SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D6.1

Technologies and Tools for Lexical Acquisition

Dissemination Level:	Public
Delivery Date:	July 16th 2010
Status – Version:	Final
Author(s) and Affiliation:	Laura Rimell (UCAM), Anna Korhonen (UCAM), Valeria Quochi (ILC-CNR), Núria Bel (UPF), Tommaso Caselli (ILC-CNR), Prokopis Prokopidis (ILSP), Maria Gavrilidou (ILSP), Thierry Poibeau (UCAM), Muntsa Padró (UPF), Eva Revilla (UPF), Monica Monachini(CNR-ILC), Maurizio Tesconi (CNR-IIT), Matteo Abrate (CNR-IIT) and Clara Bacciu (CNR-IIT)

Table of contents

1 Overview

This report describes the technologies and tools to be used for Lexical Acquisition in PANACEA. It includes descriptions of existing technologies and tools which can be built on and improved within PANACEA, as well as of new technologies and tools to be developed and integrated in PANACEA. The report touches briefly on the criteria for evaluating the results of the tools and integration (but see D7.1 for a detailed description of the evaluation criteria). It also specifies the Lexical Resources to be produced.

Four main areas of lexical acquisition are included: Subcategorization frames (SCFs), Selectional Preferences (SPs), Lexical-semantic Classes (LCs), for both nouns and verbs, and Multi-Word Expressions (MWEs). Each partner has chosen to undertake work in only those areas that are feasible given the time allowed and the availability of resources (such as corpora, parsers, and previous lexical acquisition technologies and tools) in the different languages.

The following table gives an overview of the prototypes and resources which will be developed for each language. "Types of lexical info" refers to the investigation of techniques for a given task and language, with the goal of researching whether these techniques can be developed to an accuracy which would improve Machine Translation. These areas of research, which in some cases require a significant amount of tool development in order to proceed, are investigative in nature and are not necessarily expected to result in an integrated component capable of automatically generating a lexicon. "Lexicons" refers to areas of research that are already sufficiently well-developed that the research undertaken as part of the project can be expected to result in an integrated component capable of automatically generating a lexicon. "Lexicon types" refers to whether any lexicons developed will be for general or domain-specific text (using the domain corpora obtained from WP4; the domains currently agreed upon are automotive, work legislation and environment), and "Merger" refers to whether any lexicons developed will participate in the Lexical Merger.

		English	**Spanish**	**Italian**	**Greek**
Types of lexical info	**SCF**	yes	yes	yes	yes
	SP	yes	no	yes	no
	LC verbs nouns	 yes yes	 no yes	 no no	 no no
	MWE	no	no	yes	no
Lexicons	**SCF**	yes	yes	yes	no
	SP	no	no	no	no
	LC	no	no	no	no
	MWE	no	no	no	no
Lexicon types	general domains	no yes	no yes	no yes	no no
Merger		yes	yes	yes	no

Note that unlike some of the other deliverables, the remainder of this document is subdivided by partners rather than languages (e.g. the Survey of Existing Technologies, Tools, and Resources shows the resources available to each partner, and the Work Plan focuses on each partner's work plan). This organization is motivated by the desire to show what resources and expertise each partner has available. It should be understood, however, that the primary partner responsible for English will be UCAM, for Spanish UPF, for Italian ILC-CNR, and for Greek ILSP, except that UPF will work on Lexical Classes for both English and Spanish nouns. Collaborations will be undertaken wherever relevant.

2 Survey of the State of the Art

2.1 Subcategorization Frames

Subcategorization frames (SCFs) define the potential of predicates to choose their argument slots in syntax. Most work on SCF acquisition has focused on verbs, although nouns and adjectives can also subcategorize. A knowledge of SCFs implies the ability to distinguish, given a predicate in raw text and its co-occurring phrases, which of those phrases are arguments (obligatory or optional) and which adjuncts. For example, in the sentence *Mary hit the fence with a stick in the morning*, the NP *the fence* is an obligatory argument, the instrumental PP *with a stick* is an optional argument, and the PP *in the morning* is an adjunct. SCFs describe the *syntactic*, not semantic, behaviour of predicates. Thus Chomsky's well-known example *Colorless green ideas sleep furiously* involves a violation of the selectional preferences of *sleep* but not its SCF, whereas the sentence *The parent slept the child* violates the SCF of *sleep*.

Access to an accurate and comprehensive SCF lexicon is useful for parsing (Briscoe and Carroll, 1997; Collins, 1997; Carroll et al., 1998; Arun and Keller, 2005) as well as other NLP tasks such as Information Extraction (Surdeanu et al., 2003) and Machine Translation (Hajič et al., 2002). SCF induction is also important for other (computational) linguistic tasks such as automatic verb classification, selectional preference acquisition, and psycholinguistic experiments (Schulte im Walde, 2000; Lapata et al., 2001; Schulte im Walde and Brew, 2002; McCarthy, 2001; McCarthy and Carroll, 2003; Sun et al., 2008a, 2008b).

All methods of SCF acquisition share a common objective: given corpus data, to identify (verbal) predicates in this data and record the types of SCFs taken by these predicates, and often their relative frequencies. There are two major steps: hypothesis generation and hypothesis selection. Approaches to hypothesis generation vary, depending on whether raw, partially parsed or intermediately parsed corpus data are used as input to the learning process, and how cues for hypotheses are defined and identified. Hypothesis selection is similarly subject to variation. Some systems treat hypothesised SCFs as absolute SCF indicators, while others treat them as probabilistic indicators. The latter systems typically employ a separate filtering component, with filtering frequently performed using statistical hypothesis tests. Methods vary as to whether the SCFs are pre-specified or learned, how many SCFs are targeted or learned, and how they are defined (e.g. whether they are parametrized for lexically-governed particles and prepositions, whether any semantic knowledge is incorporated, and so forth). See also Schulte im Walde (to appear) for an overview.

The first system for automatic SCF extraction was Brent (1991, 1993), who used lexical cues in raw text to acquire six SCFs from corpus data. Treating the cues as probabilistic indicators, Brent (1993) used the binomial hypothesis test (BHT) to filter hypotheses by deciding when a

verb occurs with a particular SCF often enough that the occurrences are unlikely to be errors. Brent's approach essentially generated high accuracy hypothesis at the expense of coverage, however, since lexical cues are insufficient to identify most SCFs. Subsequent systems (Ushioda et al., 1993; Manning, 1993; Gahl, 1998, Lapata, 1999) therefore aimed to increase coverage by using output from a POS tagger and a chunker, or partial parser, in the hypothesis generation step. A chunker identifies major phrases such as verb groups, NPs, PPs, etc., and thus the observed patterns from the entire corpus could be used. Ushioda (1993) used a set of six SCFs while Manning (1993) used 19, based on the OALD, LDOCE, and COBUILD dictionaries. The chunking-based approaches represented a clear improvement over Brent's approach, since extracting SCF information from chunked data increases the number of cues available and allows for low reliability cues. Their disadvantage, however, is the high level of noise in output, caused by the limitations of partial parsing. To filter the hypotheses, Ushioda used log-linear models of features in the text, while Manning used the BHT, refined by empirically seeting bounds on the probability of cues being false for certain SCFs.

The next generation of systems (Ersan and Charniak, 1996; Carroll and Rooth, 1998; Briscoe and Carroll, 1997; Sarkar and Zeman, 2000) opted for more knowledge-based hypothesis generation with the goal of maximizing both accuracy and coverage. State-of-the-art systems parse the data with an 'intermediate' parser. Rather than simply chunking the input, an intermediate parser finds singly rooted trees. Although such structures are typically built only using POS tag information, they require global coherence from syntax and therefore impose greater grammatical constraints on the analysis. In addition, the intermediate parsers used have been probabilistic, allowing weighting of analyses on the basis of the training data, which also makes them more accurate than the chunkers used in earlier work. Ersan and Charniak (1996) incorporated statistical information for verbs, nouns, and adjectives into a probabilistic context-free grammar (PCFG) parser, followed by an analysis of the PCFG rules to determine which SCFs were generated, and assigning them to a set of 15 SCFs. Carroll and Rooth (1998) used an iterative approach to training a finite state parser using the EM algorithm, where information about SCFs is fed back into each training stage, again using 15 SCFs based on the OALD dictionary.

Large-scale systems targeting a high number of SCFs were proposed by Briscoe and Carroll (1997) and Sarkar and Zeman (2000). Briscoe and Carroll's system is capable of categorizing 163 different SCFs, obtained by merging the SCF classifications of the ANLT and COMLEX dictionaries and manually adding into this set new SCFs discovered from the corpus data. Though previous approaches employed only syntactic SCFs, Briscoe and Carroll's frames also incorporate semantic information (e.g. about control of predicative arguments). The system tags, lemmatises, and parses corpus data using the Robust Accurate Statistical Parser (RASP) (Briscoe and Carroll, 2002) using a feature-based unification grammar formalism. Local syntactic frames are extracted from the parsed data and assigned to SCFs by a classifier. Although unclassifiable patterns are filtered out by the classifier, the output from the hypothesis generator is still noisy, mostly due to parser error. Briscoe and Carroll employ BHT for hypothesis selection, refining it with a priori estimates of the probability of membership in different SCFs.

The SCF extraction method of Sarkar and Zeman (2000) is unique in that it deals with Czech and that it learns previously unknown, i.e. not pre-defined, SCFs. It uses a manually derived dependency treebank (Prague Dependency Treebank, PDT; Hajič, 1998) as input data, where the dependents of a verb constitute the 'observed frame', and the correct SCFs may be subsets of

the observed frames. The hypothesis generator records the frequency of all subsets of each observed frame in treebank data, and considers them from larger (more arguments) to smaller. Large infrequent subsets are suspected to contain adjuncts, so they are replaced by more frequent smaller subsets. Small infrequent subsets may have elided some arguments and are rejected. The resulting frequency data serve as input to hypothesis selection. Sarkar and Zeman use three alternative hypothesis tests: BHT, log likelihood ratio test (LLR, Dunning, 1993) and t-score (Kalbfleisch, 1985), applied iteratively during the search for appropriate SCF subsets. Sarkar and Zeman report that their method learned 137 SCFs from corpus data.

The hypothesis tests used for filtering have been a common problem across many SCF systems. The BHT employed by many systems including Brent (1993), Manning (1993), Lapata (1999), Ersan and Charniak (1996), Briscoe and Carroll (1997), and Sarkar and Zeman (2000) is known to give unreliable performance, especially with low frequency SCFs. One issue is that the distribution of SCFs is not binomial but Zipfian. Korhonen et al. (2000) showed that the BHT and the LLR used with the SCF system of Briscoe and Carroll both perform poorly compared with a simple method which filters SCFs on the basis of their relative frequencies. This is not only because of the Zipfian distribution but because there is very little correlation between the conditional distribution of SCFs given the predicate and the unconditional distribution independent of specific predicates. Accordingly, any method for hypothesis selection (whether or not based on a hypothesis test) that involves reference to the unconditional distribution will perform badly. It is typical now for SCF systems to use relative frequency filtering instead of BHT.

Korhonen (2002) proposed a method of obtaining more accurate, semantically motivated back-off estimates for SCF distributions, and a novel approach to hypothesis selection which makes use of these estimates. Specifically, the back-off estimates were based on Levin verb classes, since verbs show subcategorization preferences similar to others in their class. She chose several representative verbs from each semantic class and merged their conditional (verb form specific) SCF distributions to obtain class-specific back-off estimates.

The state-of-the-art SCF system is Briscoe and Carroll's (1997) system as augmented by Korhonen (2002) and Korhonen and Preiss (2003). This system has been used to atomatically build the VALEX lexicon (Korhonen et al. 2006), containing SCF and frequency information for 6,397 English verbs. This system has been extended to nouns and adjectives by Preiss et al. (2007), who also uses a newer version of the RASP parser.

Some recent work has applied existing techniques to new languages. Ienco et al. (2008) use an Italian corpus annotated with syntactic dependencies, obtaining poor results using T-score to detect frames highly associated with verbs, but better results using a BBN as in Karmanidis et al. (2001).

For Spanish, Chrupala (2003) presents a system to learn subcategorization frames from a 370,000-word corpus by adopting and adapting an existing scheme of classification of subcategorization frames from the SENSEM database project (Fernández et al. 2002) and by implementing a tool that searches partially parsed corpora and detects potential verbal SCFs for 10 Spanish verbs. The detection is based on trying to find matches for “templates”, which are typical syntactic patterns associated with specific SCFs. The evaluation methodology is based on a predefined set of subcat from the SENSEM corpus. Esteve (2004) learns a set of 11 SCFs using a POS tagger, partial parser, SCF classifier, and filter, similar to the system of Korhonen (2002), but making use of information provided by clitic pronouns to assist the lexical builder in identifying verbal arguments given the somewhat flexible constituent order of Spanish. The

system is tested on a 3 million word corpus and a 50 milion word corpus using a manually annotated gold standard for 41 verbs. Pazos et al. (2009) developed a prediction SCFs for verbs based on the SCFs of their hypernyms in Spanish WordNet, but much of the annotation was manual.

Serény et al. (2008) developed a system for Hungarian similar to Brent (1993), but taking into account the rich morphological marking on verbal arguments. The best result was obtained with relative frequency filtering and use of the large Hungarian Webcorpus for training, despite the fact that it is unannotated and had to be automatically POS tagged.

Other work has learned SCFs for nouns and adjectives. Yallop et al. (2005) has developed a system to learn 30 SCFs for adjectives, using the output of RASP and a decision-tree classifier, achieving approximately 68% F-score compared to a manually annotated gold-standard. Preiss et al. (2007) also learn SCFs for adjectives and nouns.

SCF acquisition systems are typically evaluated in terms of 'types' or 'tokens' (e.g. Briscoe and Carroll, 1997; McCarthy, 2001). 'Types' are the set of SCFs acquired, whereas 'tokens' are the individual occurrences of SCFs in corpus data. For type-based evaluation, automatically acquired SCF lexicons are usually evaluated against a gold standard obtained either through manual analysis of corpus data, or from SCF entries in a large dictionary. Manual analysis is usually the more reliable method and it can be used to also evaluate the frequencies of SCFs. Obtaining a gold standard from a dictionary is quick and can be applied to a larger number of verbs, but the gold standard lexicon may be inconsistent with the usage in the corpus, particularly for low-frequency verbs. Token-based evaluation is done against manually analysed corpus tokens, either from the same corpus as the training data or a different one.

The systems that record relative frequencies of different verb and SCF combinations often evaluate the accuracy of the resulting probability distributions as well. This is done by comparing the acquired distribution against a gold standard distribution obtained from manual analysis of corpus data. Various measures of distributional similarity may be used, including the Spearman rank correlation (RC), Kullback-Leibler distance (KL), Jensen-Shannon divergence (JS), cross entropy (CE), skew divergence (SD) and intersection (IS) (Korhonen and Krymolowski, 2002).

Although the SCF acquisition systems described here differ in many ways, including number of SCFs and evaluation corpora, examining the different results can still be useful as it reveals the upper limits of performance of the various state-of-art systems. The most comparable approaches are Manning (1993), Ersan and Charniak (1996) and Carroll and Rooth (1998). They each target a similar number of SCFs and evaluate the resulting lexicons against entries obtained from the OALD dictionary. The best performer among these three is Carroll and Rooth, with 77 type F-score. Brent achieved a type F-score of 85 but on only 6 SCFs, while Briscoe and Carroll (1997) achieved only 55 type F-score but on 163 SCFs. Using Briscoe and Carroll's system but with linear interpolation with semantic back-off estimates, Korhonen (2002) achieved 78.4 F-score on a manually annotated gold standard of 45 test verbs. Korhonen et al. (2006) evaluated several sub-lexicons of VALEX against manually annotated data for 183 test verbs, selected at random from among those taking multiple SCFs. The most accurate lexicon was obtained by including high-frequency SCFs along with lower-frequency SCFs that also occurred in the ANLT or COMLEX dictionaries, semantic class-based smoothing using linear interpolation, and relative frequency filtering, achieving 87.3 F-score. (It is worth noting, however, that the low-frequency entries may not be useful in all domains.) Preiss et al. (2007)

achieved 68.9 F-score for verbs, using essentially the same system as for VALEX but without supplementing with low-frequency ANLT/COMLEX entries. For token-based evaluation, Sarkar and Zeman (2000) report an 88% token recall, but this is the percentage of SCF tokens assigned a correct argument-adjunct analysis, not a correct SCF type analysis; in addition they used manually parsed data which the other systems did not. Ushioda (1993) reported 86% token recall, but on only 6 SCFs, while Manning (1993) reported 82% token recall and Briscoe and Carroll (1997) 81%. We may conclude that, regardless of method, there is a ceiling on SCF acquisition performance for state-of-the-art systems of around 87 F-score and 88% token recall.

In terms of measures of distributional similarity between acquired and gold standard SCF distributions, the results are quite difficult to compare, as each system is evaluated using a different method. However, Preiss et al. (2007) report KL of 1.57 and IS of 0.76 when compared with a gold standard of 183 verbs, while Korhonen et al. (2006) report KL of 0.36 (lower is better) and IS of 0.95 after smoothing. This can be compared with KL of 3.24 and IS of 0.49 for Briscoe and Caroll's (1997) system when run with a new version of the parser in Preiss et al. (2007).

There have also been a small number of extrinsic, task-based evaluations. Lapata and Keller (1998) showed that acquired SCF frequencies make correct predictions about verb completion biases in a psycholinguistic study. Carroll, Minnen and Briscoe (1998) showed that SCF frequencies can significantly improve precision for a lexicalised parser.

A number of challenges remain for SCF acquisition. There is a limit to how far we can get with subcategorization acquisition merely by exploiting syntactic information. As Briscoe and Carroll (1997) point out, the ability to recognize that argument slots of different SCFs for the same predicate share selectional restrictions/preferences would assist recognition that the predicate undergoes specific diathesis alternations. Further, although SCF systems can in principle work with verb lemmas or senses, most existing systems work only with lemmas. However the relative frequency of a SCF varies depending on the relative frequency of the sense and often SCFs are different under sense extensions. For example, in *she smiled herself an upgrade*, the entire scf is only available under the extended sense (Briscoe, 2001). As tagging and parsing have improved in recent years, there may also be more work to be done on improving the initial tagging and parsing stages of the pipeline.

References

A. Arun and F. Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL*. Ann Arbor, Michigan.

Brent, M. 1991. Automatic acquisition of subcategorization frames from untagged text. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA. 209–214.

Brent, M. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics* 19.3: 243–262.

Briscoe, E. J. and Carroll, J. 1997. Automatic extraction of subcategorization from corpora. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC. 356–363.

E.J. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the RASP system. In Proc. of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Carroll, G. and Rooth, M. 1998. Valence induction with a head-lexicalized PCFG. In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing, Granada, Spain. 36–45.

Grzegorz Chrupa³a. 2003. Acquiring Verb Subcategorization from Spanish Corpora. DEA Thesis, University of Barcelona.

Ersan, M. and Charniak, E. 1996. A statistical syntactic disambiguation program and what it learns. In Wermter, S., Riloff, E., and Scheler, G. eds. Connectionist, Statistical and Symbolic Approaches in Learning for Natural Language Processing. Springer-Verlag, Berlin: 146–157.

Gahl, S. 1998. Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada. 428–432.

Eva Esteve Ferrer. 2004. Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proceedings of the ACL Workshop on Student Research.*

J. Hajič, M. Čmejrek, B. Dorr, Y. Ding, J. Eisner, D. Gildea, T. Koo, K. Parton, G. Penn, D. Radev, and O. Rambow. 2002. Natural language generation in the context of machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.

D. Ienco, S. Villata and C. Bosco. 2008. Automatic extraction of subcategorization frames for Italian. In Proc. of LREC. *Marrakech, Morocco.*

Lapata, 1999. [to be filled in]

A. Korhonen. 2002. Subcategorization acquisition. Ph.D. thesis, University of Cambridge Computer Laboratory.

A. Korhonen and Y. Krymolowski. 2002. On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In *Proc. of the Sixth CoNLL*, pages 91-97, Taipei, Taiwan.

M. Lapata, F. Keller, and S. Schulte im Walde. 2001. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30(4):419-435.

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

D. McCarthy and J. Carroll. 2003. Disambituating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics* 29)4_:639-654.

Pazos et. al. 2009. Semi-automatic Generation of Subcategorization Frames for Spanish Verbs Using Ontologies and Verbs Functional Class. Jnl of Computers. 4(8). 721-727.

Judita Preiss and Anna Korhonen. 2002. Improving Subcategorization Acquisition with WSD. In Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. Philadelphia, USA. 102-108.

S. Schulte im Walde. To appear. The induction of verb frames and verb classes from corpora. To appear as chapter 61 in A. Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, Berlin.

S. Schulte im Walde and C. Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of ACL*, Philadelphia, USA.

Sarkar, A. and Zeman, D. 2000. Automatic extraction of subcategorization frames for Czech. In Proceedings of the 19th International Conference on Computational Linguistics, Saarbrücken, Germany. 691–697.

A. Serény, E. Simon, and A. Babarczy. 2008. Automatic acquisition of Hungarian Subcategorization Frames. In *Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI)*.

Lin Sun, Anna Korhonen and Yuval Krymolowski. 2008a, "Automatic Classification of English Verbs Using Rich Syntactic Features" Third International Joint Conference on Natural Language Processing (IJCNLP 2008)

Lin Sun, Anna Korhonen and Yuval Krymolowski. 2008b. "Verb Class Discovery from Rich Syntactic Data", Ninth International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2008) PDF

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL*, Sapporo.

Ushioda, A., Evans, D., Gibson, T., and Waibel, A. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In Boguraev, B. and Pustejovsky, J. eds. SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text. Columbus, Ohio: 95–106.

Jeremy Yallop, Anna Korhonen and Ted Briscoe. 2005. Automatic Acquisition of Adjectival Subcategorization from Corpora. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan.

2.2 Selectional Preferences

Selectional preferences (SPs) describe the *semantic* restrictions imposed by a predicate on its arguments. For example, given that the verb *drink* takes an object, we may also observe that it tends to occur with objects belonging to the class of beverages. Most SP studies have focused on verbs and their nominal direct objects, although some studies also look at nominal subjects, as well as the preferences exhibited by nouns, adjectives, and prepositions (Brockmann and Lapata, 2003, Schulte im Walde, 2010, Zapirain et al., 2010, Ó Séaghdha, 2010). The task of learning selectional preferences is similar to the task of judging the plausibility of a predicate and argument occurring together. Knowledge of SPs is helpful for such NLP tasks as resolving ambiguous syntactic attachments (Hindle and Rooth, 1993), word sense disambiguation (McCarthy and Carroll, 2003, Wagner et al., 2009), semantic role labelling (Gildea and Jurafsky, 2002, Zapirain et al., 2009, 2010), natural language inference (Zanzotto et al. 2006, Pantel et al, 2007), detecting multi-word expressions (McCarthy et al., 2007), and paraphrasing metaphoric language (Shutova, 2010). The basic challenge in SP acquisition is to be able to generalize from observed predicate-argument pairs to classes of arguments, despite the sparsity of evidence for the class. For example, it may be useful to know whether *lemonade* is a plausible object for the verb *drink* even if this pair never occurred together in the training data.

Most studies have focused on English, though Brockman and Lapata (2003) and Schulte im Walde (2010) investigated German. Peirsman and Padó (2010) investigated bilingual induction

of SPs for German and Spanish using a bilingual semantic space, translating into English to obtain the plausibility judgements.

A key component of any SP acquisition system is the set of classes used for generalization. There have been two major approaches. The first is to use an existing taxonomy of semantic categories, most commonly WordNet (Miller 1995), and the other is to learn the classes automatically from the data. The latter approach is the only viable one for languages or domains where existing taxonomies are not available.

For the WordNet-based approaches, the task is to find the WordNet concept(s) that most accurately describe the selectional preferences for a given predicate, along with a statistical model of how well a predicate fits its arguments. Identifying an appropriate concept means finding one at the right level of generality in the semantic hierarchy– e.g. *beverage* rather than *liquid* or *substance* for the object of *drink*. The basic approach, introduced by Resnik (1993, 1997), is first to extract argument headwords for a given predicate and relation from a corpus, and then to generalize to other, similar words using WordNet. A number of methods have been used for generalizing. Resnik (1993, 1997) defined association strength, an information-theoretic measure of the semantic fit of a class to a predicate, based on the relative entropy of the distribution of classes with and without regard to predicate. Li and Abe (1998) used the Minimum Description Length (MDL) principle, which seeks to minimize the combined cost of encoding the model and the data, to find an appropriate cut in the WordNet hierarchy. Clark and Weir (2002) used hypothesis testing to find the appropriate level of generality for suitable generalization classes. Abney and Light (1999) used a Hidden Markov Model (HMM) to find the most likely path through the WordNet hierarchy from the root to a word sense. Ciaramita and Johnson (2000) used Bayesian belief networks to quantify SPs. See Light and Greiff (2002) and Brockmann and Lapata (2003) for overviews and comparisons of these approaches.

WordNet-based approaches are limited by the fact that the resource is of limited size, and the classes it includes are pre-defined. Even working on English the fact that the classes are static means that they may not be appropriate for a given domain or task. Thus more recent work has focused on automatically acquiring the generalization classes based on a corpus, e.g. by clustering or similarity-based methods. Four main types of approach have recently been used.

The first type of approach uses generative probabilistic models, in which each observed predicate-argument pair is assumed to be generated by a latent class variable. For Pereira et al. (1993) and Rooth et al. (1999), each class corresponds to a multinomial distribution over relations and arguments; Rooth et al. learn the model by Expectation Maximisation. Schulte im Walde et al. (2008) uses the model of Rooth et al. and incorporates the MDL principle into the EM training, so that the model explicitly models WordNet classes, but this does not provide much advantage. Padó et al (2006) use a generative probability model which jointly models the plausibility of a verb and its argument with the thematic role and grammatical function of the argument, and the verb sense. Information about thematic roles and grammatical functions is obtained from FrameNet (Baker et al. 2003), but the sparsity of FrameNet leads to low coverage and the necessity for significant smoothing. Ó Séaghdha (2010) and Ritter et al. (2010) use Latent Dirichlet Allocation (LDA), which has been proven effective for document topic modelling; in this approach the classes correspond to topics. Ó Séaghdha shows that LDA is especially effective for infrequent predicate-argument combinations, distinguishing between rare, yet plausible, combinations and ones that are genuinely implausible.

The second type of approach uses similarity-smoothed models to find the probability of a particular argument occurring with a predicate. Unlike the generative models, the similarity-based methods do not find clusters of arguments, but simply provide a model for probabilistic plausibility judgements on predicate-argument pairs. Erk (2007) and Padó (2007) treat the probability of an argument occurring with a given predicate as a weighted sum of its probability of occurring with other similar predicates, using a vector space of co-occurrences as a "semantic space". Erk (2007) uses semantic role labelling, while Padó et al. (2007) uses shallow parsing to find semantic relations. Padó et al. had success combining the low-coverage, generative Padó et al. (2006) model with the similarity-based model as a backoff. Schulte im Walde (2010) uses a second-order distributional model, which models salient properties of the argument. For example, it looks at adjectival modifiers, verb and prepositional phrase co-occurrence, e.g. for the verb *bake*, direct objects might also tend to co-occur with the adjectives *fresh*, *delicious*, etc.

The third type of approach is discriminative learning. Bergsma et al. (2008) train a collection of Support Vector Machine (SVM) classifiers to distinguish positive examples from pseudo-negative examples (created from predicate-argument combinations unobserved in a corpus). The discriminative framework makes it possible to use a large number of features in the model and Bergsma et al. use over 57,000 features including verb co-ocurrence, semantic class (from generated clusters), and fine-grained string- and token-based features such as upper vs. lowercase, number of tokens per argument, and the presence of digits, hyphens, and proper names. Discriminative training can also learn regularities across predicates, e.g. the object of *eat* is also likely to be an object of *buy* and *cook*. Like the similarity-based approaches, the discriminative approach does not produce explicit classes of arguments.

Finally, the fourth type of approach is to use simple co-occurrence frequencies to model plausibility of predicate-argument relations, but over a very large corpus. Keller and Lapata (2003) use web searches with patterns such as "v Det n" to decide whether n is a likely object for v. Chambers and Jurafsky (2010) show that a baseline using simple co-occurrence counts, backing off to a random choice when no decision can be made, and trained on approximately 1.2 billion tokens from the entire New York Times portion of the Gigaword corpus, outperforms some other state-of-the-art models. They also show that this approach can be combined with other models.

There are three main ways of evaluating SP systems: against human plausibility judgements, on a pseudo-disambiguation task, and with task-based evaluations. It is somewhat difficult to compare state-of-the-art systems since different corpora and tasks are used. However, some comparisons are possible.

The most widely used human plausibility judgement dataset is that of Resnik (1996), who used a set of 16 verbs from Holmes et al. (1989), each paired with a plausible and an implausible noun argument as judged by human subjects. Systems are judged on how well their scores agree with the human judgements. The highest scorers on the dataset from Resnik (1996) are Bergsma et al. (2008) and Ó Séaghdha (2010), both of which perform perfectly on this data (though Ó Séaghdha 2010 is trained with a smaller corpus).

The broadest pseudo-disambiguation comparison has been performed by Bergsma et al. (2008), using verb-noun pairs from the AQUAINT corpus (Voorhees, 2002) where the noun occurs at least three times in the corpus. The highest performer was the system of Bergsma et al., with a macroaverage F-score of 0.65 (averaged across each example) and 0.83 microaverage (weighted by word frequency) on positive-negative plausibility judgements, and an accuracy of 0.81 on

traditional pseudo-disambiguation, where each of the positive examples was randomly paired with a negative and the system asked to discriminate among them. Previously reported pseudo-disambiguation experiments are not comparable since the data sets are chosen with different parameters, e.g. minimum noun or verb frequency of anywhere from 30 to 500.

References

Steven Abney and Marc Light (1999): "Hiding a semantic class hierarchy in a Markov Model". In: Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing.

C. Baker, C. Fillmore, and B. Cronin. 2003. The structure of the Framenet database. *International Journal of Lexicography* 16(3):281-269.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of EMNLP*.

Carsten Brockmann and Mirella Lapata (2003): "Evaluating and combining approaches to selectional preference acquisition". In: Proceedings of the 10th Conference of the EACL.

Nathanael Chambers and Dan Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In Proceedings of ACL.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional restrictions with Bayesian networks. In *Proceedings of the 18th Inernational Conference on Computational Linguistics*, pp. 187-193, Saarbrücken, Germany.

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, simiarity-based model for selectional preferences". In: Proceedings of the 45th Conference of the Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245–288.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. Computational Linguistics, 19(1):103–120.

V.M. Holmes, L. Stowe, and L. Cupples. 1989. Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language* 28:668-689.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3):459-484.

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

Marc Light and Warren Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 87:1–13.

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

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In Proceedings of EMNLP.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. Proceedings of ACL. Uppsala, Sweden.

Sebastian Padó, Ulrike Padó and Katrin Erk. Flexible, corpus-based modelling of human plausibility judgements. Proceedings of EMNLP-CoNLL 2007. Prague, Czech Republic, 2007

Ulrike Padó, Frank Keller, and Matthew W. Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of the 28th CogSci*, pages 657–662, Vancouver, BC.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In Proceedings of NAACL 2007, Rochester, NY.

Y. Peirsman and S. Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In Proceedings of NAACL.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of ACL*. Columbus, OH.

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

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127-159.

Philip Resnik. 1997. "Selectional preference and sense disambiguation". In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?

Alan Ritter, Mausam, and Oren Etzioni. 2010. A Latent Dirichlet Allocation method for selectional preferences. In *Proceedings of ACL.*

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annoated lexicon via EM-based clustering. In *Proceedings of ACL*. College Park, MA.

Sabine Schulte im Walde. 2010. Comparing computational models of selectional preferences – second-order co-occurrence vs. latent semantic clusters. In Proceedings of LREC.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In Proceedings of ACL.

E. Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL*, Los Angeles, USA.

J. Trueswell, M. Tanenhaus, and S. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language 33:285-318.*

Ellen Voorhees. 2002. Overview of the TREC 2002 question answering track. In *Proceedings of TREC.*

Wiebke Wagner, Helmut Schmid, Sabine Schulte im Walde. 2009. Verb Sense Disambiguation using a Predicate-Argument-Clustering Model. Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts. Amsterdam, The Netherlands.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In Proceedings of COLING.

B. Zapirain, E. Agirre, and L. Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In Proceedings of ACL-IJCNLP.

Beñat Zapirain, Eniko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2010. Improving semantic role classification with selectional preferences. In Proceedings of NAACL.

2.3 Lexical-semantic Classes

2.3.1 Lexical-semantic classes for verbs

Lexical classes are defined in terms of shared meaning components and similar syntactic behavior of words (Levin, 1993). 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 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 to predict 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 which employs parsing (e.g. information extraction, machine translation).

Lexical classifications have been used to support many important NLP tasks, including e.g. computational lexicography, parsing, word sense disambiguation, semantic role labeling, information extraction, question-answering, and machine translation (Kipper et al., 2008), among others. 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 summaries decades of theoretical research on English verb classification. It classifies over 5000 verbs into 274 first level classes on the basis of their syntactic-semantic properties. Manual extension and tuning of VerbNet to different domains has 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 languages, corpora and domains.

In the recent past, several experiments have been conducted on automatic verb classification (Merlo and Stevenson, 2001; Schulte im Walde, 2006; Joanis et al., 2008; Sun et al., 2008; Li and Brew, 2008; Korhonen et al., 2008; Ó Séaghdha and Copestake, 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 existing classifications for specific tasks. Most experiments have focussed on English, although some work has also been done on other languages, in particular on German (Schulte im Walde, 2006).

The first step of lexical classification is to extract from text corpora linguistic features which may indicate verb classes. English syntactic-semantic verb classification has been traditionally based on diathesis alternations (Levin, 1993) where syntactic subcategorization frames (SCFs) alternate, but the verb meaning stays 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 SCF s (Schulte im Walde, 2006; Li and Brew, 2008; Sun and 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 which 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 and Brew, 2008; Sun and 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 focussed 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 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) have employed classes in the semantic network of WordNet (Miller, 1995) as SP models, and recently, Sun and Korhonen (2009) have experimented with automatically acquired SPs. These were obtained by clustering potential arguments of verbs in parsed data.

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 pre-defined 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; Sun et al., 2008; Li and Brew, 2008; Ó Séaghdha and Copestake, 2008). The majority of these are well-known ML methods which 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 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 is 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 and Schulte im Walde, 2002; Schulte im Walde, 2006; Korhonen et al., 2008; Sun and Korhonen, 2009; Vlachos et al., 2009). These include both hard and soft clustering methods. The former assign a verb into a single class while the latter assign it to several classes which can be useful when the verb has many meanings (e.g. the financial sense vs. the motion sense of the verb charge). However, soft clustering has not proved successful in this task yet.

Automatic verb classification has been typically applied to large cross-domain corpora and evaluated against a manually constructed gold standard. Two gold standards based on Levin (1993)'s verb classes have been used to evaluate much of the recent work on English: Joanis et al. (2008) provides a classification of 205 verbs in 15 (some broad, some fine-grained) Levin classes, and Sun et al. (2008) classifies 204 medium-high frequency verbs into 17 fine-grained Levin classes, so that each class has 12 member verbs. In both cases the verbs have been selected based on their frequencies in corpus data. Most works report accuracy and F-measure on the gold-standard data. 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: for example, Merlo and Stevenson (2001) have estimated that the accuracy of classification performed by human experts in lexical classification is likely to be around 85%.

On the gold standard of Joanis (2008), the best performing supervised method reported so far is that of Li and 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: 66.3 accuracy. Joanis et al. (2008) report the second best supervised result (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 to verb classification. Finally, the recent unsupervised method of Sun and Korhonen (2009) performs quite similarly with the supervised approach of Joanis et al. (2008), yielding 57.6 accuracy. Sun and Korhonen used a variation of spectral clustering and experimented with a variety of fea-tures (e.g. COs, SCFs, LPs, voice, tense), including also semantic ones (SPs). The features were extracted using a SCF acquisition system which makes use of a parser. The 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 the gold standard of Sun et al (2008), the best performing supervised method so far is that of Ó Séaghdha and Copestake (2008) which employs a distributional kernel method to classify SCF features parameterized 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 recent unsupervised approach of Sun and Korhonen (2009) outperforms both these methods on the same data when SCFs are used in conjunction with automatically acquired SPs, producing 80.4 F-measure. The better result using an unsupervised method can be attributed to the use of a more accurate parser and a SCF system, and a more comprehensive feature set (see (Sun and 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 perfomance, 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 applications such as machine translation or information extraction.

References

Brew, C. and S. Schulte im Walde, 2002. Spectral clustering for german verbs. In Proc. Of EMNLP.

Joanis, E., 2002. Automatic Verb Classification Using a General Feature Space. Master's thesis, University of Toronto.

Joanis, E., S. Stevenson, and D. James, 2008. A general feature space for automatic verb classification. Natural Language Engineering.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer, 2008. A large-scale classification of english verbs. Language Resources and Evaluation.

Kipper-Schuler, K., 2005. VerbNet: A broad-coverage, comprehensive verb lexicon.

Korhonen, A., Y. Krymolowski, and T. Briscoe, 2006. A large subcategorization lexicon for natural language processing applications. In Proc. of the 5th LREC.

Korhonen, A., Y. Krymolowski, and N. Collier, 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. In Proc. of COLING.

Levin, B., 1993. English verb classes and alternations: A preliminary investigation. Chicago, IL.

Li, J. and C. Brew, 2008. Which Are the Best Features for Automatic Verb Classification. In Proc. of ACL.

Merlo, P. and S. Stevenson, 2001. Automatic verb classification based on statistical distributions of argument structure. Computational Linguistics, 27:373–408.

Miller, G. A., 1995. WordNet: a lexical database for English. Communications of the ACM.

Ó Séaghdha, D. and A. Copestake, 2008. Semantic classification with distributional kernels. In *Proc. of COLING*.

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

Sun, L. and A. Korhonen, 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proc. of EMNLP*.

Sun, L., A. Korhonen, and Y. Krymolowski, 2008. Verb class discovery from rich syntactic data. Lecture Notes in Computer Science, 4919:16.

Vlachos, A., A. Korhonen, and Z. Ghahramani, 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. In Proc. of the Workshop on Geometrical Models of Natural Language Semantics.

2.3.2 Lexical-semantic classes for nouns

In contrast with verbs, the topic of proposing classes of nouns has not been addressed in the works that dealt with noun semantics. Traditionally noun lexical-semantic meaning has been addressed in frameworks more related to knowledge representation such as taxonomies and ontologies. WordNet (Fellbaum, 1998) or Generative Lexicon, (Pustejovsky, 1995) are exceptions but they still make use of different theoretical constructs (synsets in Wordnet, complex types in GL, etc.) whose final goal is not to define groups of related syntactic and lexical properties in the way we are considering here. Our work during the first year will concentrate on the definition and learning of such classes, and the proposal of classes for Spanish and English nouns.

The acquisition of lexical information for nouns has also been less addressed than for verbs. For instance, Light (1996) used information from derivational affixes to classify nouns. Baldwin and Bond (2003) induced mass/count information from a parsed English corpus, using parallel supervised classifiers that took into account different syntactic cues: head number, modifier number, subject-verb agreement, the occurrence in 'N of N' constructions, etc. Bel et al. (2007) used Decision Trees with morphosyntactic and lexical cues for training a classifier to identify mass nouns in Spanish (as well as their pattern of complementation, including bounded prepositional phrases). With some technical differences, Bel et al. (2010) also used the frequency of ad-hoc, linguistically motivated, morphosyntactic and lexical cues for building a classifier for identifying a subclass of event nouns in Spanish and English.

Some work on lexical semantics that is worth mentioning although some how different of classification of nouns into classes has been carried out in the area called "Word Space Models", see for instance Baroni et al. (2008). Authors in this area share the assumption that the statistical analysis of the contexts in which words co-occur gives a representation of the semantic content of words. These works, however, require of very large amount of data for computing lexical co-occurrences.

As for evaluation, the systems we have mentioned have been assessing accuracy of type classification. Accuracy of Baldwin and Bond (2003) system was measured in terms of F-score[1]: 0.89 in classifying English nouns as mass, with a gold standard test set that, however, accepted a double classification, i.e. a noun could be both mass and count. Bel et al evaluated their results and declared an accuracy of 67%, although allowing only one class per noun in the gold standard. Following this approach, the most recent experiments with event nouns for

[1] F-score is the harmonized mean value of precision and recall.

English and Spanish obtained an accuracy of 80% for Spanish and 79% in English (the experiment for English used a small corpus).

References

Baldwin, T. and F. Bond. 2003. "Learning the Countability of English Nouns from Corpus Data". Proceedings of the 41st. Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan.

Baroni, M.; Evert, S. and Lenci, A. (eds.) 2008. ESSLLI Workshop on Distributional Lexical Semantics.

Bel, Núria; Coll, Maria; Resnik, Gabriela (2010), Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production, to appear in Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010, Beijing, China.

Resnik, Gabriela; Bel, Núria (2009). "Automatic Detection of Non-deverbal Event Nouns in Spanish" in -- Proceedings of the 5th International Conference on Generative Approaches to the Lexicon. Pisa: Istituto di Linguistica Computazionale.

Bel, Núria; Espeja, Sergio; Marimon, Montserrat. "Automatic Acquisition of Grammatical Types for Nouns" dins *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York: Association for Computational Linguistics. 2007. Pàg. 5-8. ISBN 1-932432-94-

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

2.4 Multi-Word Expressions

MWEs are often defined as "idiosyncratic interpretations that cross word boundaries (or spaces)", i.e., adjacent or non adjacent combinations of words (e.g., *by and large, have a bath, high school*) that refer to a single concept (Sag et al., 2002; Mona and Bhutada, 2009; among others). Many types of MWEs have been identified so far in the literature, depending on the parameters that are taken into consideration (some works addressing the classification problem are Baldwin et al., 2003; Katz and Giesbrecht, 2006; Mona and Bhutada 2009). The parameters that are mostly used for classification are the degree of (both semantic and syntactic) idiomaticity and the frequency of the expression. Sag et al. (2002) distinguish between lexicalized phrases, that show some degree of semantic and syntactic idiosyncrasy, and institutionalized phrases, that are compositional but highly frequent. Lexicalized phrases showing the highest idiomaticity and syntactic fixedness are *fixed expressions* (e.g., *ad hoc*) and *semi-fixed expressions* (e.g., *speak of the devil*); other types of lexicalized phrases, showing some degree of compositionality and flexibility, are *light* (or *support*) *verb constructions* (e.g., *take a shower*) and *verb-particle constructions* (e.g. *get over*). On the other hand, institutionalized phrases, or collocations, are compositional, but tend to occur together with a statistically idiosyncratic frequency (e.g. *traffic light*, Sag et al. 2001).

In the literature on NLP the notion of MWE is often overlapping with the one of collocation. From this perspective, collocations can be seen as the super-set of MWEs: collocations are seen as lexical affinities identified by calculating strong word associations in corpora using various association measures (AMs), they are not necessarily lexicalized phrases and therefore may be

compositional; MWEs instead are lexicalized phrases, showing some degree of semantic opacity are generally seen as semantic units, i.e. denoting a specific concept or entity in the real world.

It is often stressed that the identification of MWEs and collocations is extremely relevant for many NLP tasks, especially for those related with some kind of semantic processing, such as information retrieval and machine translation. The increasing interest for MWEs is reflected in many dedicated events, such as the "Workshops on Multiword Expressions" (organized at ACL, LREC and COLING from 2003 to 2010), and research projects, such as the Stanford Multiword Expression Project and the Identification (http://mwe.stanford.edu/) and Representation of Multiword Expressions (IRME) project (for Dutch, http://www-uilots.let.uu.nl/irme/). The reason for this interest relies mainly on the fact that MWEs and collocation are extremely frequent in language. Jackendoff (1997) claims that in the general lexicon of speakers the number of MWEs and that of single words are comparable. When it comes to text from specific domains, the number of MWEs is even larger (Nakagawa and Mori 2003 show that 85% of the entries in specialized lexicons are MW terms).

Given the importance of MWEs for NLP applications, much research has been conducted for their automatic acquisition, with the aim of building or expanding lexica, both general and domain-specific.

Extraction procedures usually involve the following two steps: (1) the identification of candidates, and (2) the candidates ranking according to the collocational strength or association score. To that end, different methods have been proposed in the existing literature. Older approaches make use of plain text corpora and identify candidate on the basis of *n*-grams; some of them then use POS filtering to clean the candidate lists. More recent methods make use of parsed data in order to improve precision (see the nice review in Seretan and Wehrli 2009: 73-74). Through a POS tagger, for example, it is possible to first identify all words tagged as particle, then to identify the head verb associated to those words: this way verb-particle constructions are identified (cf. Baldwin 2005). The problem, in this approach, is that it is not possible to distinguish true MWE and word combinations with literal meaning (Baldwin and Kim 2010). Another approach refers to the "fixedness" of many (although not all) MWEs. True MWE, as opposed to combinations with literal meaning, are assumed not to undergo morphologic or syntactic variation. For example, if the system finds "kicking the buckets", it will not consider this combination to be a MW. The problem, in this case, consists in the large amount of manual work to determine the degree of variability a given MWE can undergo (Baldwin and Kim 2010).

The ranking of candidates is then achieved by applying some association measure (hereafter AM) calculated on the basis of co-occurrence frequency of the content words involved in candidates. AMs are formulas used to determine the degree of association between constituents of phrases: MWE candidates are those groups of words co-occurring with a frequency that is significantly higher as compared to that of the individual words forming them. To each extracted collocation candidate is attributed an association score, either for ranking (candidates with the higher probability to be a collocation at the top) or for classification (candidates below a given threshold are discarded) (cf. Pecina 2010). Some of the most commonly used AMs are: Mutual Information (MI), Pointwise MI, Dice, Pearson's chi-squared, log-likelihood ratio, odds ratio, Fisher's exact tests, left and right context entropy, Permutation Entropy. Several works have also carried out detailed comparisons of the methods used in the literature, evaluating the association measures used. Among them, Pearce (2002), Evert (2004), Hoang *et al.* (2009), and

Pecina (2010). In Pecina (2010), 82 AMs are evaluated, using data sets of collocation candidates extracted from the *Prague Dependency Treebank* and from the *Czech National Corpus*. Different AMs for MWE extraction are compared and their performance evaluated by precision-recall curves and by mean average precision scores.

It emerges that the efficacy of a given AM cannot be stated in absolute terms: it depends on factors like the language being analysed and the type of MWE that has to be identified (Evert and Krenn 2005). Moreover, different AMs may be used to isolate different properties of the association between words. In general, it is claimed that the better choice is to combine different AMs together, since in this way both precision and recall of the extraction procedure are improved. Hoang et al. (2009) evaluate AMs for extracting verb-particle and light-verb constructions using a data set from the Wall Street Journal section of the Penn Tree Bank (the method followed to build the data set was: a) for verb-particle construction, first particles were identified, then the head verbs; b) for light verb constructions, first occurrences of light verbs were identified, then the nearest noun on the right of the verb. As a result of evaluation, the authors divide AMs into two main classes: one class of AMs (including MI, Pointwise MI, T score, Pearson's chi-squared, and others) is suitable for detecting the degree of institutionalization; the other class of AMs (including cosine, dice similarity, and others) use context information to measure non-compositionality. Other authors demonstrate that the success of a single AM depends on the specific type of MWE to be identified. For example Krenn and Evert (2001), looking at precision and recall scores, show that support verb constructions in German are best extracted through Mutual Information, while for figurative expressions mere co-occurrence frequency is more suitable.

Another approach that has been experimented for MWE extraction is the alignment based method (Melamed, 1997; Caseli et al., 2009; Zarrieß and Kuhn Caseli et al., 2010). In this case, two parallel texts (one in the source and the other in the target language) are automatically aligned. Candidate MWEs are those sequences of two or more words in the source language that are aligned with one or more words in the target language. For example, the English sequence *human being* may be found aligned with both *essere umano* and *persona* in an Italian translation. Caseli et al. (2009) show that this method is characterized by low costs as concerns the tools and resources required, because collocation candidates come as a by-product of automatic word alignment.

For most extraction methods, after the generation of a first list of MWE candidates, the next step is to filter them. This process can be done automatically (for example, by deciding a minimal threshold of occurrences to remove infrequent candidates, cf. Caseli et al., 2009) and/or manually (cf. Pecina, 2010).

As already noticed, MWEs are numerous in general language, but in specialized domain language they are even more frequent (Sag et al., 2002). Therefore, much research has been done to extract MWEs from text from specific domains, such as pediatrics (Caseli et al. ,2009), history of art and legal texts (Bonin et al., 2010). The connection with terminology extraction is clearly tight. Ramisch et al. (2010), for example, propose a Multiword Expression Toolkit for the identification of MWEs, and apply it to domain-specific text corpora. In particular, they worked on the biomedical domain. The extraction of candidates is based either on row n-grams or on morphosyntactic (POS) patterns (that may contain wildcards, so that it is possible to extract also discontinuous MW terms). The list is then filtered using a set of four different AMs. Using a frequency threshold of 5 (i.e., considering only candidate that occur at least 5 times in the corpus), precision is 74.14%, recall is 6.42%, F-measure is 11.82%. If a higher recall is

needed (for instance, if the aim is the creation of a terminological dictionary), the threshold can be lowered to 1, thus obtaining a recall of 20.91%. The authors observe that their domain-specific MWE extraction methods achieve higher results than the baseline systems used for comparison (the general-purpose tool Xtract and Yahoo! terms).

Although much work is on English data, research on MWE extraction has been carried out also for many other languages, such as German (Krenn and Evert, 2001; Zinsmeister and Heid, 2003), Dutch (Villada Moiron, 2005; Grégoire, 2010), Czech (Pecina et al., 2009), French (Laporte et al., 2008), Portuguese (Villavicencio et al., 2010), among others.

For Italian, a first work on collocation extraction used a window method for identifying candidate in a plain text corpus and use MI for ranking (Calzolari and Bindi, 1990). Recently, efforts have been made to create MWE resources. Bentivogli and Pianta (2002) extracted from the Collins English-Italian dictionary MWE ("hidden" MWEs, i.e. MWEs that are not explicitly marked as such in the dictionary) in a semi-automatic way, thus compiling a list of 18,800 Italian MWEs. Also Zaniello and Nissim (2010) extracted MWEs from an existing dictionary (the monolingual *De Mauro-Paravia* online dictionary), creating a lexicon encoded in XML. Each MWE contained in the lexicon was then used as a query, to extract an example corpus from the large web-based corpus ItWac. It was also created a relational database of MWE, encoding morphosyntactic patterns. Spina (2010) reports on the creation of a *Dictionary of Italian Collocation* (DICI) to be integrated in a Virtual Learning Environment for learners of Italian as a second language. A list of collocations were first extracted from LIP (a spoken corpus) and from ItalWordNet. From this list, the 10 most frequent POS patterns were selected. These patterns were then used to extract collocation candidates from the Perugia Corpus. After a filtering process, a list of 1553 collocations has been selected to be included in the dictionary. Bonin *et al.* (2010) extracted MW terminology for the Art History and Legal domains adopting a contrastive approach in order to identify domain-specific multi-words and filtering out open-domain ones. The resulting list has been evaluated against gold standard resources (domain-specific dictionaries) and through validation by domain experts.

References

Baldwin, T. 2005. The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions* 19 (4), 398–414.

Baldwin, T., Kim, S. N. 2010, Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau (eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267—292.

Baldwin, T., Bannard, C., Tanaka, T. Widdows, D. 2003. An empirical model of multiword expression decomposability. In *Proceedings of Workshop on Multiword expressions, ACL 2003*, pages 89–96, Morristown, NJ, USA.

Bentivogli, L., E. Pianta. 2002. Detecting Hidden Multiwords in Bilingual Dictionaries. *Proceedings of the tenth EURALEX International Congress*, Copenhagen, Denmark, August 14-17, 2002. 785-793.

Bonin, F., Dell'Orletta, F., Venturi, G., Montemagni, S. 2010. A Contrastive Approach to Multi–word Term Extraction from Domain Corpora. *Proceedings of LREC 2010*, Valletta, Malta.

Bouma, G., B. Villada. 2002. Corpus-based Acquisition of Collocational Prepositional Phrases. Computational Linguistics in the Netherlands (CLIN) 2001. University of Twente.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A. 2002. Towards best practice for multiword expressions in computational lexicons. *Proceedings of LREC 2002,* Las Palmas, Canary Islands.

Calzolari, N., Bindi, R. 1990. Acquisition of Lexical Information from a Large Textual Italian Corpus. In *Proceedings of the 13th International Conference on Computational Linguistics.* Helsinky, Finland. pp. 54-59.

Caseli, H. de M. et al. 2009 . Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains. In *Proceedings of ACL 2009* pp.1-8.

Caseli, H. de M. et al. 2010. Alignment-based Extraction of Multiword Expressions. *Language Resources & Evaluation* 44: 59–77.

Evert, S. 2004. The statistics of word cooccurrences: Word pairs and collocations. PhD Thesis, University of Stuttgart.

Evert, S. 2008. The MWE 2008 Shared Task: Ranking MWE Candidates. *LREC 2008,* Marrakech, Morocco.

Evert, S., & Krenn, B. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language* 19 4 , 450–466.

Grégoire, N. 2010 . DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44: 23–39.

Hoang, H. H. et al. 2009. A Re-examination of Lexical Association Measures. P*roceedings of ACL 2009.*

Jackendoff, R.,1997. Twistin' the night away. *Language,* 73: 534–59.

Katz, G., Giesbrecht , E. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July. Association for Computational Linguistics.

Krenn, B., S. Evert. 2001. Can we do better than frequency? A case study on extracting pp-verb collocations. In *Proceedings of the ACL Workshop on Collocations*. Toulouse, France.

Laporte, E., Nakamura, T., Voyatzi, S. 2008. A French Corpus Annotated for Multiword Nouns. *Proceedings of LREC 2008,* Marrakech, Morocco.

Melamed, I. D., 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing EMNLP-2* , Brown University, USA, August 1997. Association for Computational Linguistics.

Mona T. Diab, Pravin Bhutada, "Verb Noun Construction MWE Token Supervised Classification" *In Proceedings of ACL 2009*

Nakagawa, H., T. Mori 2003. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. *Terminology*. Vol.9, No. 2: 201–219.

Pearce, D. 2002 A comparative evaluation of collocation extraction techniques. In *Third international conference on language resources and evaluation*. Spain, Las Palmas.

Pecina, P. 2010 . Lexical association measures and collocation extraction. *Language Resources and Evaluation*. 44:137–158.

Ramisch, C., Villavicencio, A., Boitet, C. (2010). "mwetoolkit: a Framework for Multiword Expression Identification", *Proceeding of LREC 2010*, Valletta, Malta.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In Proceedings of the third international conference on computational linguistics and intelligent text processing *CICLing-2002 , Lecture Notes in Computer Science,* London, UK, Vol. 2276 pp. 1–15 .

Seretan, V., Wehrli, E. 2009 . Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation* 43: 71–85.

Spina, S. 2010. The *Dictionary of Italian Collocations*: Design and Integration in an Online Learning Environment. *Proceeding of LREC 2010*, Valletta, Malta.

Villada Moiron, M.B. 2005. Data-driven identification of fixed expressions and their modifiability. PhD Thesis. University of Groningen.

Villavicencio, A., Ramisch, C., Machado, A., Caseli, H. de M., Finatto, M. J. 2010. Identificação de Expressões Multipalavra em Domínios Específicos. *Linguamática* 2 1 :15-33.

Zaninello, A., Nissim, M. 2010. Creation of lexical resources for a characterisation of multiword expressions in Italian. *Proceedings of LREC 2010*, Valletta, Malta. 654-661.

Zarrieß, S., Kuhn, J. 2009, Exploiting Translational Correspondences for Pattern-Independent MWE Identification. *Proceedings of ACL 2009*

Zinsmeister, H., Heid, U. 2003. Significant triples: Adjective+Noun+Verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003).* Budapest.

3 Survey of Existing Technologies, Tools and Resources

This section describes the technologies, tools and resources needed for each lexical acquisition task. It describes the existing technologies, tools and resources available to each partner for each task being undertaken. In areas where there is as yet no tool or resource available, this tool or resource will be developed during the course of the project.

3.1 Subcategorization Frames

This section describes the technologies and tools used for the automatic acquisition of subcategorization frames from corpora.

The basic resource requirements for SCF acquisition are: raw corpora (min. 100 occurrences per verb), text processing tools (including a tagger, a tokeniser, a lemmatiser, and a shallow parser or chunker – the parser/chunker must not already use SCFs), and SCF dictionaries for development and evaluation. The additional tools required are: a subcat classifier which extracts SCFs from parsed data, a lexical builder which constructs SCF entries from classified data, a filter which removes noisy SCFs, and evaluation resources (dictionaries and/or manually constructed resources).

3.1.1 University of Cambridge (UCAM)

UCAM has the following resources and tools available for English.

Required resources and tools	Available resources and tools	Comments

Raw corpora (min. 100 occurrences per verb)	yes	Several large corpora available, to be supplemented with project domain data
Text processing tools: tagger tokeniser lemmatiser shallow parser/chunker	yes	RASP
Subcat classifier	yes	
Lexical builder	yes	
Filter	yes	
Evaluation resources (SCF dictionaries)	yes	But need to develop domain-specific resources

Tools

We have a system for subcategorization frame (SCF) acquisition which can be used to acquire comprehensive lexicons for verbs, nouns and adjectives from un-annotated corpus data (Preiss et al., 2007). The system makes use of the RASP toolkit (Briscoe et al., 2006). RASP is a modular statistical parsing system which includes a tokenizer, tagger, lemmatizer, and a wide-coverage unification-based tag-sequence parser. We use the standard scripts supplied with RASP to output the set of grammatical relations (GR) for the most probable analysis returned by the parser or, in the case of parse failures, the GRs for the most likely sequence of subanalyses. The dependency relationships which the GRs embody correspond closely to the head-complement structure which subcategorization acquisition attempts to recover, which makes GRs ideal input to the SCF classifier.

The rule-based classifier incrementally matches GRs with the corresponding SCFs. The rules were manually developed by examining a set of development sentences to determine which relations were actually emitted by the parser for each SCF. The classifier identifies 168 verbal, 37 adjectival and 31 nominal frames. The SCFs recognized by the classifier were obtained by manually merging the frames exemplified in the COMLEX Syntax (Grishman et al., 1994), ANLT (Boguraev et al., 1987) and NOMLEX (Macleod et al., 1997) dictionaries and including additional frames found by manual inspection of unclassifiable examples during development of the classifier. These consisted of e.g. some occurrences of phrasal verbs with complex complementation and with flexible ordering of the preposition/particle, some non-passivizable words with a surface direct object, and some rarer combinations of governed preposition and complementizer combinations. The frames were created so that they abstract over specific lexically-governed particles and prepositions and specific predicate selectional preferences but include some derived semi-predictable bounded dependency constructions.

Lexical entries are constructed for each word and SCF combination found in the corpus data. Each lexical entry includes the raw and relative frequency of the SCF with the word in question, and includes various additional information e.g. about the syntax of detected arguments and the argument heads in different argument positions.

Finally the entries are filtered to obtain a more accurate lexicon. The system integrates a number of (relative) frequency–based and statistical filtering techniques. When filtering is done by using a very simple method, i.e. by setting empirically determined thresholds on the relative frequencies of SCFs, the system achieves state-of-the-art performance (over 70 F-measure) on

all the three sets on cross-domain corpus data. In addition, we have pioneered the use of weakly-supervised methods which can boost the baseline performance over 85 F-measure by smoothing verb specific SCF frequency distributions using back-off estimates based on relevant lexical semantic classes. Currently this technology is only applicable to verbs.

Evaluation resources

The performance is evaluated against a gold standard based on a manual analysis of some of the test corpus data (300 occurences per word), supplemented with additional frames from the ANLT, COMLEX and NOMLEX dictionaries. We have such gold standard data for 200 verbs, 30 nouns and 30 adjectives. We have also a merged version of ANLT, COMLEX and NOMLEX, but this purely dictionary–based gold standard does not include frequency data and is therefore not ideal for evaluation. In addition, domain-specific resources will need to be created.

References

B. Boguraev, J. Carroll, E. J. Briscoe, D. Carter, and C. Grover. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In Proc. of the 25th Annual Meeting of ACL, pages 193–200, Stanford, CA.

E. J. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the rasp system. In Proc. of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Preiss, J., Briscoe, E. J. and A. Korhonen. 2007. A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In the Proc. of ACL.

R. Grishman, C. Macleod, and A. Meyers. 1994. COMLEX Syntax: Building a Computational Lexicon. In COLING, Kyoto.

A. Korhonen and Y. Krymolowski. 2002. On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In Proc. of the Sixth CoNLL, pages 91–97, Taipei, Taiwan.

A. Korhonen. 2002. Subcategorization acquisition. Ph.D. thesis, University of Cambridge Computer Laboratory.

C. Macleod, A. Meyers, R. Grishman, L. Barrett, and R. Reeves. 1997. Designing a dictionary of derived nominals. In *Proc. of RANLP*, Tzigov Chark, Bulgaria.

3.1.2 University Pompeu Fabra (UPF)

UPF has the following resources and tools available for Spanish.

Required resources and tools	Available resources and tools	Comments
Raw corpora (min. 100 occurrences per verb)	yes	UPF has a 30M-word corpus that could be used, but is also interested in the induction of SCFs from a smaller corpus to reproduce the actual conditions of tuning to a new domain.
Text processing tools: tagger	yes – tagger, tokeniser, lemmatiser.	In 2011 UPF will build a treebank that will be used to

tokeniser lemmatiser shallow parser/chunker	no statistical parser	train a statistical parser, as required by this task.
Subcat classifier	no	
Lexical builder	no	
Filter	no	
Evaluation resources (SCF dictionaries)	yes	Evaluation to be done during the 2nd trimester of 2012.

Tools

UPF has no specific tools for verbal SCF acquisition, and the previous work on SCF acquisition for Spanish (see section 1.1) provides a reference only, rather than any particular tools or components.

Nevertheless, UPF is interested in using UCAM methods (and tools when possible) to create a system for subcategorization frame acquisition for Spanish verbs. The tools will need to be updated for Spanish. In what follows we analyse the required resources and language-dependent tools to assess the work that has to be done.

Acquisition experiment with 30M-word IULA-UPF corpus.

· Acquisition with smaller corpus to tune domain- dependent dictionaries.

· Research on "domain tuning"

Because of the availability of the statistical parser, the SCF acquisition experiment should be performed at the end of 2011.

Evaluation resources

The SCF dictionaries already available are: UPF has 2 general dictionaries used for MT and parsing with the following distribution:

INCYTA: 4887 verbs, 29782 nouns, 11992 adjectives, 2967 adverbs.

SRG's: 4329 verbs; 27755 nouns and 10212 adjectives

The available dictionaries will supply the basis and two domain specific evaluation corpora will be developed.

References

Alonso,L., I. Castellón, N. Tinkova (2007). Adquisición de subcategorizaciones verbales mediante un clasificador automático, *Revista de la SEPLN*

3.1.3 ILC-CNR

ILC-CNR has the following resources and tools available for Italian.

Required resources and tools	**Available resources and tools**	**Comments**
Raw corpora (min. 100 occurrences per verb)	yes	A general domain/newspaper corpus of 5M and possibly 20M is available. Plus domain corpora will come

		from WP4
Text processing tools: tagger tokeniser lemmatiser shallow parser/chunker	yes – tagger, tokeniser, lemmatiser. Dependency parser	Syn-SG. a rule-based parser up to dependency level.
Subcat classifier	no	adapt UCAM technology
Lexical builder	no	adapt UCAM technology and/or adopt LMF standard and adapt past experiences (BOOTStrep)
Filter	no	adapt UCAM technology
Evaluation resources (SCF dictionaries)	yes	Some lexica with SCF available.

Tools

Apart from the experience in LE-SPARKLE, no tool is at the moment available for automatic SCF acquisition. We plan to develop such tool in the context of PANACEA. We have raw corpora, a parser (dependency parser) and lexica with SCF information which can be used as gold standards. Previous works like Federici et al. 1998 and Lenci et al. 2008 can be taken as reference.

ILC-CNR is interested in extending UCAM methods for acquiring SCF information (development of tool and corresponding lexicon) for Italian verbs and nouns.

ILC-CNR has at its disposal a 5M word corpus and an additional 20M word corpus which can be used to induce a general system for SCF. Tuning domain will be done by exploiting the domain specific 1M word monolingual corpora which will be acquired in the context of PANACEA, WP4.1.

ILC-CNR has at its disposal the Synthema Slot Grammar (Syn SG), a multilingual rule-based parser, performing document and sentence segmentation, word tokenization, Part-of-Speech tagging, lemmatisation, Chunking and Dependency Parsing.

Evaluation resources

ILC-CNR has at its disposal a generic syntactic dictionary, the LE-PAROLE lexicon. the LE-PAROLE at the syntactic level consists of 20,051 unique on word-entries selected from the most frequent words I the ILC Italian Reference Corpus (Bindi et al. 1991). The lemmas belong to the following part of speech: verbs (3,120), nouns (13,212), adjectives (2,997), adverbs (562) and empty words (160).

A PAROLE syntactic entries encodes the specific properties /restrictions of a lemma and of its subcategorizing elements in a given syntactic construction. All the general properties shared by whole word classes (e.g. for verbs, passivization, pro-drop, postponed subjects etc) are assumed to be within the competence of the grammar. In the Italian lexicon predicate arity has been limited to 4 arguments maximum. The PAROLE Linguistic Specification proposes a liberal definition of frame: a distinction is drawn between lexically governed syntactic context and non lexically governed ones rather than between arguments and adjuncts. A position filler is considered as syntactically strongly bound provided that it is lexically selected by the head, no matter if it is an argument or an adjunct. However, fillers are distinguished between obligatory and optional. As for nouns complements, simple noun complements were considered as

optional, while object-like deverbal noun complements were marked as obligatory. In figurative meanings simple and deverbal noun complements were considered as obligatory. Different syntagmatic realization of a paradigmatically-related alternating slot filler in a frame was clustered in a single description.

Summing up, the LE-PAROLE syntactic lexicon can be used as a gold standard resource for subcategorization frames for verbs, nouns and adjectives. Previous experiments (subcategorization acquisition in the LE-SPARKLE project) has proved its validity. shortcomings are common to the use of a static resource for evaluating this kind of linguistic information, namely the absence in the resource of subcategorization frames automatically acquired.

A further dictionary, VERBAT, which encodes information on 12,000 Italian verbs at level of sense and subcategorization information is in phase of recovery. However, we could not commit to the use of this additional resource in PANACEA.

References

Bindi R., Monachini M., Orsolini P. 1991. Italian Reference Corpus - General information and key for consultation of contexts**.** NERC Working paper: WP7 - Acquisition and Reusability. ILC-TLN-1991-1, Istituto di Linguistica Computazionale, Pisa, Dicembre 1991

Federici S., Montemagni S., Pirrelli V. 1998. Chunking Italian: Linguistic and Task-oriented Evaluation. LREC 1998: Workshop on the Evaluating of Parsing Systems. Paris, ELRA.

Lenci A., McGillivray B., Montemagni S., Pirrelli V. 2008. Unsupervised Acquisition of Verb Subcategorization Frames from Shallow-Parsed Corpora**.** LREC 2008: *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), Marrakech, Morocco, May 26-1 June 2008, 3000-3006. CD-ROM.

Ruimy N., Corazzari O., Gola E., Spanu A., Calzolari N., Zampolli A. 2003. **T**he *PAROLE* model and the Italian Syntactic lexicon**.** In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. *Linguistica Computazionale*, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 793-820.

3.1.4 ILSP

ILSP has the following resources and tools available for Greek.

Required resources and tools	**Available resources and tools**	**Comments**
Raw corpora (min. 100 occurrences per verb)	yes	
Text processing tools: tagger tokeniser lemmatiser shallow parser/chunker	yes	
Subcat classifier	no	
Lexical builder	no	
Filter	no	
Evaluation resources (SCF	yes	

dictionaries)		

Tools

Available resources and tools that will be utilized are:

- large (100 mws) general domain corpora,

- a tool processing chain developed at ILSP that consists of a tokeniser, a tagger, a lemmatizer and a shallow parser,

Evaluation resources

- LEXIS, a Greek computational lexicon of general language created from a general language corpus, which comprises approximately 69,000 entries containing morphological information, of which a subset of 32,000 entries also contains syntactic information and a further subset of 15,000 includes semantic information. The syntactic level contains around 8,000 verbal syntactic units, all bearing subcat information: information is provided as regards the number of complements that each syntactic unit can subcategorise for, as well as their identification, i.e. their syntactic function, morphosyntactic realisation and optionality The LEXIS lexicon is an extension of the PAROLE/ SIMPLE lexica as regards the size, but also as regards the model as such, in order to cater for the idiosyncrasies of the Greek language. This lexicon will serve as validation resource against which the subcat frames acquired from the corpora will be validated.

References

Anagnostopoulou D., E. Desipri, P. Labropoulou, E. Mantzari, M. Gavrilidou. 2000. "LEXIS - Lexicographical infrastructure: systematising the data". *Comlex2000*, Patras, Greece.

Boutsis S., P. Prokopidis, V. Giouli, S. Piperidis. 2000. A robust parser for unrestricted Greek text. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 467 – 474.

3.2 Selectional Preferences

This section describes the technologies and tools used for the automatic acquisition of Selectional preferences from corpora.

The basic resource requirements for SP acquisition are: raw corpora and a parser. An optional tool is: a SCF acquisition system.

3.2.1 University of Cambridge (UCAM)

UCAM has the following resources available for English.

Required resources and tools	**Available resources and tools**	**Comments**
Raw corpora	yes	

Parser	yes	
SCF aquisition system	yes	

Tools

Most work on corpus-based induction of Selectional Preferences (SPs) has involved collecting argument headwords from data and generalizing to semantic classes in lexical resources like WordNet (Miller, 1990). However, WordNet-based approaches do not always outperform simple frequency-based models in SP acquisition (Brockmann and Lapata, 2003), and reliance on manually-compiled resources is not optimal in specific domains or languages.

In our recent experiments on English (Korhonen et al., 2008; Sun and Korhonen, 2009) we inferred semantic classes directly from corpus data: we acquired SPs from argument head data stored in a SCF lexicon extracted using RASP and our English SCF system. Two types of SP models were compared: raw argument head types and classes obtained using clustering (spectral clustering). The latter yielded a better result. The model was evaluated in a task-based setting where it improved the performance of lexical classification and via qualitative analysis which showed that it captured semantically meaningful preferences.

Evaluation resources

We have so far done task-based and qualitative evaluation. More investigation needs to be done into readily-available datasets as well as the development of domain-specific resources..

References

Brockmann, C. and M. Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In Proc. of EACL.

Miller, G. A. 1990. WordNet: An on-line lexical database. International Journal of Lexicography, 3(4):235–312.

Korhonen, A., Y. Krymolowski, and N. Collier. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. In Proc. of COLING.

Sun. L. and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In Proc. Of EMNLP. Singapore.

3.2.2 ILC-CNR

ILC-CNR has the following resources available for Italian.

Required resources and tools	**Available resources and tools**	**Comments**
Raw corpora	yes	
Parser	yes	Dependency parser
SCF aquisition system	no	Also being developed as part of the project

Tools

No tool is available for the acquisition of selectional preferences. Such a tool will be developed in the context of PANACEA. For its development we have at our disposal raw corpora and the semantic lexicon PAROLE/SIMPLE/CLIPS which can be used as a gold standard.

For WN-like approaches, ILC-CNR has at its disposal ItalWordNet (IWN). As an additional semantic resource we have at our disposal a rich semantic lexicon, PAROLE/SIMPLE/CLIPS. The SIMPLE lexicon is a four-layered computational lexicon developed under two EU-sponsored project (PAROLE and SIMPLE) and extended under the Italian government founded project CLIPS. It represents the largest computational lexical knowledge base of Italian language, containing over 45,000 lemmas and more than 57,000 word senses, or semantic units. At the semantic layer of information, lexical units are structured in terms of a semantic type system and are characterized and interconnected by means of a rich set of semantic features and relations. A SIMPLE/CLIPS lexical entry consists of a bundle of information, expressed in terms of valued features and relations between semantic units. For each entry it is possible to identify up to eight different levels of information. As for SPs acquisition, the most relevant level of information is represented by the argument structure. At this level, each predicative semantic unit, be it a verb, deverbal, deadjectival or simple noun, is assigned a lexical predicate. For verbs and simple, i.e. non derived, predicative nouns, the predicate names coincides with the semantic unit naming, e.g. SemU correre ←→ Pred correre. On the other hand, deverbal nouns share with their verbs the same predicates, thus "accusatore" [accuser], "accusato" [accused] and "accusa" [accusation] all point to the verb predicate "accusare" [to accuse], no matter their semantic type. Moreover, each predicative semantic unit is assigned a predicate-argument structure in terms of predicate's arity, semantic role and semantic type preference of each argument. For instance, the predicate for "guidare" [to drive] contains two arguments. The first argument has the semantic role "Agent" and two semantic preferences, corresponding to two ontological semantic types, "Human — HumanGroup". The second argument has the semantic role "Patient" and preference for the semantic type "Vehicle". It is worth noting that the encoding of preferences on arguments entails that the lexical resource provides information not only on word senses (ontological classification and rich semantic description) but also on their semantic context. For the use of the PAROLE/SIMPLE/CLIPS lexicon, preliminary refinement work will be necessary.

For corpus based approaches we have at our disposal a 5M word corpus and possibly a 20M word corpus which can be used to develop test and training data.

Evaluation Resources

The semantic lexicon PAROLE/SIMPLE/CLIPS can be used as a gold standard.

References

Bindi R., Monachini M., Orsolini P. 1991. Italian Reference Corpus - General information and key for consultation of contexts . NERC Working paper: WP7 - Acquisition and Reusability. ILC-TLN-1991-1, Istituto di Linguistica Computazionale, Pisa, Dicembre 1991

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., Rossi S. 2003. A computational semantic lexicon of Italian: *SIMPLE***.** In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. *Linguistica Computazionale*, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 821-864.

3.3 Lexical-semantic Classes

This section describes the technologies and tools used for the automatic acquisition of lexical-semantic classes from corpora.

3.3.1 University of Cambridge (UCAM) – lexical-semantic classes for verbs

The basic resource requirements for LC acquisition for verbs are: raw corpora, text processing tools (including a tagger, a tokeniser, a lemmatiser, and a shallow parser), and an SCF acquisition system.

UCAM has the following resources and tools available for English.

Required resources and tools	**Available resources and tools**	**Comments**
Raw corpora	yes	
Text processing tools: tagger tokeniser lemmatiser shallow parser	yes	
SCF acquisition system	yes	

Tools

We have a system which discovers lexical (syntactic-semantic) verb classes of the style found in (Levin, 1993) and VerbNet (Kipper-Schuler, 2005) in corpus data (Sun and Korhonen, 2009). The system extracts features from corpora which can indicate lexical classes. We employ a wide range features extracted from raw, tagged, lemmatized and/or parsed corpus data: co-occurrences, prepositional and lexical preferences (of verbs), tense (POS tags of verbs), voice (passive or active), SCFs parameterized for prepositions and other information, including verb selectional preferences. For classification we employ various methods. We have implemented both unsupervised methods (e.g. nearest neighbours, information bottleneck, information distortion, PLSI, spectral clustering) as well as supervised ones (e.g. SVMs, Gaussian). We have so far reported our best result using SCF+SP features and spectral clustering (Sun and Korhonen, 2009): around 80 F-measure when evaluated on the dataset of Sun et al., (2008).

Evaluation resources

The resources that can be used for evaluation include Levin's (1993) classification, its extended version in VerbNet (Kipper-Schuler, 2005), and the datasets of Joanis et al. (2008) and Sun et al., (2008), which include subsets of Levin classes. To identify error types and discover novel classes missing in gold standards, evaluation against gold standards is often supplemented with qualitative analysis of data.

References

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. Natural Language Engineering, 2008.

Beth. Levin. 1993. English verb classes and alternations: A preliminary investigation. Chicago, IL.

Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania.

Korhonen, A., Y. Krymolowski, and N. Collier. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. In Proc. of COLING.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. Lecture Notes in Computer Science, 4919:16.

Sun. L. and A. Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. In Proc. Of EMNLP. Singapore.

3.3.2 University Pompeu Fabra (UPF) – lexical-semantic classes for nouns

The basic resource requirements for LC acquisition for nouns are: raw corpora, cues for classes, and a decision tree classifier.

UPF has the following resources and tools available for Spanish and English.

Required resources and tools	Available resources and tools	Comments
Raw corpora	yes	
Cues for classes	yes Spanish / no English	
Decision tree classifier	yes	

Tools

The ultimate goal of UPF tools for lexical classification of nouns is to develop a system, which users can take for building a dictionary according to their needs. The basic idea is that users define the cues that can identify the class they are trying to annotate lexica with. These cues, probably in conjunction with other more general ones, will be sought in corpus data for known members of the class (selected by the user) in order to prepare a training test-set (as small as possible). Once trained, the system will classify the rest of nouns that the user wants to encode. The experiment has to find the feasibility for new classes (for instance, emotion related nouns for new opinion mining systems…).

For this purpose, UPF already has a series of components that perform these different tasks:

- Definition and access to corpus data
- Definition of cues with Regular Expressions tuned to the annotated corpus
- Development of the training set for a particular class/feature
- Training of a Decision Tree
- Execution of the classification exercise

The main functionality is to build a vector that represents whether or not a number of contexts, as expressed by means of regular expressions, have been matched in the word occurrences in a corpus. The system first builds a binary vector for every occurrence of a particular type in a corpus, and one vector is built for every occurrence. This first vector of vectors can be later transformed into a frequency based unique vector for each word type, or into a smoothed vector, for instance (Bel, 2010). The vector of binary vectors is wrapped in an XML file that indicates the number and name of the features (this has to be specified in a file) and the number of times each vector is repeated. This means that the number of different vectors and the number of times

that are produced sum up the information obtained by running the regular expressions in the concordances file.

A second module (Legolab) transforms, as indicated by the user, the binary vectors into a flat vector that sums up all the occurrences. Legolab can also deliver the set of vectors in a Weka format, which can be also be used from Legolab to train and test the J48 decision tree classifier (Witten and Frank, 2005).

References

Bel, Núria (2010). "Handling of Missing Values in Lexical Acquisition" dins Calzolari, Nicoletta et al. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Paris, France: European Language Resources Association (ELRA). Pàg. 2728-2735. ISBN 2-9517408-6-7

Witten, I. H. and Frank E. 2005. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.

3.4 Multi-Word Expressions

The basic resource requirements for MWE acquisition are: raw or part-of-speech tagged corpora, tools for calculating statistical co-occurrences of words with different association measures, filtering/classification techniques.

In order to explore whether we can obtain better accuracy, the possibility of exploiting a chunked or dependency parsed corpus will also be explored, although it is still an open question whether more sophisticated linguistic information significantly improves the accuracy of results. In the context of the project, however, given that in a platform such as PANACEA higher precision is of greater value, even a small improvement could be a big benefit.

3.4.1 ILC-CNR

ILC-CNR has the following resources and tools available for Italian.

Required resources and tools	**Available resources and tools**	**Comments**
Raw or POS-tagged corpora	yes	
Tools for calculating statistical co-occurrences	no	
Filtering/classification techniques	no	

Tools

Currently, there is no stable system for full MWE acquisition, but some methods for identifying candidates on a chunked corpus are available, and research on MWE and collocation extraction and representation is being carried out independent of the project. Also available is the MultiwordTagger developed within the Kyoto project, a multilingual multiwords tagger which uses information in wordnets and domain resources to tag multi word terms in texts. As this is not properly an acquisition module, it will not be employed as is in PANACEA, but it may be useful as a source for evaluation of our results.

4 Evaluation

For information on the evaluation plans for the components and resources described in this deliverable, see D7.1.

5 Resource Building

As mentioned in Section 1, not all of the lexical acquisition tasks investigated during the project will necessarily result in a PANACEA fully-integrated component and thus produce a full lexicon, as at this early stage it is not possible to anticipate the performance of the tools to be developed or adapted for some tasks. For the tasks and languages resulting in a fully-integrated component, the results of the lexical acquisition components will be encoded in XML and, where possible, will be compliant to the LMF standard.

Domain-specific lexicons with the entries resulting from monolingual corpora analysis (WP4) will be created for verb SCFs for English, Spanish and Italian. The format of the SCF lexicons will be in the form of a list of components and sets of lists of components as possible alternations associated to the syntactic behaviour of verbs in a corpus. The format will be as flexible as possible and will be compliant with LMF[2] specifications. It will be possible to customize lexicons to include only information for which the system has a minimum confidence level, to increase precision, thus in addition to frequency information related to each SCF assigned to a particular type, a confidence score will also be supplied. Domain-specific SCF resources for Greek will also be developed as part of the outcome of the research on building an SCF component. If the research on this component is successful then a Greek lexicon may be created using the same format as the other languages.

Domain-specific SP resources will be issued for English and Italian as the outcome of the research on building a SP component. If the research on this component is successful, an LMF-compliant format will be defined later in the project.

Domain-specific LC resources for English verbs and English and Spanish nouns will be created as the outcome of the research on building an LC component. They will contain the verbs and nouns (type) found in the monolingual reference corpus. The format of the resource will be a type associated to one or more classes. The resource will also contain a confidence score that will make it possible to customize them to select only a set of high precision results. If the research on this component is sucessful, an LMF-compliant format will be defined later in the project.

Domain-specific MWE resources will be issued for Italian as the outcome of the research on building a MWE extraction component. The resource, compliant with LMF specifications, will not be a simple list of MWEs, but it will specify other relevant information that could be useful in applications such as MT. Obligatory information will be: the multiword unit, its component lemmas or forms, frequency of co-occurrence and association measure. Optional information could be: head of the multiword, syntactic structure (i.e. PoS pattern, dependency structure) and semantic relation between the content words in the unit.

[2] Lexical Markup Framework, www.lexicalmarkupframework.org , ISO-24613:2008.

6 Lexical Merger

By *lexical merging*, or *merging of dictionaries*, we refer to the process of composing a new lexicon out of two or more existing lexica.

Electronic (or printed) lexica are often based on information taken from different sources, and are created for different purposes. Therefore, the kind of information stored in the respective input lexica of a merging operation may be very different (e.g. one may contain syntactic information, whereas another contains semantic information). Thus, the resulting data of such an operation may contain overlapping and possibly inconsistent information. Lexica may also be structured differently, so that it may be necessary to convert them to a standard model such as LMF (Lexical Markup Framework) prior to the actual merging. These considerations make clear that the merging process is a nontrivial task.

However, merging resources together is becoming an increasingly important task, since it allows to have different levels of information wrapped up into a single powerful resource which can be easily usable by different NLP (Natural Language Processing) systems, or to obtain custom resources suitable to address a specific problem. Often the available resources are unbalanced with respect of the type of lexical information encoded, focusing on a particular type and not providing enough coverage of other aspects. In some other cases, they are too much or too little detailed for the specific purposes of applications.

The community is increasingly calling for new types of lexical resources that are openly customizable: lexicons that can be built rapidly, possibly by combining certain types of information while discarding others, and tailored to specific needs and requirements. Rather than building new lexical resources, the new trend focuses on trying to exploit the richness of existing lexicons.

6.1 Current techniques

Chan and Wu (1999) present a basic method to automatically generate a set of mapping rules between lexicons that employ different incompatible part-of-speech (POS) categories such as the ones found in the Brill's tagger and in the Moby lexicon. The authors look specifically at the problem that different lexicons employ their own POS tagsets that are incompatible with each other, owing to their different linguistic backgrounds, application domains, and lexical acquisition methods. Their strategy is to inspect the co-occurrence of tags on those lemmas that are found in both lexicons, and to use that information as a basis for creating POS mapping rules. The key steps of the algorithm are four:

1. generation of POS feature vectors;
2. generation of what the authors call an *anti-lexicon* containing *anti-lexemes* which are simple pairs that associate a lemma with an *anti-tag* (A POS tag is called an *anti-tag* a of a lemma if it can never be a tag of that lemma);
3. mapping rule learning algorithm: the idea is to assume that a mapping rule between two POS tags holds if the similarity between their feature vectors exceeds a preset threshold;
4. merging of the entries using the mapping rule.

In Monachini et al. (2006) the authors focus on merging the phonological layer of the PAROLE-SIMPLE-CLIPS lexicon and the LCSTAR pronunciation lexicon. They present a specific framework that provides a method to create new language resources via unification and

combination of different independently created existing sources. Their method consists of 3 steps:

1. conversion of native data structures and formats to a uniform structure and format (an LMF-compliant Interchange format);
2. identification of those parameters that detect equivalence between lexical entries in Lexicon A and Lexicon B and perform one-to-one mappings. The mapping is perforformed by an automatic routine that, given mapping rules, compares two entries from Lexicon A and Lexicon B (entry a and entry b) and tests their equivalence over a mapping window (ortography, lemma, transcription, IF). Entry a and entry b are considered equivalent and candidates to become an entry in the unified lexicon, if all fields of the mapping window perfectly coincide;
3. fusion of source entries candidate to the merging into one unified entry.

Ruimy and Roventini (2005), Ruimy (2006) and Roventini et al. (2007) describe the efforts done to map ItalWordNet and the semantic and lexical level of PAROLE-SIMPLE-CLIPS. The authors' aim is to semi-automatically link and eventually merge the two lexicons so that the end user can dispose of a more exhaustive and in-depth lexical information combining the potentialities features offered by the two lexical models. Mapping is performed on a semantic type-driven basis. A semantic type of the SIMPLE ontology is taken as starting point. Considering the type's SemUs along with their PoS and 'isa' relation, the IWN resource is automatically explored in search of linking candidates with same PoS and whose ontological classification matches the correspondences established between the classes of both ontologies. The mapping process consists of the following steps:

1. selection of a PSC semantic type and definition of the loading criteria, i.e. either all its SemUs or only those bearing a given information;
2. selection of one or more mapping constraints on the basis of the correspondences established between the conceptual classes of both ontologies, in order to narrow the automatic mapping;
3. human validation of the automatic mapping and storage of the results;
4. if necessary, relaxation/tuning of the mapping constraints and new processing of the input data.

The work described in Crouch and King (2005) is particularly interesting. The goal is to merge the information coming from XLE syntactic lexicon, WordNet, Cyc, and VerbNet and put it in a uniform format to build a Unified Lexicon (UL) with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning and to understand where gaps in information arise across the merged resources. Merging is achieved via four steps:

1. the data is automatically extracted from the external resources;
2. the extracted data is merged into the UL entries;
3. the UL entries are corrected with hand-coded and automatically created patch files;
4. mapping rules are extracted from the UL.

It is worth noting that WordNet class information is crucially used to determine whether entries from Cyc and VerbNet could be merged.

LEXUS (Kemps-Snijders et al., 2006) is a web-based application based on LMF aimed at providing a flexible framework for maintaining structure and content of lexica. According to the authors, users can perform advanced cross-lexica operations, such as searching, comparing and merging of lexica. LEXUS proposes a general model for the process of merging that incorporates tasks such as the identification of related lexical entries, restructuring of lexical information, and handling of inconsistent data, all of which can be done automatically or manually. Users may monitor every step of the merging process and override values that have been produced automatically. LEXUS is available at http://www.lat-mpi.eu/tools/lexus/.

Soria et. al. (2006) present LeXFlow, a web application framework where lexica already expressed in LMF semiautomatically interact by reciprocally enriching themselves. LeXFlow is intended as an instrument for the development of dynamic multi-source lexica and as a way to promote the adoption of standards. In a way similar to the one implemented in document workflow (Marchetti et al, 2005), lexical entries move across *agents* and become dynamically updated. *Agents* can be either human or software actors. An entry of a lexicon A becomes enriched via basically three steps:

1. it is mapped onto a corresponding entry belonging to a lexicon B;
2. the entry inherits the semantic relations available in lexicon B;
3. the relations acquired are integrated into the entry and proposed to the human encoder.

As a result of the lexical flow, in addition, for each starting lexical entry (LA) mapped onto a corresponding entry (LB) the flow produces a new entry representing the merging of the original two.

Molinero et al. (2009) describe a method for building a large morphological and syntactic lexicon (the Leffe - Lexico de formas flexionadas del espanol) by merging existing resources. The methodology is based on the work of Sagot et al. (2006), Sagot and Danlos (2008) which applied it first to French. In order to allow the merging of the resources, their original formats were first converted to a common format developed in the Alexina framework. The conversion to the Alexina format is done by applying specific solutions on the basis of the lexicon type (morphological vs. syntactic). The merging of the morphological lexica rely on lemmas which are common to the two original resources (Multext [Ide and Veronis, 1994] and the USC lexicon [Alvarez et al., 1998]). Exceptions were resolved by giving priority to the Multex lexicon, which was considered as the baseline. The merging of the syntatic lexica (ADESSE lexicon [Garcia-Miguel and Albertuz, 2005] and SRG lexicon [Marimon et al., 2007]) exploited the fully specified syntactic information, i.e. no alternatives and no facultative arguments. The two lexica thus expanded can be easily merged by observing common expanded syntactic frames and then factorized to reduce the size. The results is a new syntactic lexicon which is trivially merged with the morphological one. Those morphological entries which missed syntactic information were assigned a default transitive syntactic frames.

References

Concepcion Alvarez, Pilar Alvarino, Adelaida Gil, Teresa Romero, Maria Paula Santalla, and Susana Sotelo. “Avalon, una gramatica formal basada en corpus”. In Procesamiento del Lenguaje Natural (Actas del XIV CONGRESO de la SEPLN), pages 132–139, Alicante, Spain, 1998.

Daniel Ka-Leung Chan and Dekai Wu. 1999. "Automatically Merging Lexicons that have Incompatible Part-of-Speech Categories".

Crouch and King. 2005. "Unifying lexical resources" Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes; 2005 February 28 - March 1; Saarbruecken; Germany. pp. 32-37.

José M. Garcia-Miguel and Francisco J. Albertuz. "Verbs, semantic classes and semantic roles in the ADESSE project". In Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.2005.

Marc Kemps-Snijders, Mark-Jan Nederhof, and Peter Wittenburg. 2006. "LEXUS, a web-based tool for manipulating lexical resources." *Proceedings of LREC2006*, Genoa, Italy.

Nancy Ide and Jean Véronis. "Multext: Multilingual text tools and corpora", in Proceedings of COLING-94, 1994

Andrea Marchetti, Maurizio Tesconi and Salvatore Minutoli. 2005. "XFlow: an Xml-Based Document-Centric Work- flow", in Web Information Systems Engineering – WISE 2005, pp. 290-303.

Montserrat Marimon, Natalia Seghezzi, and Nuria Bel. "An open-source lexicon for spanish". In Sociedad Espanola para el Procesamiento del Lenguaje Natural, n. 39, 2007.

Molinero, A.Miguel, Benoit Sagot, Lionel Nicolas. "Building a morphological and syntactic lexicon by merging various linguistic resources", in Proceeding of the NODALIDA Conference, pp 126-133.

Monica Monachini, Nicoletta Calzolari, Khalid Choukri, Jochen Friedrich,Giulio Maltese, Michele Mammini, Jan Odijk & Marisa Ulivieri. 2006. "Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian", in Calzolari et al. (eds.), LREC2006: 5th International Conference on Language Resources and Evaluation: Proceedings, pp. 1852-1857, Genoa, Italy.

Adriana Roventini, Nilda Ruimy, Rita Marinelli, Marisa Ulivieri, Michele Mammini. 2007. "Mapping Concrete Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results." Proceedings of the ACL 2007 Demo and Poster Sessions, pages 161–164, Prague, June 2007.

Nilda Ruimy. 2006. "Merging two Ontology-based Lexical Resources". LREC Proceedings, 1716--1721.

Nilda Ruimy and Adriana Roventini. 2005 "Towards the linking of two electronic lexical databases of Italian", In Zygmunt Veutulani (ed.), L&T'05.

Benoit Sagot and Laurence Danlos. "Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français". In Proceedings of the workshop "Lexicographie et informatique : bilan et perspectives", Nancy, France, 2008

Benoit Sagot, Lionel Clément, Eric Villemonte de La Clergerie, and Pierre Boullier. "The Lefff 2 syntactic lexicon for French: architecture, acquisition, use". In Proceedings of LREC'06, 2006.

Claudia Soria, Maurizio Tesconi, Francesca Bertagna, Nicoletta Calzolari, Andrea Marchetti and Monica Monachini. 2006. "Moving to Dynamic Computational Lexicons with LeXFlow", *Proceedings of LREC*.

7 Work Plan

This section summarises the areas in which each partner intends to explore the development of lexical resources. Where tools and resources already exist, the focus is on improving and adapting these tools and resources; whereas for languages without existing resources, the focus is on the initial development of a prototype which can work with the overall PANACEA architecture.

All development will focus on the domain-specific data obtained from WP4 rather than general text.

Note that some effort in WP6 may need to be dedicated to modifying components as required so that they can be wrapped as a web service in WP3. This is not specifically noted under each individual task below.

7.1 Subcategorization Frames

7.1.1 University of Cambridge (UCAM)

We plan to build domain-specific lexicons for SCFs for automotive and legal text. We will use the system described in Section 2 (as in Preiss et al., 2007), but will investigate ways of improving this system and adapting it to new domains. We focus on improving both the hypothesis generation and hypothesis selection steps of SCF acquisition.

The tagger and parser used for pre-processing in the hypothesis generation step of SCF acquisition have a large impact on the final accuracy of SCFs. Statistical techniques can be used to correct for noise in the parser output, but fundamentally the accuracy of this first stage remains crucial since detecting SCFs depends on syntactic analysis. As SCF systems have evolved, pre-processing has moved from lexical cues, to partial parsing, to full intermediate parsing. However, even in the last few years there have been further developments in tagging and parsing which could be important for SCF detection. Preiss et al. (2007) has already shown that using the latest version of the RASP toolkit (Briscoe et al., 2006) improved performance significantly. In addition to improvements in RASP, there are now other broad-coverage, high-accuracy unlexicalized parsers such as the Stanford parser (Klein and Manning, 2003) and the Berkeley parser (Petrov et al. 2006). (We focus on unlexicalized parsers since they do not already have knowledge about SCFs, which is what we want to learn; although it may be possible to use lexicalized parsers for SCF acquisition in a self-training context.) We plan to use the latest version of RASP and also to investigate whether other unlexicalized parsers can provide alternative views of the data, or be used in an ensemble for more accurate pre-processing. This will involve some re-engineering of the classifier in the existing SCF acquisition tool to work with other parser formalisms. Parser ensembles have been successfully used to improve parsing accuracy on both intrinsic (Sagae and Lavie, 2006) and extrinsic measures (Miyao et al., 2008) and for such tasks as pre-processing French text for manual annotation as part of a large corpus (Paroubek et al. 2010).

We will also look at retraining the POS tagger used in the RASP toolkit. A number of techniques for classifier domain adaptation have been introduced in the last few years (e.g. Daumé III, 2007) which make it possible to minimize the amount of manual annoation required in the new domain. We plan to investigate the use of such a technique. In general there has been increasing interest over the last few years in predicting the cross-domain performance of NLP tools based on text features (Rimell and Clark, 2008; McClosky et al., 2010; Van Asch and Daelemans, 2010) and it may be possible to model the automotive and legal domains in order to

predict at which part of the pipeline domain adaptation is most important. This can be done in conjunction with ongoing research at UCAM into lexical acquisition for different biomedical subdomains. Given that we already have an SCF acquisition tool for general English, the domain adaptation issue is a crucial one; Roland and Jurafsky (2002) compared SCF frequencies obtained from five different corpora and found that corpus variation was a major factor in SCF differences.

For hypothesis selection, the current state-of-the-art system (Korhonen et al., 2006) provides several smoothing and filtering techniques to improve the quality of automatically acquired SCF distributions and/or to create sub-lexicons suitable for different purposes. First, it is possible to customise the selection of verbs by frequency or according to a verb list. Second, the automatically acquired SCF distributions for individual verbs can be smoothed by add-one smoothing (Laplace, 1995), Katz backing-off (Katz, 1987), or linear interpolation (Chen and Goodman, 1996). For the latter two, smoothing uses the back-off estimates of the verb class of the most frequent WordNet sense of the verb. Finally, a subset of SCFs can be selected based on empirically defined filtering thresholds based on the absolute or relative frequencies of SCFs, statistical confidence tests, or the SCFs in the COMLEX and ANLT dictionaries. We plan to improve these methods further and to re-train them so that they work optimally with the modified classifier resulting from the improvements to hypothesis generation. In addition, we plan to investigate whether smoothing can make use of lexical-semantic class information obtained automatically from domain-specific corpus data, making use of automatically-acquired selectional preferences (see section 3.2.1).

We also plan to investigate whether extrinsic evaluations can help identify an appropriate level of precision in SCF acquisition for extrinsic tasks such as MT. Carroll, Minnen and Briscoe (1998) have shown that SCF frequencies can improve precision for a lexicalised parser. It may be possible to investigate the interface between SCFs and other applications including parsing, MT, or IE to help determine the level of precision in SCFs required by the application.

It will be necessary to develop an evaluation corpus of 20-30 verbs from each domain, with manually annotated examples from a relevant corpus. We plan to measure human annotation time in order to determine how much benefit can be obtained by automatic acquisition of SCFs.

References

H. Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

D. Klein and C.D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

D. McClosky, E. Charniak, and M. Johnson. Automatic Domain Adapatation for Parsing. North American Chapter of the Association for Computational Linguistics - Human Language Technologies 2010 Conference (NAACL-HLT 2010), Los Angeles, CA.

Y. Miyao, R. Sætre, K. Sagae, T. Matsuzaki, and J. Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL*. Columbus, Ohio.

P. Paroubek, O. Hamon, E. de La Clergerie, C. Grouin and A. Vilnat. 2010. The Second Evaluation Campaign of PASSAGE on Parsing of French. In *Proceedings of LREC.* Malta.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In Proceedings of COLING-ACL 2006).

L. Rimell and S. Clark. 2008. Adapting a Lexicalized-Grammar Parser to Contrasting Domains. In *Proceedings of EMNLP*. Honolulu, Hawaii.

D. Roland and D. Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In S. Stevenson and P. Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam: John Benjamins, 325-346.

K. Sagae and A. Lavie. 2006. Parser combination by reparsing. In *Proceedings of HLT-NAACL*. New York.

V. Van Asch and W. Daelemans. 2010. Using domain similarity for performance estimation. In P*roceedings of the ACL 2010 Workshop on Domain Adaptation for Natural Language Processing*. Uppsala, Sweden.

7.1.2 University Pompeu Fabra (UPF)

We plan to develop the components and resources missing to provide a component for domain-specific lexicon for Spanish with the methods used by UCAM (Preiss et al. 2007).

As motivated in section 2.1.2, UPF has to develop a statistical parser in order to be able to do it. The statistical parser needs a Treebank whose development is beyond the scope of this project. However, UPF is planning to have a Treebank (because of its participation in another project) in late 2011. We plan to use it to derive the parser required for this task. Thus, the exercise of developing components for SCF acquisition for Spanish to build domain-tuned lexica will not start until the last trimester of 2011. This planning will only have the positive consequence that UPF will be able to test the new developments made by UCAM for Spanish (cf. 3.1.1), further validating UCAM improvements.

Also in line with UCAM workplan, UPF is interested in participating in extrinsic evaluations by using the SRG grammar (Marimon et al. 2007), which requires SUBCAT information to produce rich information parses.

For the evaluation of the domain-based exercise, UPF will develop an evaluation corpus from two domains, with manually annotated examples for Spanish.

References

Marimon, Montserrat; Bel, Núria; Espeja, Sergio; and Seghezzi, Natalia (2007). "The Spanish Resource Grammar: Pre-processing Strategy and Lexical Acquisition" dins Baldwin, Timothy et al. (eds.) *Proceedings of the ACL2007 Workshop on Deep Linguistic Processing*. Stroudsburg, PA 18360: Association for Computational Linguistics. Pàg. 105-111. 2007. ISBN 978-1-932432-88-6

7.1.3 ILC-CNR

As stated in section 2.1.3 ILC-CNR plans to develop a subcategorization acquisition system. We plan to adapt UCAM technologies for lexical builder and filtering methods to remove noisy SCFs. Evaluation for the domain-based exercise will be conducted as stated in D7.1. In the second trimester of 2010 we plan to start the development of a general domain SCF acquisition system. This system will be evaluated against dictionaries and manual inspection. For the domain based exercise for SCF, ILC-CNR will develop an evaluation corpus from two domains (20-30 verbs from each domain, with manually annotated examples from a relevant corpus). In line with UCAM workplan, human annotation time will be measured in order to determine how much benefit can be obtained by automatic acquisition of SCFs. Furthermore, in the encoding

format of SCF we will work in the perspective of the LMF standard, which will use to facilitate the merging of dictionaries.

Tentative timeline: end of 2010 have a general domain SCF acquisition system and lexicon evaluated. Domain adaptation task will be performed as soon as the crawled monolingual corpora will be available and after the creation of the domain specific data set.

7.1.4 ILSP

Since ILSP does not have an SCF acquisition tool for Greek, we plan to develop one by first examining portability of algorithms developed by partners in the consortium. We plan to initially use a general domain corpus of 100+M EL corpus, and an existing subcat frame lexicon for evaluation on a small set of verbs. After that we will examine tuning this tool to one of the two domains targeted by the project, by a) selecting or, creating manually, entries for 20-30 verbs in the domain and b) extract SCF information from a small EL corpus in the domain.

7.2 Selectional Preferences

7.2.1 University of Cambridge (UCAM)

We plan to investigate whether we can achieve sufficient accuracy in automatic acquisition of selectional preferences to be useful for rule-based or statistical Machine Translation.

Recent work (Ó Séaghdha, 2010; Ritter et al., 2010) uses Latent Dirichlet Allocation (LDA) to model selectional preferences. We plan to investigate whether this technique can be applied to new domains. There is ongoing work at UCAM to apply these techniques to selectional preference modelling for biomedical data; and it will be informative to compare that domain with legal and automotive.

We plan to develop a manually annotated gold standard for 20-30 verbs. In addition, we will perform evaluation using pseudo-disambiguation with domain-specific examples. We will include experiments that follow the recommendations of Chambers and Jurafsky (2010) for pairing positive and negative examples according to frequency, and for including seen as well as unseen words in the test data.

References

N. Chambers and D. Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of ACL*. Uppsala, Sweden.

D. Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL*. Uppsala, Sweden.

A. Ritter, Mausam, and O. Etzioni. 2010. A Latent Dirichlet Allocation method for selectional preferences. In *Proceedings of ACL*. Uppsala, Sweden.

7.2.2 ILC-CNR

We plan to develop a system for SPs in Italian. Work in this area in Italian is at its beginning, since to the best of our knowledge the only work is that of Lenci et al (2010).

We will develop our system by exploiting corpus based techniques, in coordination with UCAM, by adapting their tools, when possible, to Italian.

As reported in D7.1, evaluation of the SP system will be done by means of pseudo-disambiguation. We will concentrate on developing a domain specific SP system for the PANACEA domains. As agreed a manually annotated corpus of 20-30 domain specific verbs will be developed.

As for the encoding format of SPs, we will work in the perspective of the LMF standard, which will use to facilitate the merging of dictionaries.

Timeline: work on SPs will start in the first trimester of 2011.

References

Lenci, A., Johnson M, and Lapesa, G. Building an Italian FrameNet through Semi-automatic Corpus Analysis. LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), La Valletta, Malta.

7.3 Lexical-semantic Classes

7.3.1 University of Cambridge (UCAM) – lexical-semantic classes for verbs

We plan to investigate whether we can achieve sufficient accuracy in automatic acquisition of lexical-semantic classes to be useful for rule-based or statistical Machine Translation.

Recent work uses SCF and SP features and spectral clustering to identify lexical-semantic classes (Sun and Korhonen, 2009). We plan to investigate the use of hierarchical clustering (Jardine and van Rijsbergen, 1971; Duda and Hart, 1973; Heller and Ghahramani, 2005; Yu et al., 2005), which returns a hierarchy of clusters rather than a flat, unstructured set. As with spectral clustering, the clusters are learned from the data rather than pre-specified. It is possible to choosing a level in the resulting hierarchy so as to yield clusters that are more or less fine-grained, meaning that an appropriate level of precision can be chosen for a given application. Automatic clustering is particularly relevant for domains where the Levin verb classes may not be appropriate, and when little training data is available, the flexibility provided by hierarchical clustering may be important. Preliminary work shows that automatic clustering is more accurate for specialised domains such as biomedical text than for general text, because there is less interference from multiple word senses. Hierarchical clustering has been used most successfully in Information Retrieval (Willett, 1988; Masłowska, 2003; Cowans, 2004; Haffari and Teh, 2009).

We will evaluate the automatically generated clusters for approximately 200 verbs against human judgements and, where appropriate, against VerbNet classes.

References

P. J. Cowans. 2004. Information retrieval using hierarchical Dirichlet processes. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR).

R.O. Duda and P.E. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley.

G. Haffari and Y.W. Teh. 2009. Hierarchical Dirichlet trees for Information Retrieval. In *Proceedings of NAACL*. Boulder, Colorado.

K.A. Heller and Z. Ghahramani. 2005. Bayesian Hierarchical Clustering. In *Proceedings of ICML*.

N. Jardine and C.J. van Rijsbergen , The Use of Hierarchic Clustering in Information Retrieval. Information Storage and Retrieval 7 (1971), pp. 217–240.

I. Masłowska. 2003. Phrase-based hierarchical clustering of web search results. In *Advances in Information Retrieval*. Springer.

P. Willett. 1988. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management* 24(5):577-597.

K. Yu, S. Yu, and V. Tresp. 2005. Soft Clustering on Graphs. In *Proceedings of NIPS*.

7.3.2 University Pompeu Fabra (UPF) – lexical-semantic classes for nouns

UPF has been working with methods that identify lexico-semantic classes of nouns and adjectives from corpus data. As Levin (1993) for verbal classes, our approach is to take similar syntactic behaviour and shared meaning components as a basis for the proposal for a particular class. If the class is well defined, then it has to be possible to recognize its members by observing its syntactic behaviour, and thus it is possible to train classifiers to do it (Bel et al. 2007, Resnik and Bel 2009, and Bel et al. 2010).

In more practical terms, our work wants to solve the problem of manual annotation for describing the meaning components of nouns in so that this annotation can contribute to solve a number of NLP tasks. Somehow inductively, meaning components or semantic features have been used as labels to assist rule-based components in the identification of arguments in a sentence, i.e. selectional restrictions, transfer rules in MT systems or inference mechanics for topic identification, etc.

Following Pustejovsky and Hanks (2001), we want to work with semantic features that have been empirically found to be prototypical in the description of selectional restrictions of different types of verbs. We want to motivate the existence of classes of nouns that correspond to these features and which can be justified in terms of similar syntactic behaviour, or, as Pustejovsky and Hanks (2001) suggest, similar "selection contexts", i.e. stereotypical syntagmatic patterns where nouns that, we can say, belong to the same class can be inserted. We also follow Jackendoff's (1983) proposal (pag. 139): "A word meaning, then, is a large heterogeneous collection of such (typicality) conditions dealing with form, function, purpose, personality or whatever else is salient. Taxonomic […] information also plays a role. As the importance of information for individuation and categorization drops off (as weighting, observability, or frequency of occurrence decreases), it shades toward "encyclopaedia" rather than "dictionary" information, with no sharp line drawn between the two types". Thus, we will not try to justify a possible evident taxonomical relation behind these features that we want to promote to classes in the terms defined before. Brandeis Shallow Ontology, BSO, is an attempt at doing it with classes that directly map onto the ones we are proposing here.

Our first selection of classes, very much based on the semantic features prototypically used for selectional restrictions, has been driven by practical motivations, and cannot be considered a model of lexical meaning. The following list of classes is based on the labels that a rule based MT system (INCYTA) and a rule based rich grammar for Spanish (SRG, Marimon et al. 2007) have used in order to define parsing rules. The granularity of this selection is motivated by the range of phenomena that current RMT systems can deal with. Thus, we have also included EVENT, which is not in the first list but that together with Process can account for a large range of language phenomena, and MASS, which is considered to be a separate grammatical feature rather than a semantic type. Our first list for classes of nouns to be learned is the following:

Abstract,
Animal (including microorganisms & animal groups),
Body part,
Concrete,
Event,
Human,
Location,
Mass,
Matter,
Plant,
Pot (machines, tools, technologies and natural phenomena),
Process,
Semiotics,
Social entities,
Temporal,
Units of Measurement.

As these classes have been used in an actual MT system, we have the possibility of evaluating our exercise intrinsically, by using the actual list of nouns that have been labelled and tested for years by this MT system, as well as extrinsically, by using them in parsing to obtain correct analyses in a particular grammar (Marimon et al., 2007). We are particularly interested in experimenting with domain tuning.

Besides, we have also noticed that lexico-semantic information of nouns (and other PoS) can be defined from different dimensions and for different purposes, especially when we consider domain dependent knowledge. Thus, besides this first list of classes, we want our system to consider also the possibility of having new or different classes. The system that has to be integrated into PANACEA must address the possibility of a user defining a new class. In order to allow it, we want to experiment with two scenarios:

Following Merlo and Stevenson's (1999) previous work, how to enable the user to define a reduced, ad-hoc linguistically motivated set of features that bring about distinctions among lexico-semantic classes.

Following Joanis et al. (2007), our intention is to investigate further the possibility of having a large, general and multipurpose set of linguistically motivated features that can be used to learn and classify any possible lexico-semantic class as defined by a user. We have tried the classification with different ML methods, specially Decision Trees, but also Bayesian methods that try to use information derived from a linguistic lexical model rather than from training data.

Our system, which works with Decision Trees (C45 implemented as J48 in Weka, by Witten and Frank 2005) achieves an accuracy of around and 80% in classifying EVENTS (for English and Spanish in Bel et al., 2010), and 65% in classifying MASS nouns (in Bel et al. 2007), for instance.

Our workplan is:

1. To carry out research to improve the achieved results in nominal lexical classification (for Spanish and English) by considering the following aspects:

- Bayesian methods for dealing with the problem derived from the similar frequency distribution of noise and significant patterns (sparse data problem), (Bel et al. 2008 and Bel 2010).
- Dealing with ambiguity in the classification of classes and how to handle it in the GL framework by taking into account regular polysemy.

2. To enlarge the coverage of the current models for lexical classification and to address all the semantic classes that we have proposed as initial list for Spanish and English.
3. To experiment the feasibility of a general approach in the lines of Joanis et al. (2007) and selection of better supervised techniques for larger dimensional spaces.
4. In addition to improvements in coverage and methods, UPF plans to develop the necessary changes to allow modules to be deployed as web services that can be integrated in the platform. The chaining of the services has to constitute a kind of laboratory where the user can define classes and train and test a classifier. The following services are being proposed:
    - Regular Expression matching, given a concordance file and a RE file, the system returns binary vectors.
    - Concordancer, given a lemma and a category, and the URI of an indexed and PoS tagged corpus, the system delivers a file with the concordances where the word that has been looked for is marked with ##. This Web service can be replicated in another one in which the corpus is not PoS tagged.
    - Vector transformers. The vector transformation modules are less interesting as webservices of a general purpose but for our goal will be also deployed as web services:
        - Frequency based transformer. Given a set of binary vectors for a particular type, the system returns a unique vector that sums up all frequency information
        - Mean Smoothed vector transformer. given a set of binary vectors for a particular type, the system returns a unique vector that sums up all frequency information and smoothes zero values with a calculated mean
        - Trimmed Mean Smoothed vector transformer. Given a set of binary vectors for a particular type, the system returns a unique vector that sums up all frequency information and smoothes zero values with a calculated treammed mean
    - Other webservices that will be required after the changes in methods. For instance, new smoothers.
    - Trainer webservice. Given a training set built from the collection of *n* samples of positive and negative examples of a class supplied by the user, the system has to produce the training data and to deliver a trained system.
    - N CLASSxLANGUAGE Classifier webservices. Given a corpus and a list of lemmas/PoS, the system produces a list of the lemmas, the classifier prediction for each one and its likelihood. This operation can also be

deployed as an open classifier, where the user also supplies the model as produced by the Trainer webservice.

References

Bel, Núria; Coll, Maria; Resnik, Gabriela (2010), Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production, to appear in Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010, Beijing, China.

Bel, Núria (2010). "Handling of Missing Values in Lexical Acquisition" dins Calzolari, Nicoletta et al. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Paris, France: European Language Resources Association (ELRA). Pàg. 2728-2735. ISBN 2-9517408-6-7

Resnik, Gabriela; Bel, Núria (2009). "Automatic Detection of Non-deverbal Event Nouns in Spanish" in -- Proceedings of the 5th International Conference on Generative Approaches to the Lexicon. Pisa: Istituto di Linguistica Computazionale.

Bel, Núria; Espeja, Sergio; Marimon, Montserrat (2008): Automatic acquisition for low frequency lexical items. Proceedings of the 6th International Conference on Language Resources and Evaluation. Paris: European Language Resources Association (ISBN 2-9517408-4-0)

Bel, Núria; Espeja, Sergio; Marimon, Montserrat. "Automatic Acquisition of Grammatical Types for Nouns" dins *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York: Association for Computational Linguistics. 2007. Pàg. 5-8. ISBN 1-932432-94-

Joanis, E; Stevenson, S; and James, D. 2007. A General Feature Space for Automatic Verb Classification. Natural Language Engineering.

Marimon, Montserrat; Natalia Seghezzi and Núria Bel. An Open-source Lexicon for Spanish. Procesamiento del Lenguaje Natural, n. 39, pp. 131-137. September, 2007. ISSN 1135-5948.

Marimon, Montserrat; Bel, Núria; Espeja, Sergio; and Seghezzi, Natalia (2007). "The Spanish Resource Grammar: Pre-processing Strategy and Lexical Acquisition" dins Baldwin, Timothy et al. (eds.) *Proceedings of the ACL2007 Workshop on Deep Linguistic Processing*. Stroudsburg, PA 18360: Association for Computational Linguistics. Pàg. 105-111. 2007. ISBN 978-1-932432-88-6

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.

Stevenson, Suzanne and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In Proc. of the 9th Conference of the European Chapter of the ACL, Bergen, Norway.

Witten, I. H. and Frank E. 2005. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.

7.4 Multi-Word Expressions

7.4.1 ILC-CNR

The ultimate goal is to build a component for acquiring MWes from domain corpora for building or enriching lexica with collocational information.

Since MWE is a wide area including various types of structures, work will be focussed on those MWEs which may benefit multilingual applications more and for which more robust methods exist: namely, nominal collocations (i.e. noun compounds, complex nominals, and adjective noun pairs). Thus, target MWEs will have the form NN (which in Italian is not very productive, but is still salient in domain terminologies), AdjN or Nadj, and N prep N.

First a system following the n-gram with POS filtering corpora will be built as baseline and various AMs will be used for ranking the candidates as in most common state-of-the-art methods described above. Then the system will be adapted to work on chunked coupus data in order to reduce noise in the candidate list and experiments will be done also with dependency parsed data to assess improvements in performance. In fact, it must be assessed on the specific case of PANACEA whether a syntactic approach is better, as the potential errors in the parsed input corpus may affect MWE extraction (cfr. Seretan and Wehrli, 2009:78).

Given that no ready-made tool will be used in this task, it design will try to take into account directly the requirements of the platform and in particular the fact that it should run as a webservice, which is no trivial issue given that MWE methods require processing of large quantities of data.

Work will start on Italian, using first general purpose corpora for developing and the acquisition component based on the state-of-the-art methods, and then the monolingual domain corpora obtained in WP4 for domain tuning. The official evaluation will be performed on domain data only.

Finally, given that the methods are relatively language independent (although in the literature it is reported that AM ranking works differently in different languages), the possibility of applying the same technology to English (and possibly other languages for which domain corpora will be crawled) will be assessed.

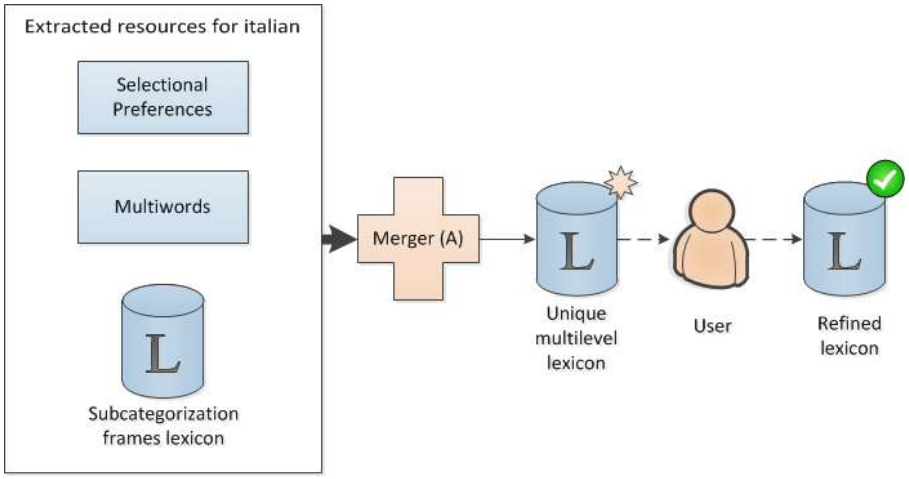
7.5 Lexical Merging

In task 6.3, a merger component will be integrated in the PANACEA platform. The merging process will regard both lexicons and lexical resources acquired in PANACEA itself.

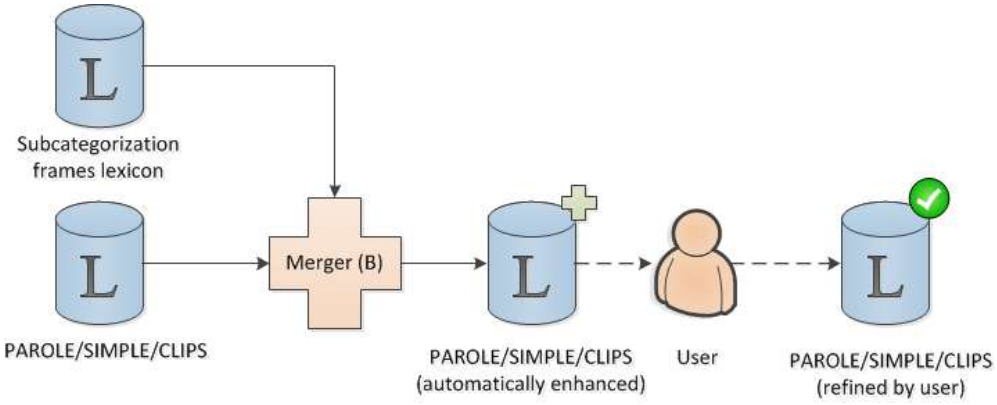
As stated in the overview of D6.1 document, “Lexicons” refers to areas of research that are already sufficiently well-developed: the research undertaken as part of the project can be expected to result in a relative good quality resource. The automatic acquisition of lexical information (SP - for English, Italian and Spanish, Lexical Classes - verbs for English and nouns for Spanish, and MWE - for Italian) is still in a phase that can be defined “experimental”: the methodologies used to acquire them and the results obtained require refinements. Part of the research in the PANACEA project aims, on the one hand, at improving the results of the systems involved and, on the other hand, to extend and develop systems and preliminary resources for less-resources languages such as Italian and Spanish. Nevertheless, although for these latter types of lexical information (SPs, MWEs and Lexical Classes) we will not commit to produce “Lexicons”, as a side effect of the development of the dedicated components we still obtain lexical resources.

The merging component in PANACEA will be two-fold:

1 it will integrate (join) the lexical resources produced by the PANACEA components into a unique multi-level lexicon. This newly obtained lexical resource will be performed for Italian for SCFs, SPs and MWEs. This resource will keep track as far as possible of the reliability of the information stored (e.g. by assigning confidence scores);



2 it will merge the information of the SCF lexicon - considered to be a relative good quality lexicon - into an existing lexical resource, namely PAROLE/SIMPLE/CLIPS, thus producing an enhanced version of it. This result can be further processed by a human agent, to obtain a more polished version of the new lexicon.



To develop such a component, we need to:

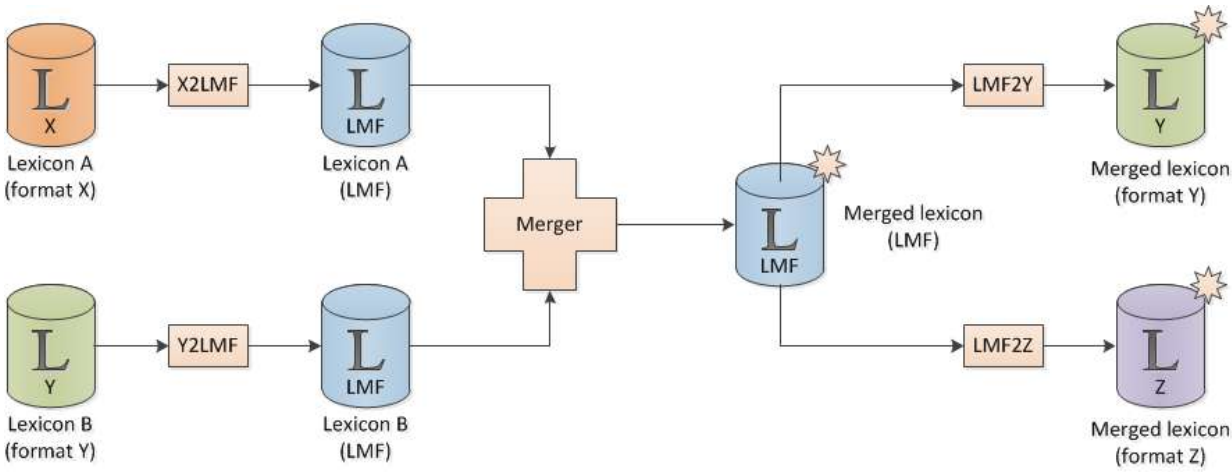
1. define the content of the input lexical resources and the content we want in the result;
2. define the format of the input lexical resources and the format we want for the result.

Addressing the first point is not easy, since the exact content of the input resources will be determined by the extractors procedures that are to be developed in tasks 6.1. E.g. it would be useful to know if selectional preferences and subcategorization frames are somehow directly connected, or if a connection to the sentence they were extracted from will be present. One issue

that makes the merging not trivial is the different word senses. When we need to merge a newly acquired lexicon with an existing one, we need to decide to which sense of a word the newly acquired information has to be added. To do so, it is necessary to compare the new information with the existing information for each sense and decide whether they are compatible with each other. Thus, the new information will be added to those senses that do not present an incompatibility. We think that is worth to study if this compatibility can be approached using graph theory (e. g. following a proposal similar to Graph Annotation Format (GrAF) (Ide and Suderman, 2007)). Other unification techniques will be evaluated, as soon as the content and the format of information to be handled will be clearer.

A related issue regards the need to decide how the lexical entries will be unified, i.e. which are the data categories to be mapped and how, determining the features that define whether the newly acquired information is compatible with the existing one or not. It is necessary to define the set of features that are decisive in the behavior of a word, thus ignoring the other ones. From these features, we will need to establish which ones have to be shared by the newly acquired information and by the set of information already associated to each word: these features will define the compatible set of information. We have to investigate more in that line, but we think that an interesting way to explore is the use of heuristics to determine the importance of each feature. Once again, though, other techniques can be investigated.

As for the formats, we envision the use of automatic components that performs conversions to and from LMF (or a yet to be investigated ad-hoc internal format) both in the input layer and in the output layer. Such a scenario is depicted in the figure below:



If the involved formats are XML-based, a conversion component can be implemented by using an XSLT (eXtensible Stylesheet Language Transformations) processor. An XSLT processor takes as its input an XML document and a special document written in the XSLT language (stylesheet) describing the conversion process, and produces a document in standard XML syntax or in another text format. Through this method we will be able to provide support for a wide variety of formats, without uselessly complicating the merger component.

In conclusion, in order to define a future workplan, we think that we should:

- continue exploring/evaluating merging techniques for Language Resources in general;

- as soon as a first data set is ready for testing, define merging techniques that better fit to automatic acquisition of PANACEA Lexical Resources.

References

Ide, N. and Suderman, K. (2007). GrAF: a graph-based format for linguistic annotations. In Proceedings of the Linguistic Annotation Workshop (Prague, Czech Republic, June 28 - 29, 2007). ACL Workshops. Association for Computational Linguistics, Morristown, NJ, 1-8.