante.net/.

have associated usage information, there is scope for doing experiments linking sense to syntactic behaviour. Moreover, as well as a start point for acquisition, the resource can be used as a gold standard for evaluation of automatic acquisition of information contained therein such as SCF, sense induction, sense disambiguation and usage, for example domain.

5 Conclusion

In this paper we have presented the DANTE lexical resource which we believe will prove a useful resource for computational linguistics, particularly at the syntax-semantics interface but elsewhere also. We have suggested ways in which the data therein could be used as a starting point for research at the syntax-semantics interface, for example alternation detection and selectional preference acquisition, and also as a resource for lexical acquisition evaluation.

There are a multitude of resources for English dealing with predicate-argument structure and word sense. No one resource is a panacea and researchers have already highlighted the merits of combining resources (Merlo and van der Plas, 2009). SemLink [10] is a great initiative in this direction with mappings between VerbNet and propbank and VerbNet and FrameNet. Atkins (2010) proposes possibilities in this direction for combining DANTE with FrameNet using syntactic information common to both and distributional thesauruses (such as those in Sketch Engine) for relating lexical units. We believe that interesting research will result from such endeavours and that, as well as automatic approaches for linking these resources should prove interesting in their own right.

[10] http://verbs.colorado.edu/semlink/

Acknowledgments

Thanks to my colleagues on the DANTE project, particularly Sue Atkins, Adam Kilgarriff, Cathal Convery, Michael Rundell, Diana Rawlinson and Valerie Grundy. I am indebted to Sue Atkins, Adam Kilgarriff and Anna Korhonen for comments on an earlier draft. Any mistakes in this article are solely my own responsibility.

References

Sue Atkins, Charles Fillmore, and Christopher Johnson. 2003. Lexicographic relevance: selecting information from corpus evidence. *International Journal of Lexicography*, 16(3):251–280.

Sue Atkins, Michael Rundell, and Adam Kilgarriff. 2010. Database of ANalysed Texts of English (DANTE). In *Proceedings of Euralex*.

Sue Atkins. 2010. The DANTE database: Its contribution to English lexical research, and in particular to complementing the FrameNet data. In Gilles-Maurice de Schryver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha.

Collin F. Baker and Josef Ruppenhofer. 2002. FrameNet's frames vs. Levin's verb classes. In J. Larson and M. Paster, editors, *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, pages 27–38.

Michael R. Brent. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 209–214.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 356–363.

John Carroll, Guido Minnen, and E. Briscoe. 1998. Can subcategorisation probabilities help a statistical parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126, Bergen, Norway.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Charles Fillmore. 1967. The grammar of *hitting* and *breaking*. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 120–133. Ginn and Company: a Xerox Company, Waltham M.A.

Patrick Hanks. Forthcoming. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of Euralex*, pages 105–116, Lorient, France.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the new corpus for Ireland. *Language Resources and Evaluation Journal*, 40(2):127–152.

Aadam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychly. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Euralex Proceedings*, Barcelona.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*, pages 199–206, Hong Kong. ACL.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy.

Anna Korhonen. 2002. *Subcategorization Acquisition*. Ph.D. thesis, University of Cambridge.

Claudia Kunze and Lothar Lemnitzer. 2002. Germanetrepresentation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain.

Maria Lapata. 1999. Acquiring lexical generalizations from corpora: a case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 397–404.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Beth Levin. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago and London.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Christopher Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242.

Mitchell Marcus, Beatrice Santorini, and Mary Marcinkiewicz. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2).

Lluís Màrquez, Xavier Carreras, Ken Litkowski, and Suzanne Stevenson. 2008. Semantic role labelling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Diana McCarthy and Anna Korhonen. 1998. Detecting verbal participation in diathesis alternations. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1493–1495.

Diana McCarthy, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics. (NAACL)*, pages 256–263, Seattle,WA.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3):373–408.

Paola Merlo and Lonneke van der Plas. 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore, August. Association for Computational Linguistics.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106, sept.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 912–919, Prague, Czech Republic, June. Association for Computational Linguistics.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the* SIGLEX *Workshop on Tagging Text with Lexical Semantics: Why What and How?*, pages 52–57, Washington, DC.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended theory and practice. Technical report, International Computer Science Institute, Berkeley, June. http://framenet.icsi.berkeley.edu/.

Sabine Schulte im Walde and Chris Brew. 2002. Inducing german semantic verb classes from purely syntactic subcategorization information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the Nineth Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–52.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Singapore, August. Association for Computational Linguistics.

Vivian Tsang and Suzanne Stevenson. 2010. A graph-theoretic framework for semantic distance. *Computational Linguistics*, 36(1), March.