Automatic lexical classification: bridging research and practice

Anna Korhonen

References	**This article cites 8 articles** http://rsta.royalsocietypublishing.org/content/368/1924/3621.full.html#ref-list-1
Rapid response	Respond to this article http://rsta.royalsocietypublishing.org/letters/submit/roypta;368/1924/3621
Subject collections	Articles on similar topics can be found in the following collections human-computer interaction (9 articles)
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

Review

Automatic lexical classification: bridging research and practice

By Anna Korhonen*

*Computer Laboratory and RCEAL, University of Cambridge,
15 J J Thomson Avenue, Cambridge CB3 0FD, UK*

Natural language processing (NLP)—the automatic analysis, understanding and generation of human language by computers—is vitally dependent on accurate knowledge about *words*. Because words change their behaviour between text types, domains and sub-languages, a fully accurate static lexical resource (e.g. a dictionary, word classification) is unattainable. Researchers are now developing techniques that could be used to automatically acquire or update lexical resources from textual data. If successful, the automatic approach could considerably enhance the accuracy and portability of language technologies, such as machine translation, text mining and summarization. This paper reviews the recent and on-going research in automatic lexical acquisition. Focusing on lexical classification, it discusses the many challenges that still need to be met before the approach can benefit NLP on a large scale.

Keywords: natural language processing; automatic lexical acquisition;
lexical classification

1. Introduction

Natural language processing (NLP) is a growing, interdisciplinary field of computer science that develops key technologies for analysing, understanding and generating human language, as well as useful applications to support the processing, mining and extraction of knowledge from large collections of textual data (e.g. text mining, summarization, question answering, machine translation, human–computer dialogue). Looking into the future, NLP is a particularly timely and important field to work on. Cognitive scientists have not yet succeeded in decoding the enigma of human language acquisition and processing. Providing rich models of language use, NLP can contribute to solving this problem, which is of interest for a wide range of scientific disciplines. Owing to the growing problem of information overload, there is also a great demand for NLP-based applications in many areas of society (e.g. communication, healthcare, science).

*alk23@cam.ac.uk

After decades of research, basic NLP techniques are now sufficiently developed to be integrated into practical applications. Yet, a major intellectual challenge is still involved in improving and tuning the techniques further to meet the needs of demanding real-world applications.

Achieving a major breakthrough in NLP is difficult because human language is highly complex, ambiguous, subtle and subject to constant change. Embedded in a social, cultural and physical context, it is challenging to model using computers. When aiming to improve NLP further, one key area that requires improvement is the lexicon. Successful language processing requires accurate information about words, i.e. the structure, meaning and relative frequency of nouns, verbs, adjectives and other types of words in specific texts. For example, accurate machine translation of legal texts from English to French requires a dictionary (one for each language) specific to the legal language. A dictionary specific, for example, to financial, sports, biomedical or general language would exclude relevant terminology or include it but with wrong meanings (e.g. the verbs *prohibit*, *bind* and *regulate* have entirely different meanings in legal and biomedical texts).

High-quality *lexical resources* (e.g. dictionaries, word classifications, thesauri) are therefore core components of many language technologies. Currently, most lexical resources used in specific systems are developed manually by linguists. Manual lexicography is extremely costly, taking several person years to complete. In addition, since lexical information varies between sub-languages, domains and over time, manually built resources require extensive labour-intensive porting to new NLP tasks. They often also lack information important for NLP systems that is difficult to gather by hand, such as statistical information about the likelihood of words and their meanings in specific texts.

The ultimate solution to computational lexicography would be an automatic system that takes relevant (e.g. legal) texts as input, returns a lexicon including all and only the relevant words as output, and then possibly supplements this lexicon with additional information from manually built dictionaries. Such automatic acquisition or updating of lexical information from relevant repositories of text (such as the Web, corpora of published text, etc.) can avoid the expensive overhead of manual work. This approach is now viable and gathers statistical information as a side-effect of the acquisition process. The statistical information can easily be adapted to new domains and usage patterns, provided that relevant corpus data are available. The resources and techniques required for the automatic approach are now available. Several large corpora have been constructed for many languages (e.g. the two-billion-word Gigaword corpus for English), along with the Web and large databases of domain-specific articles (e.g. the Medline database for biomedical articles, see http://medline.cos.com/). The methods for automatic text analysis and machine learning have now also developed to the point that they can be usefully deployed.

After two decades of intensive research, advances have been made in many areas of lexical acquisition, including the development of techniques for identifying and extracting terms, word senses, lexical classes, semantic relations and multi-word expressions, among many others, in texts (for surveys of recent research see Villavicencio *et al.* 2005; Agirre & Edmonds 2006; McCarthy 2006; Marquez *et al.* 2008; Schulte im Walde 2009). Although the acquisition of some types of information (e.g. word senses) has proved

more challenging than that of others (e.g. terms), the best available techniques are now capable of extracting large-scale lexical frequency data with promising accuracy.

Despite this, the practical usefulness of lexical acquisition for NLP remains largely unattested. Only a few publicly available lexicons have been built using this technology, and there has been little task-based evaluation or application of the techniques to real-world NLP systems. For example, current machine translation technology makes no use of automatically acquired lexical information. The fact that nearly no acquisition technology has moved from research laboratories into widespread application after decades of research is surprising, given that this line of research is largely motivated by the need to produce better lexicons for practical tasks. We discuss the reasons for this situation, and highlight the challenges that need to be addressed before the automatic approach can be used to enhance the performance and portability of NLP technology on a large scale.

Since the scope of this paper is limited, we will illustrate our discussion with one sub-area of lexical acquisition that can be particularly useful for NLP applications and represents many of the issues typical for the area at large: *lexical classification*. The following section presents an overview of this sub-area. It acts as an introduction to the subsequent section, which discusses the current challenges and the future of lexical acquisition.

2. Automatic lexical classification: the state of the art

Lexical classes, defined in terms of shared meaning components and similar syntactic behaviour of words (Levin 1993), have attracted a great deal of interest in NLP. These classes are particularly useful for their ability to capture generalizations about a range of linguistic properties. For example, 'manner of motion' verbs, such as *travel*, *run* and *walk*, not only share the meaning of 'manner of motion', but also behave similarly in texts, e.g. they appear in similar syntactic frames, such as *I travelled/ran/walked*, *I travelled/ran/walked to London* and *I travelled/ran/walked five miles*. Lexical classes can be identified across the entire lexicon (e.g. 'change of state', 'manner of speaking', 'sending', 'removing', 'learning', 'building' and 'psychological' verbs, among many others) and they may also apply across languages.

Such classes can benefit NLP systems in a number of ways. One of the biggest problems in NLP is the sparse-data problem: for many tasks, only small text corpora are available, and many words are rare even in the largest corpora. Lexical classifications can help to compensate for this problem by predicting the likely syntactic and semantic analysis of a low-frequency word. For example, if *simple* occurs infrequently in the data in question, the knowledge that this word is likely to belong to the class of 'easy' adjectives will help in predicting that it takes similar syntactic frames to the other class members (e.g. *difficult, convenient*). This can improve the likelihood of correct syntactic analysis, which can in turn benefit any NLP system that employs parsing (e.g. information extraction, machine translation). Similarly, the knowledge that the verbs *activate, induce, stimulate* and *up-regulate* are

likely to belong to the class of 'activate' verbs in biomedical texts can help in identifying semantically similar statements in texts (e.g. those describing activation events) and this, in turn, can improve information extraction from biomedical literature.

Offering a powerful tool for generalization, abstraction and prediction, lexical classifications have been used to support many important NLP tasks, including, for example, computational lexicography, parsing, word sense disambiguation, semantic role labelling, information extraction, question answering and machine translation (Kipper *et al.* 2008). However, the exploitation of classes in real-world or highly domain-sensitive tasks has been limited because only general, manually built classifications are available. The largest such classification is VerbNet (Kipper-Schuler 2005). Building on the well-known classification of Levin (1993), VerbNet summarizes decades of theoretical research on English verb classification. It classifies over 5000 verbs into 274 first-level classes based on their syntactic–semantic properties. Manual extension and tuning of VerbNet to different domains have proved very costly because class-based differences are manifested in differences in the statistics over usages of a variety of syntactic–semantic features. This information is time-consuming to collect by hand. It is also highly domain-sensitive, i.e. it varies with predominant word senses, which change across corpora and domains.

In the recent past, several experiments have been conducted on automatic verb classification (Merlo & Stevenson 2001; Schulte im Walde 2006; Joanis *et al.* 2008; Korhonen *et al.* 2008; Li & Brew 2008; Ó Séaghdha & Copestake 2008; Sun *et al.* 2008; Vlachos *et al.* 2009). This work is exciting since it opens up the possibility of inducing novel verb classifications from corpus data, and tuning the existing classifications for specific tasks. Most experiments have focused on English, although some work has also been done on other languages, in particular on German (Schulte im Walde 2006). In what follows, we will mainly survey recent work on general English, but will discuss work on other languages and domains later in §3.

(*a*) *Methods*

(i) *Features*

The first step of lexical classification is to extract from text corpora such linguistic features that may indicate verb classes. English syntactic–semantic verb classification has been traditionally based on *diathesis alternations* (Levin 1993), where syntactic sub-categorization frames (SCFs) alternate, but the meaning of the verb remains the same (or gets modified only slightly). For example, 'break' verbs share a number of alternations, one of which is the causative/inchoative alternation, where two SCFs alternate (*Tony broke the window* ↔ *The window broke*), preserving the basic meaning of the verb *break*. Requiring evaluation of verb meanings, automatic detection of diathesis alternations is very challenging. Therefore, most works on automatic verb classification have used syntactic frames as basic features, exploiting the fact that verbs taking similar alternations take similar SCFs. For example, Joanis *et al.* (2008) have used shallow syntactic slots (e.g. the relative frequency of noun phrases following specific verbs) to approximate the frames. Such slots

can be extracted from corpora using fast, inexpensive NLP processing. Others have used SCFs (Schulte im Walde 2006; Li & Brew 2008; Sun & Korhonen 2009). These correspond better with the frames involved in alternations, but their extraction requires deeper and more costly processing (parsing). Recent research has also experimented with features that may be meaningful, although they have not been used in manual verb classification: co-occurrences (COs) of verbs with other words, e.g. the number of times *break* co-occurs with *Tony, window* and *hammer* within a window of five words; or lexical preferences (LPs), e.g. the number of times *Tony* occurs as a subject of *break* (Li & Brew 2008; Sun & Korhonen 2009). Some experiments have also used verb tense (e.g. the number of times *break* occurs in the past or present tense) and voice (e.g. how often *break* occurs in active and passive) (Joanis *et al.* 2008; Korhonen *et al.* 2008). While most works have focused on syntactic or lexical features, a few attempts have been made to refine syntactic features with semantic information about selectional preferences (SPs), i.e. the semantic preferences that verbs have for their arguments (e.g. the direct object of the verb *break* is often a breakable physical object such as *window*). For example, Joanis (2002) has employed classes in the semantic network of WordNet (Miller 1995) as SP models, and recently, Sun & Korhonen (2009) have experimented with automatically acquired SPs. These were obtained by clustering potential arguments of verbs in parsed data.

(ii) *Classification*

The second step of lexical classification is to classify the linguistic features using machine learning (ML). Both supervised and unsupervised methods have been used for this. Supervised methods assign verbs into a set of predefined classes. They can be useful for NLP tasks, where the set of target classes is known in advance. They tend to perform better than unsupervised methods, but only when hand-labelled training data are available for each target class, which can guide the classification of unseen data. A wide range of supervised methods have been employed so far, including the K nearest neighbours, maximum entropy, support vector machines, Gaussian, distributional kernel methods and Bayesian multinomial regression, among others (Joanis *et al.* 2008; Li & Brew 2008; Ó Séaghdha & Copestake 2008; Sun *et al.* 2008). The majority of these are well-known ML methods that have been successfully applied to related NLP tasks.

Unsupervised methods uncover verb classes in corpus data. They are more exploratory in nature: they can be used to learn novel classifications, e.g. for languages or domains, where no manually built classifications are available, or to supplement the existing classifications (e.g. VerbNet) with novel classes. Unsupervised methods do not require any training data. This is beneficial in tasks where no labelled data are available or would be costly to obtain. Various well-known methods have been tried, e.g. the K means, expectation maximization, spectral clustering, information bottleneck, probabilistic latent semantic analysis and cost-based pairwise clustering (Brew & Schulte im Walde 2002; Schulte im Walde 2006; Korhonen *et al.* 2008; Sun & Korhonen 2009; Vlachos *et al.* 2009). These include both hard and soft clustering methods. The latter assigns a verb into a single class, while the former assigns it to several classes, which can be

useful when the verb has many meanings (e.g. the financial sense versus the motion sense of the verb *charge*). However, soft clustering has not yet proved to be successful in this task (see the discussion in §3).

(*b*) *Evaluation*

(i) *Methods*

Automatic verb classification has been typically applied to large cross-domain corpora and evaluated against a manually constructed gold standard. Two gold standards (GS1 and GS2) based on the verb classes of Levin (1993) have been used to evaluate much of the recent work on English:

GS1 The gold standard of Joanis *et al.* (2008). When frequency-based selection criteria are applied and the class imbalance is restricted, this gold standard provides a classification of 205 verbs in 15 (some broad, some fine-grained) Levin classes.

GS2 The gold standard of Sun *et al.* (2008). This classifies 204 medium- to high-frequency verbs to 17 fine-grained Levin classes, so that each class has 12 member verbs. The verbs have been selected on the basis of their relatively frequent occurrence in corpus data.

A variety of measures have been used to compare automatic classification against a gold standard. Most works have reported at least accuracy (the proportion of labelled classifications that are correct) and *F*-measure. A frequently used measure in NLP, *F*-measure is a weighted harmonic mean of two measures: precision (the proportion of acquired classes that are correct) and recall (the proportion of gold standard classes that were found). Although these measures are calculated slightly differently for supervised and unsupervised approaches (the details of which can be found in respective published papers), we will use them to compare the results of some recent approaches to give a rough idea of the state of the art in this research area. The results should be compared against a random baseline (e.g. 1/number of classes) and a realistic upper bound for the task: e.g. Merlo & Stevenson (2001) have estimated that the accuracy of classification performed by human experts in lexical classification is likely to be around 85 per cent.

(ii) *Performance*

We will now look at the performance of those recently supervised and unsupervised approaches to general English verb classification, which were evaluated on GS1 (using accuracy) and on GS2 (using *F*-measure). On GS1, the best-performing supervised method reported so far is that of Li & Brew (2008). Li and Brew used Bayesian multinomial regression for classification. A range of feature sets integrating COs, SCFs and/or LPs were extracted from a large corpus using a parser. The combination of COs and SCFs gave the best result: an accuracy of 66.3. Joanis *et al.* (2008) report the second best supervised result on GS1 (58.4) using support vector machines for classification. They compared various features extracted using shallow syntactic processing: syntactic slots, slot overlaps, tense, voice and animacy of NPs. They concluded that syntactic

information about core constituents occurring with a verb (syntactic slots) is most important for verb classification. Finally, the unsupervised method of Sun & Korhonen (2009) performs quite similarly to the supervised approach of Joanis *et al.* (2008), yielding an accuracy of 57.6. Sun & Korhonen used a variation of spectral clustering and experimented with a variety of features (e.g. COs, SCFs, LPs, voice, tense), also including semantic ones (SPs). The features were extracted using an SCF acquisition system that makes use of a parser. SPs were obtained by clustering nouns in potential argument positions in parsed data. The best result was obtained when using SCFs in conjunction with SPs.

On GS2, the best-performing supervised method so far is that of Ó Séaghdha & Copestake (2008), which employs a distributional kernel method to classify SCF features parametrized for prepositions in the automatically acquired Valex SCF lexicon (Korhonen *et al.* 2006). It yields 67.3 *F*-measure. Using exactly the same data and feature set, Sun *et al.* (2008) obtained a slightly lower result when using another supervised method (Gaussian): 62.5. The unsupervised approach of Sun & Korhonen (2009) (discussed above with GS1) outperforms both these methods on the same data, when SCFs are used in conjunction with automatically acquired SPs, producing an *F*-measure of 80.4. The better result using an unsupervised method can be attributed to the use of a more accurate parser and an SCF system, and a more comprehensive feature set (see Sun & Korhonen (2009) for details and discussion).

Although this brief comparison focuses on recent work on English classification and does not cover approaches evaluated on other gold standards, languages or domains, it does give a picture of the state of the art: current approaches perform at their very best around 66 accuracy and 80 *F*-measure, when evaluated against relatively small gold standards containing known classes only. While this performance is clearly better than the chance performance, it is still much lower than the realistic upper bound on the task. Also, these figures tell us little about how well the methods would scale-up and perform in the context of NLP application tasks, such as machine translation or information extraction.

3. The way forward

We will now discuss the many challenges that need to be met in order to improve lexical classification further, so that it can be used to benefit real applications. Although we will keep the focus on lexical classification and on how to bridge the gap between research and practice in this area, much of our discussion is relevant to lexical acquisition in general.

(*a*) *Features*

In most experiments, lexical or syntactic features have proved to be most useful for lexical classification. While they may really be the most relevant features for this task, their relatively good performance is partly due to the fact that they can be extracted from corpora quite reliably using the current NLP technology (part-of-speech taggers, parsers and/or SCF acquisition systems). Semantic features, on the other hand, may not perform well simply because they are more challenging to extract from the corpora. For example, although semantic features play a key role in manual verb classification, until recently,

no significant additional improvement was reported using verb SPs (Joanis 2002; Schulte im Walde 2006), although SPs are strong indicators of diathesis alternations (McCarthy 2001) and fairly precise semantic descriptions can be assigned to the majority of verb classes (Kipper-Schuler 2005). However, in their recent experiment, Sun & Korhonen (2009) obtained a considerable improvement using SPs in conjunction with syntactic features, although they used a fully unsupervised approach to both verb clustering and SP acquisition. This may suggest that NLP and ML techniques have now developed to the point where the extraction of at least some semantic features is becoming feasible. Of course, the main semantic features in manual verb classification are diathesis alternations. Although diathesis alternation detection has proved challenging (McCarthy 2001), recent improvements in parsing and lexical (e.g. SCF, SP) acquisition might now also facilitate this research. In general, the integration of different types of lexical acquisition (syntactic and semantic) to support each other could lead to richer and better-quality features.

(*b*) *Classification*

We mentioned a number of supervised and unsupervised ML methods that have been used for verb classification. Many of these methods (e.g. support vector machines) were chosen for the task because they had proved suitable for the classification of natural language data in other NLP tasks and could thus be expected to perform well also in lexical acquisition. However, for optimal results, an ML method should be chosen to match the particular data and task at hand. For example, Sun & Korhonen (2009) obtained promising results in their recent experiment with SP features, not only because the features made theoretical sense but also because the clustering method (spectral clustering) was particularly suited for the resulting, high-dimensional feature space.

Novel clustering and classification methods have been imported from ML recently that have desirable properties thinking of lexical acquisition, for example, methods that combine clustering with an element of guidance based on a prior intuition or methods that do not require defining the number of clusters in advance (e.g. unsupervised and constrained Dirichlet process mixture models for verb clustering by Vlachos *et al.* (2009)). However, semi-supervised or active learning methods have not been explored, although they are well known in NLP (Abney 2008). Combining the benefits of supervised and unsupervised approaches, they could port more easily between domains while maintaining good performance—which is important thinking of real-world applications.

(*c*) *Polysemy*

Many words (e.g. *run*) have several meanings (e.g. *Tom ran to school*, *Tom ran a company*) and can therefore be members of several classes (e.g. 'manner of motion', 'manage'). This phenomenon—*polysemy*—is challenging for NLP. No reliable technique exists for word sense disambiguation yet. Most work on verb classification has by-passed this issue by assuming a single class for each verb—usually the one corresponding to its most frequent sense in a small manually annotated corpus. However, this approach is not realistic thinking of real-world applications. Many important verbs in language do not have a single predominating sense, and for the ones that do, predominating sense is never static

but varies across domains and sub-languages. Few attempts have been made to address this problem. Korhonen *et al.* (2003) performed a clustering experiment with highly polysemous verbs. They constructed a polysemous gold standard for *ca* 200 English verbs and examined whether soft clustering could be used to assign verbs to several classes. The majority of verbs ended up in one class only, but the experiment showed that polysemy has a considerable impact on verb classification. It is clearly an issue that needs to be addressed in lexical acquisition. In verb classification, this amounts to finding a suitable ML method, e.g. one capable of multi-label classification (Boleda *et al.* 2007) or modelling the overlap between lexical categories.

(*d*) *Other languages and domains*

Most lexical acquisition research has focused on English. Verb classification is no exception. Considerable research has been conducted on German (Schulte im Walde 2006), but only small-scale studies exist for other languages, e.g. Spanish (Ferrer 2004), Japanese (Oishi & Matsumoto 1997), Chinese and Italian (Merlo *et al.* 2002). Evaluating the applicability of techniques to several languages would be critical for both theoretical and practical reasons: for (i) improving the accuracy, scalability and robustness of techniques; (ii) advancing work in other languages; (iii) gaining a better understanding of the language-specific versus cross-linguistic components of lexical information (e.g. the extent to which the features used for one language are also valid for other languages); and (iv) improving the performance of NLP applications, including challenging multilingual applications (e.g. machine translation). The same can be said about domains and sub-languages. To our knowledge, the only experiment that has applied verb clustering to a specific domain is that of Korhonen *et al.* (2008). This experiment focused on biomedicine. It revealed that domain-specific classifications can be very different from general classifications. The fact that many domains tend to be quite conventionalized in terms of language use has many consequences, which require further investigation before classification techniques can be successfully applied to domain-specific tasks.

(*e*) *Evaluation and application*

Most lexical acquisition techniques have been evaluated quantitatively on small gold standards, such as GS1 and GS2, although many applications require a large verb classification for optimal performance. For English, the classification of over 5000 verb senses in VerbNet would therefore offer a much more realistic gold standard for evaluation. For many languages and domains, no gold standards are available, but semi-automatic methods exist, which can be used to build them from scratch with adequate linguistic and/or domain expertise (Korhonen *et al.* 2008). Some of the works have supplemented quantitative evaluation with qualitative analysis (e.g. Schulte im Walde 2006; Korhonen *et al.* 2008; Vlachos *et al.* 2009). This is time-consuming, but vitally important, because it can reveal the true potential of automatic acquisition, e.g. the ability of the techniques to discover novel classes and class members (i.e. information missing in current gold standards). It can also help to identify error types, which is important for further development of the techniques. Equally important would be evaluation in the context of practical tasks and applications. To the best of our knowledge,

none of the existing verb classification techniques have been evaluated in this manner yet, although many are capable of producing large-scale classifications and although VerbNet classes have been used to support various applications. Task-based evaluation is lacking in most areas of lexical acquisition. This is surprising, given that the entire line of research is largely motivated by the need to produce better resources for NLP applications. Extrinsic evaluation could provide a more objective measure of performance, especially as the optimal granularity of lexical information (e.g. how fine-grained verb classes should be) may vary from one task to another. Further, since much of the current NLP relies heavily on manually built lexical resources, one way of demonstrating the usefulness of lexical acquisition would be to use the techniques to supplement and tune the existing resources for specific tasks. This is another area where little work has been conducted so far.

4. Conclusion

During the past two decades, a lot has been achieved in automatic lexical acquisition. The initial techniques targeting a very small number of lexical categories have developed into techniques that in some areas of lexical acquisition are capable of extracting large lexical frequency data with promising accuracy. We have discussed the state of the art in this research area, illustrating our discussion with one key area of lexical acquisition—lexical classification— and have described the challenges that need to be met before lexical acquisition can benefit NLP on a large scale. One of the biggest challenges is to improve the accuracy of existing techniques further and to replace small-scale techniques with more powerful and portable techniques. Without this leap, technologies will always be limited in what they can achieve. Meeting this challenge requires extending the current technology further, applying it to larger datasets and novel (sub-)languages, and evaluating the results in richer, novel and more realistic ways. From the perspective of NLP, the ultimate evaluation of lexical acquisition should be based on its impact on practical applications.

References

Abney, S. 2008 *Semisupervised learning for computational linguistics*. London, UK: Chapman & Hall.

Agirre, E. & Edmonds, P. 2006 *Word sense disambiguation: algorithms and applications*. New York, NY: Springer.

Boleda, G., Schulte im Walde, S. & Badia, T. 2007 Modelling polysemy in adjective classes by multi-label classification. In *Proc. Conf. on Empirical Methods in Natural Language Processing, and Conf. on Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007*, pp. 171–180.

Brew, C. & Schulte im Walde, S. 2002 Spectral clustering for German verbs. In *Proc. Conf. on Empirical Methods in Natural Language Processing, Philadelphia, PA, 6–7 July 2002*, pp. 117–124.

Ferrer, E. E. 2004 Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proc. 42nd Annu. Meet. of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004, Workshop on Student Research*, pp. 37–42.

Joanis, E. 2002 Automatic verb classification using a general feature space. Master's thesis, University of Toronto, Canada.

Joanis, E., Stevenson, S. & James, D. 2008 A general feature space for automatic verb classification. *Nat. Lang. Eng.* **14**, 337–367. (doi:10.1017/S135132490600444X)

Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. 2008 A large-scale classification of English verbs. *Lang. Res. Eval.*. **42**, 21–40.

Kipper-Schuler, K. 2005 VerbNet: a broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania, PA.

Korhonen, A., Krymolowski, Y. & Marx, Z. 2003 Clustering polysemic subcategorization frame distributions semantically. In *Proc. 41st Annu. Meet. of the Association for Computational Linguistics, Sapporo, Japan, 7–12 July 2003*, pp. 64–71.

Korhonen, A., Krymolowski, Y. & Briscoe, T. 2006 A large subcategorization lexicon for natural language processing applications. In *Proc. 5th Int. Conf. on Language Resources and Evaluation, Genoa, Italy, 24–26 May 2006*, pp. 345–352.

Korhonen, A., Krymolowski, Y. & Collier, N. 2008 The choice of features for classification of verbs in biomedical texts. In *Proc. 22nd Int. Conf. on Computational Linguistics, Manchester, UK, 18–22 August 2008*, pp. 18–22.

Levin, B. 1993 *English verb classes and alternations: a preliminary investigation*. Chicago, IL: University of Chicago Press.

Li, J. & Brew, C. 2008. Which are the best features for automatic verb classification. In *Proc. 46th Annu. Meet. of the Association for Computational Linguistics, Columbus, OH, 15–20 June 2008*, pp. 434–442.

Marquez, L., Carreras, X., Litkowski, K. & Stevenson, S. 2008 Semantic role labeling: an introduction to the special issue. *Comput. Linguist.* **34**, 145–157. (doi:10.1162/coli.2008.34.2.145)

McCarthy, D. 2001 Lexical acquisition at the syntax–semantics interface: diathesis alternations, subcategorization frames and selectional preferences. PhD thesis, University of Sussex, UK.

McCarthy, D. 2006 Lexical acquisition. In *Encyclopedia of language and linguistics* (ed. K. Brown), pp. 61–68. Oxford, UK: Elsevier.

Merlo, P. & Stevenson, S. 2001 Automatic verb classification based on statistical distributions of argument structure. *Comput. Linguist.* **27**, 373–408. (doi:10.1162/089120101317066122)

Merlo, P., Stevenson, S., Tsang, V. & Allaria, G. 2002 A multilingual paradigm for automatic verb classification. In *Proc. 40th Annu. Meet. of the Association for Computational Linguistics, Philadelphia, PA, 6–12 July 2002*, pp. 207–214.

Miller, G. A. 1995 WordNet: a lexical database for English. *Commun. ACM* **38** (11), 39–41.

Oishi, A. & Matsumoto, Y. 1997 Detecting the organization of semantic subclasses of Japanese verbs. *Int. J. Corpus Linguist.* **2**, 65–89.

Ó Séaghdha, D. & Copestake, A. 2008 Semantic classification with distributional kernels. In *Proc. 22nd Int. Conf. on Computational Linguistics, Manchester, UK, 18–22 August 2008*, pp. 649–656.

Schulte im Walde, S. 2006 Experiments on the automatic induction of German semantic verb classes. *Comput. Linguist.* **32**, 159–194.

Schulte im Walde, S. 2009 The induction of verb frames and verb classes from corpora. In *Corpus linguistics. An international handbook* (eds A. Lüdeling & M. Kytö), pp. 61–68. Berlin, Germany: Mouton de Gruyter.

Sun, L. & Korhonen, A. 2009 Improving verb clustering with automatically acquired selectional preferences. In *Proc. Conf. on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009*, pp. 648–647.

Sun, L., Korhonen, A. & Krymolowski, Y. 2008 Verb class discovery from rich syntactic data. In *Computational linguistics and intelligent text processing*. Lecture Notes in Computer Science, no. 4919, pp. 16–27. Berlin, Germany: Springer. (doi:10.1007/978-3-540-78135-6_2)

Villavicencio, A., Bond, F., Korhonen, A. & McCarthy, D. 2005. Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Comput. Speech Language* **19**, 365–377.

Vlachos, A., Korhonen, A. & Ghahramani, Z. 2009 Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proc. EACL Workshop on Geometrical Models of Natural Language Semantics, Athens, Greece, 31 March 2009*, pp. 74–82.

AUTHOR PROFILE

Anna Korhonen

Anna Korhonen is a Royal Society University Research Fellow in the University of Cambridge where her fellowship is hosted jointly by the Computer Laboratory and the Research Centre for English and Applied Linguistics (RCEAL). She holds an MA in Theoretical Linguistics from the University of Reading (Department of Linguistics, 1995), an MPhil in Computer Speech and Language Processing from the University of Cambridge (Department of Engineering, 1997) and a PhD in Computer Science from the University of Cambridge (Computer Laboratory, 2002). Anna has conducted research on Natural Language Processing at research institutions in the UK (University of Cambridge), USA (University of Pennsylvania) and Japan (National Instutute of Informatics, Tokyo). She is interested in syntactic and semantic analysis of texts, and has made major contributions to automatic acquisition of lexical information from corpora—an important, timely area of NLP which is aimed at developing high-quality lexical resources for real-world NLP applications. She has developed novel techniques and tools for automatic lexical acquisition for English and other languages, and has used them to help key NLP application tasks as well as advance research in related fields (e.g. cognitive and biomedical sciences). She has conducted this research in the context of research fellowships, several EPSRC, MRC, BBSRC-funded projects and international collaborations. She has recently held visiting positions or carried out major joint research with researchers in Colorado, Paris, Stockholm and Tokyo. She has published over 30 articles, chaired several workshops and edited special journal issues, and acted as programme committee member for more than 20 international conferences and associated workshops in her area.