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

D6.2

Integrated Final Version of the Components for Lexical Acquisition

Dissemination Level:	Public
Delivery Date:	28/12/12
Status – Version:	V1.10 **(Final)**
Author(s) and Affiliation:	Laura Rimell (UCAM), Núria Bel, Muntsa Padró (UPF), Francesca Frontini (CNR-ILC), Monica Monachini (CNR-ILC), Valeria Quochi (CNR-ILC).

Relevant Panacea Deliverables

D6.1	Technologies and Tools for Lexical Acquisition
D6.3	Monolingual lexica for English, Spanish and Italian tuned to aparticular domain (LAB and ENV)

Table of contents

1 Introduction

The PANACEA project has addressed one of the most critical bottlenecks that threaten the development of technologies to support multilingualism in Europe, and to process the huge quantity of multilingual data produced annually. Any attempt at automated language processing, particularly Machine Translation (MT), depends on the availability of language-specific resources. Such Language Resources (LR) contain information about the language's *lexicon*, i.e. the words of the language and the characteristics of their use. In Natural Language Processing (NLP), LRs contribute information about the syntactic and semantic behaviour of words – i.e. their grammar and their meaning – which inform downstream applications such as MT.

To date, many LRs have been generated by hand, requiring significant manual labour from linguistic experts. However, proceeding manually, it is impossible to supply LRs for every possible pair of European languages, textual domain, and genre, which are needed by MT developers. Moreover, an LR for a given language can never be considered complete nor final because of the characteristics of natural language, which continually undergoes changes, especially spurred on by the emergence of new knowledge domains and new technologies. PANACEA has addressed this challenge by building a factory of LRs that progressively automates the stages involved in the acquisition, production, updating and maintenance of LRs required by MT systems. The existence of such a factory will significantly cut down the cost, time and human effort required to build LRs.

WP6 has addressed the *lexical acquisition* component of the LR factory, that is, the techniques for automated extraction of key lexical information from texts, and the automatic collation of lexical information into LRs in a standardized format. The goal of WP6 has been to take existing techniques capable of acquiring syntactic and semantic information from corpus data, improving upon them, adapting and applying them to multiple languages, and turning them into powerful and flexible techniques capable of supporting massive applications. One focus for improving the scalability and portability of lexical acquisition techniques has been to extend exiting techniques with more powerful, less "supervised" methods. In NLP, the amount of supervision refers to the amount of manual annotation which must be applied to a text corpus before machine learning or other techniques are applied to the data to compile a lexicon. More manual annotation means more accurate training data, and thus a more accurate LR. However, given that it is impractical from a cost and time perspective to manually annotate the vast amounts of data required for multilingual MT across domains, it is important to develop techniques which can learn from corpora with less supervision. Less supervised methods are capable of supporting both large-scale acquisition and efficient domain adaptation, even in the domains where data is scarce.

Another focus of lexical acquisition in PANACEA has been the need of LR users to tune the accuracy level of LRs. Some applications may require increased precision, or accuracy, where the application requires a high degree of confidence in the lexical information used. At other times a greater level of coverage may be required, with information about more words at the expense of some degree of accuracy. Lexical acquisition in PANACEA has investigated confidence thresholds for lexical acquisition to ensure that the ultimate users of LRs can generate lexical data from the PANACEA factory at the desired level of accuracy.

This deliverable, D6.2, describes the development of lexical acquisition components for PANACEA, including the tools and technologies for each lexical acquisition task, and the integration of an appropriate subset of these tools into the PANACEA platform. In particular, the 3rd version of the platform has been released and represents the key time for lexical acquisition components to be deployed within PANACEA. This deliverable thus describes the lexical acquisition components that have been integrated in the 3rd version of the platform (see D3.4 for a full description of the platform itself), along with the Common Interfaces and Travelling Objects of the components. The tasks addressed in this work package are: Subcategorization Frame Acquisition (SCF), Lexical-semantic Classification (LC), Multiword Expression Acquisition (MWE), and Selectional Preference Acquisition (SP), as laid out in D6.1. There are currently 22 web services and 13 workflows related to lexical acquisition, divided among the tasks of SCF, MWE, and LC (lexicon merging components are presented separately, in D6.4). The documentation for these components is largely in the form of scientific papers.

The number and breadth of services deployed represents an improvement over the work plan laid out in D6.1. D6.1 called for research on all four lexical acquisition tasks, with a specific focus on integrating SCF components in the platform. Although it was at that time unknown whether tools for the other lexical acquisition tasks would be appropriate for integration, in fact there has also been a successful deployment of state-of-the-art MWE components (ILC-CNR) and LC components (UPF) as well. (As anticipated, SP components were not integrated due to the complexity of the underlying model being unsuitable for a web platform, although SP experiments have proceeded as planned.) We have also been able to deploy two language-independent SCF components, a strategy laid out at the WP6 technical meeting in December 2012, and consequently have planned a cross-linguistic SCF acquisition experiment.

D6.3 and D6.5 present lexicons based on multi-level merging, which will be presented separately in those documents.

It has been a goal of PANACEA to present experimental results in the form of scientific papers whenever possible, since a scientific paper not only codifies the results but has the potential to disseminate the results more widely among the scientific community. Throughout this deliverable, therefore, references are made to published or prepared scientific papers, and summaries of the techniques and main results in each paper are provided. The papers themselves are also provided as Annex A to this deliverable.

The following abbreviations are used throughout the deliverable. They are collected in this table for easy reference.

Abbreviation	**Definition**
LR	Language Resource.
NLP	Natural Language Processing.
SCF	Subcategorization Frame. Subcategorization refers to the tendency of a word to select the syntactic phrase types it co-occurs with.
SP	Selectional Preference. The tendency of a word to select the semantic types of its co-occurring phrases (arguments).

MWE	Multi Word Expression. A sequence of morphosyntactically separate words which form a semantic unit, often with a meaning unpredictable from the meanings of its component words.
LC	Lexical Class(ification). The tendency of words to be grouped by classes, which share semantic and syntactic behaviour.
CoNLL	Conference on Computational Natural Language Learning. The shared tasks for this annual NLP conference use a format for dependency parser output which has become a standard in the field; thus some of the tools used in PANACEA produce or accept CoNLL format. See http://conll.cemantix.org/2012/data.html.
LMF	Lexical Markup Framework (Francopoulo et al. 2008)
ARFF	Attribute-Relation File Format (ARFF, http://weka.wikispaces.com/ARFF), which is the file format used by Weka (Witten and Frank, 2005)

This document is organized as follows: first of all a summary of the research undertaken for each lexical acquisition task is presented with the explanation of the corresponding developed tools and the corresponding scientific papers. Then, we specify how those tools have been included in the PANACEA platform: the list of deployed web services, the common interfaces and the travelling objects defined for acquired lexica. Finally, we present the developed workflows and the list of deliveries associated to this work package.

2 Summary

Here we give a brief summary of the results from WP6, and an overview of the remainder of this deliverable.

This deliverable begins by recounting the research undertaken as part of WP6. Work on Subcategorization Frames included “inductive” approaches, in which frames are learned from the corpus data (English, Spanish, Italian); merging of lexicons by parser combination (English); SCF acquisition in the biomedical domain (English); unsupervised SCF acquisition using tensor factorization (English); unsupervised SCF acquisition using graphical models (English); and a practical application of SCF lexica for improving parsing accuracy (English). The SCF research resulted in three web services (in addition to parser web services), five conference/workshop papers, two journal papers, and one MPhil dissertation. The accuracy of the resulting lexica (as for all the lexical acquisition components) is described in D7.4.

Work on Selectional Preferences included unsupervised SP acquisition using tensor factorization (English, Italian), and an approach making use of a lexical hierarchy. The SP research resulted in two conference papers.

Work on Multiword Expressions included approaches to dealing with noisy data (Italian) and an evaluation of three methods including correspondence asymmetries between different versions of Wikipedia, translation of WordNet, and lexical association measures (Arabic). The MWE research resulted in one web service and three conference papers.

Work on Lexical-semantic Classes included nominal semantic class identification using decision trees and Bayesian models (Spanish, English); nominal semantic class identification using Bayesian inference (Spanish); identification of deverbal event nouns (Italian); identification of verbal semantic classes using hierarchical verb clustering (English, French); identification of adjective classes in a novel task (English); a review paper; and a practical application of LC lexica for metaphor processing (English). The LC research resulted in 15 web services, six conference papers, three journal papers, and an MPhil dissertation.

After describing the research, this deliverable provides a consolidated list of the scientific articles and workshops which were presented in the different sections. It lists the web services developed for WP6, again consolidated from the different sections. It describes the Common Interfaces adopted to ensure interoperability among lexical acquisition components, and it presents an extensive definition of the Travelling Object format for acquired lexica, with examples. Finally, this deliverable describes the lexical acquisition workflows produced for WP6, and describes the deliveries of monolingual lexica.

3 Development of Tools and Technologies

In this section we present the research conducted to develop the different tolos fot lexical acquisition.

3.1 Subcategorization Frames (WP6.1)

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. 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 Schulte im Walde (to appear) for an overview.

Some recent work has applied SCF techniques to various languages including Ienco et al. (2008) for Italian, Chrupala (2003) and Esteve Ferrer (2004) for Spanish, Messiant (2008) for French.

3.1.1 Approaches

For PANACEA it was decided that the majority of SCF acquisition would focus on "inductive classifiers" as exemplified by (Messiant, 2008). Such classifiers are relatively domain-independent because they take parsed data as input and learn SCFs based on observed verbal argument patterns, without any preconceived inventory of SCFs. They are also lightweight and suitable for large scale web service provision. The goal here was not necessarily to improve on state-of-the-art accuracy, but to make the tools available in a lexical acquisition platform.

Two different inductive classifiers have been developed, using slightly different methods for deciding on the features to be included in the SCFs. The initial versions of these inductive classifiers were developed and tested on two different languages, Italian and English, respectively, but the newest versions implemented generalized, language- and tagset-independent SCF classifiers that can use the native tagset of the parser output to learn SCFs. These classifiers adopt the same acquisition methodology as the language-specific ones, but rely on very general extraction rules, with a customizable interface. These classifiers have been deployed as web services and integrated in the PANACEA platform. Using these classifiers, SCF lexica for three languages (Italian, English, and Spanish) have been acquired.

In addition to inductive classifiers, several other approaches have been pursued within PANACEA. The remainder of this section details different approaches to SCF acquisition which have been investigated in PANACEA. When web services have been deployed, they are described under each approach. A full summary table of web services is collected in Section 3. The major papers associated with each approach are summarized here. A full list of papers is collected in Section 2.7, and the papers themselves are available in Annex A.

3.1.2 Inductive SCF Acquisition

This section describes the inductive classifiers developed for PANACEA.

3.1.2.1 Web services

Three web services for inductive SCF extraction have been deployed.CNR has deployed two web services for SCF extraction, one for Italian and two language independent.

3.1.2.1.1 estrattore_scf_it (CNR)

The IT-SCF Extractor (Extractor henceforth) takes as input dependency parsed data in the CoNLL format and is composed of three core modules: a) a pattern extractor which identifies possible SCF patterns for each verb; b) a SCF builder, which assigns a list of candidate SCFs to each verb and, finally; c) a filter which removes SCFs that are considered incorrect. The raw data is morpho-syntactically analyzed through the FreeLing suite for Italian (Padro et al., 2010) and then parsed by the DeSR parser (Attardi) through an input/outputformat converter. The DeSR parser is one of the most accurate dependency parsers for Italian (it achieved first position in both Dependency Parsing tasks at Evalita 2009). The parser builds dependency structures and chooses at each step whether to perform a shift or to create a dependency between two adjacent tokens. The dependency annotation schema is based on the ISST syntactic-functional annotation schema and does not fully distinguish between core arguments and adjuncts.

Module 1: The Pattern Extractor

The pattern extractor (PE) collects the dependencies found by the parser for each occurrence of a (target) verb. Some cases receive a special treatment, namely:

- - the reflexive pronouns si, mi, ti, ci and vi are always extracted when they have the relations "obj" (direct object), "clit" (clitic), "comp-ind" (indirect object) and "arg" with a verb. Their presence does not give rise to a different verb entry, i.e. reflexives are not considered as separate verb entries;

- - modifiers realized by adverbs, gerunds and past participles which normally are not part of an SCF of a verb are extracted and stored in a dedicated slot within the verb SCF;

- - when a preposition is a dependent of the verb, the pattern extractor explores its dependent to discover the PoS which follows it (either a NP or a verbal clause in the infinitive form);

- - the extraction is interrupted after a maximum of four dependent elements or when a complement clause is identified.

Module 2: The SCF Builder

The SCF builder stores the information provided by the pattern extractor as lists of eligible SCF for each verb entry. Each extracted SCF is then ordered alphabetically according to the syntactic constituents involved in order to have a normalized form of the SCF for the evaluation of the Extractor (i.e. the position of the arguments relative to the verb is not distincitive). Nevertheless, each occurrence of an SCF (including its frequency) is stored in a dedicated cache (SCF variants). In the lexicon, the variant with the highest frequency will be promoted as the canonical SCF form. To clarify this, let"s consider the examples 1. and 2. (notice that the SCF builder output is partial, i.e. auxiliary information is not reported). The dollar symbol ($) in front of each syntactic constituent is a device to facilitate the identification of SCFs.

1. Hanno accusato Giovanni di furto. "They accused Giovanni of theft"

Pattern Extractor Output: $OBJGiovanni $COMPDIdifurto

SCF Builder: ACCUSARE $COMP-DI_$OBJ SCF FREQ=1 V-SCF FREQ=1

SCF Variants: ACCUSARE $OBJ_ $COMP-DI

FREQ=1

2. Hanno accusato di furto Giovanni. "They accused of theft Giovanni"

Pattern Extractor Output: $COMP-DIdif urto $OBJGiovanni

SCF Builder: ACCUSARE $COMP-DI_$OBJ SCF FREQ=2 V-SCF FREQ=2

SCF Variants : ACCUSARE $COMP-DI_$OBJ

FREQ=1

Due to language specific issues, i.e. the fact that Italian is a pro-drop language, and to the fact that subjects are external verb arguments, they have not been extracted at this stage of development.

Module 3: The Filter

Apart from processing errors, the output of the Extractor is noisy due to the task itself, i.e. the acquisition of verb SCFs. The most debatable issue in this task is related to the argument - adjunct distinction. Following Messiant et al. (2008), we assume that arguments tends to occur in argument position more frequently than adjuncts. Thus, frequent SCFs are assumed to be correct. The identification of these items, i.e. filtering, is accomplished in two steps by means of empirical measures based on the maximum likelihood estimate (MLE) (Korhonen et al., 2006). In this context, MLE barely corresponds to the relative frequency of the V-SCF couple. To compute Figure 1: SCF acquisition service in the Taverna workflow editor MLE we apply the formula used by Messiant et al. (2008). Where corresponds to the frequency of with the verb Vi, i.e. the V-SCF couple, and |Vi| corresponds to the overall frequency of the verb Vi. According to a given MLE threshold, whatever is below the empirical threshold will be rejected as probably incorrect. In addition to this first filter, we introduce a further MLE filter, which we will call percentage on verb frequency, (PVF) for clarity's sake. Thus, for every VSCF couple which is below the initial MLE threshold, the system reduces the length of the syntactic dependents of the SCF by taking into account all the possible combinations. Once a newly created V-SCF couple is found that already exists, then it re-assigns the associated frequency to the existing V-SCF with the highest frequency. If the updated V-SCF are above the PVF, then they are accepted, otherwise the SCF length reduction process is restarted until the V-SCF couple is above the PVF ratio. For instance, in case we have a V-SCF couple of this kind Vx - $SCF1 $SCF2, the system splits the couple in Vx - $SCF1 and Vx -$SCF2 and assign both the frequency of the old V-SCF couple. If at least one of the newly proposed couple already exists, its assigns the frequency to the already existing frame and computes the PVF ratio. Otherwise, a new reduction process is performed until the frame is assigned. Both the MLE and the PVF thresholds can be set by the user (they are passed as a parameter to the service), in order to allow for various types of output accuracy, depending on the specific uses the extracted data is intended for. The higher the threshold, the higher the accuracy, but obviously the lower the number of retrieved Verb-SCF pairs. In our experiments, we established that MLE >=0.008 and PVF = 2.5% are the best filters for reaching a good balance between precision and recall.

Parameters required

Input files in CoNNL format

List of verbs (optional)

3.1.2.1.2 estrattore_scf_lang_indip (CNR)

The language independent version of the SCF extractor works on a CoNNL like parsed input and processes it along the lines that are defined works along the same steps that have been described; the differences are that the PoS for verbs can be set and that no language dependent special rules are in place during the extraction. The extractor thus locates all instances of verbs and retrieves all their relations. A list of verbs can still be passed by the user.

Parameters required

Input files in CoNNL format

Value of the PoS for the Verb in the input file

List of verbs (optional)

3.1.2.1.3 tcp_subcat_inductive (UCAM)

Until recently, state of the art SCF acquisition systems used handcrafted rules to match natural language parser output to a set of pre-defined SCFs (Briscoe and Carroll, 1997; Korhonen, 2002; Preiss et al., 2007). Such approaches achieved F-measures of about 70 against a manually annotated gold standard. Recently, however, it has become more common to use an 'inductive' approach, in which the inventory of SCFs is induced directly from parsed corpus data (O'Donovan et al., 2005; Cesley and Salmon-Alt, 2006; Ienco et al., 2008; Messiant, 2008; Lenci et al., 2008; Altamirano and Alonso I Alemany, 2010; Kawahara and Kurohashi, 2010). Candidate frames are identified by grammatical relation (GR) co-occurrences, sometimes aided by language-specific heuristics. Statistical filtering or empirically-tuned thresholds are used to select frames for the final lexicon. These inductive approaches have achieved respectable accuracy (60-70 F-measure against a dictionary) and are more portable than earlier methods. They are suitable for languages and domains in which no pre-defined inventory of SCFs is available, as long as a parser is available. They are also highly scalable to large volumes of data, since the identification and selection of frames for the lexicon generally takes minimal time and resources compared to step of parsing the data.

Due to the portability and scalability of inductive SCF acquisition, it was decided that the PANACEA web services would use this approach. UCAM developed and deployed the web service tpc_subcat_inductive, which was used for English and Spanish SCF acquisition. Here we describe the operation and parameters of tpc_subcat_inductive.

The input to the web service is the output of a parser. The web service can accept parser output in one of two formats: either the output format of the RASP parser, or the CoNLL format [insert references]. These formats were chosen based on the parsers in use at UCAM and UPF, but adding input formats to the web service is quite straightforward, requiring only a definition of the GR format used by the parser.

The web service user can define a set of verbs of interest. These are the verbs whose SCFs will appear in the lexicon. After reading the input file, the web service proceeds by identifying parsed sentences containing the target verb lemmas. For each verb lemma, the set of co-occurring GRs is tallied, and relative frequencies calculated. For instance, if the verb lemma *consider* appears eight times with a dobj (direct object) GR and two times with both a dobj and xcomp (which includes adjectives and non-finite clauses), then the resulting lexicon will show a relative frequency of 0.8 for the frame DOBJ, and a relative frequency of 0.2 for DOBJ-XCOMP.

The GR types of interest can also be defined by the web service user. In this way the user can decide which GR types are likely to be arguments of the verb, and hence part of the subcategorization frame, and which are likely to be modifiers, or adjuncts, and should not be considered. However, the user does not define specific frames, as in earlier SCF acquisition work. Rather, if the user specifies DOBJ and XCOMP as GR types of interest, but not MODIFIER, then the SCF inventory will consist of all observed combinations of DOBJ and XCOMP, and MODIFIER will never appear in the SCF. Thus a minimal amount of linguistic expertise is required to set the parameters defining the resulting SCF inventory.

A few additional parameters allow refinement of the GR types; these are described in the table below. The more parameters used, the more fine-grained the resulting SCF inventory, and the more SCFs will be detected by the system. Increasing the level of granularity in the inductive SCF inventory, while it allows the acquisition of more detailed and fine-grained SCFs, tends to have the result of decreasing measured accuracy, since it is easier for the system to mistake one SCF for another. At present there is no evidence that a very fine-grained inventory is more useful to a downstream application than a relatively coarse-grained one.

The tpc_subcat_inductive web service also performs maximum likelihood filtering, using a uniform threshold (which is set as a parameter) and filtering out all SCFs that are attested at a relative frequency below this threshold for any given verb. Adjusting this threshold will affect the precision-recall tradeoff.

The full set of parameters used by tpc_subcat_inductive is as follows:

Parameter Name	Description
Target Verb	List of verbs to include in the lexicon
Threshold	Minimum relative frequency for filtering the lexicon (discard all SCFs with a per-verb relative frequency below this threshold)
Parser Format	RASP or CoNLL
Target GR Types	The set of GR types from which to inductively build the inventory of SCFs. User should specify GR types which are typically arguments rather than adjuncts.
Ignore Instance GR Types	Ignore any verb instances where the verb participates in this GR type. For instance, the user can choose to ignore sentences where the verb is the head of a 'passive' GR.
POS Groups	A way of generalizing over POS tags. For example, if the user cares whether the dependent of a GR is a noun, but not which type of noun, they can create a Noun group including NN1, NN2, etc. Pronouns can also be grouped with nouns, for example. Or POS tags with verbs representing different tenses may be grouped together. Any POS tags not falling into one of the POS Groups will be output as "OTH" (other) within the SCF.
GR Types to Dep POS	For these GR types, consider the POS group of the dependent as part of the SCF.
GR Types to Child	For these GR types, consider the child of the dependent as part of

	the SCF (where the user must specify which types of children are of interest).
GR Types to Lex	For these GR types, lexicalize the dependent. Typically used for prepositions and particles.

3.1.2.2 Related Papers

The SCF system for Italian has been presented as a paper at LREC 2012.

- Caselli, Tommaso; Rubino, Francesco; Frontini, Francesca; Russo, Irene; Quochi, Valeria. (2012). **Flexible Acquisition of Verb Subcategorization Frames in Italian.** In Proceedings of LREC 2012, Istanbul, Turkey.

This paper describes a system for the automatic (unsupervised) acquisition of verbal SCFs in Italian, to be integrated in a distributed platform for the automatic creation of Language Resources. The methodology used is similar to those described in Messiant et al. (2008) and Lenci et al. (2008). The system is completely unsupervised, in the sense that it does not assume any pre-defined list of SCFs, but learns them from data instead. One of the most interesting features of this work is the possibility the final users have to customize the results of the SCF extractor and obtaining different SCF lexica in terms of size and accuracy. The tool is made available as a web service through the PANACEA Platform.

The IT-SCF Extractor (Extractor henceforth) takes as input dependency parsed data in the CoNLL format and is composed of three core modules: a) a pattern extractor which identifies possible SCF patterns for each verb; b) a SCF builder, which assigns a list of candidate SCFs to each verb and, finally; c) a filter which removes SCFs that are considered incorrect.

The raw data is morpho-syntactically analyzed through the FreeLing suite for italian (Padro et al., 2010) and then parsed by the DeSR parser (Attardi and Ciaramita 2007; Attardi and Dell'Orletta 2009), through an input-output format converter. The DeSR parser is one of the most accurate dependency parsers for Italian (it achieved first position in both Dependency Parsing tasks at Evalita 2009.)

Multiple evaluations were performed, and confidence levels investigated; see D7.4 for a detailed description of the evaluation.

3.1.2.3 Related Experiments

Experiments with English and Spanish inductive SCF acquisition have not yet been published as scientific papers. The experiments are described here and the evaluations are described in D7.4.

3.1.2.3.1 Details on using tpc_subcat_inductive webservice for Spanish

In order to perform the SCF acquisition for Spanish, UPF has used the language independent web service developed in UCAM. This web service takes a parsed text and extracts statistics about the occurrences of verbs with a number of particular complements. The type of complements is previously stated by tuning a set of parameters. Here we present the parameters used to develop the Spanish SCF lexicon.

The parameters required by the SCF extractor include information about the output of the parser (PoS tag set, label of the complements) and on which kind of complements we want the SCF extractor to capture (e.g. direct object, indirect object, etc).

In order to extract SCF from corpus, first of all we need to create a parsed corpus. For Spanish, we used the Spanish Malt parser PANACEA web service. This instance of Malt, uses FreeLing PoS tags, this is, the EAGLES tag set. Regarding complement labels, it uses 27 function labels, from which 12 are selected for the SCF extractor (those related to syntactic functions other than modifiers) as we will see below. For more information about the Spanish parser output, see the Spanish Malt parser PANACEA web service documentation[1].

Relevant complements and PoS tags for Spanish SCFs

In order to extract SCFs for Spanish, we needed to define which complement labels and PoS tags it was necessary to extract. We selected this information according to the information encoded in the gold-standard[2] (Necsulescu et al., 2011) which encodes verb frames by position in a list and information about the syntactic category of the possible fillers: whether they can be a noun phrase, a clause, a prepositional phrase, etc.

From the parsed corpus, we needed to identify the syntactic functions we wanted to extract and some information about the PoS tag in order to know which kind of clause realizes the complement. In the following table we present the information present in the gold-standard and how we can obtain it from a parsed corpus.

Complement type in gold-standard	**Corresponding syntactic function in parsed corpus**	**PoS of the complement**	**Other information to be extracted**
np	Direct Object (DO) or Subject (SUBJ)	noun, pronoun	
cp		verb in infinitive or relative pronoun (that introduces a relative clause)	
ppa	Indirect Object (IO)		
pp	Oblic complement (OBLC): the complement has a bounded preposition		Child of the preposition to determine the realisation of the pp (np, cp, etc) Lemma of the preposition to get the bounded preposition
	Loc or Dir complement (PP-LOC, PP-DIR)		Child of the preposition to determine the realisation of the pp (np, cp, etc)
	Predicative complements (PRD or OPRD)	preposition	

[1] http://registry.elda.org/services/249

[2] http://panacea-lr.eu/en/info-for-researchers/gold-standards/subcategorization-frames/spanish-scf-gold-standard

adj	Predicative complements (PRD or OPRD)	adjective or past participle verb	
adv	Adverbial complement (ADV)		

This information is encoded in the parameters of the web service “tcp_subcat_inductive”, that allows us to define which complements, PoS, and children need to be extracted. Here we give a general description of what needs to be encoded in the parameters. For the concrete values of those parameters, consult one of the developed workflows for Spanish SCF extraction (e.g. http://myexperiment.elda.org/workflows/80).

Parameters:

- target_gr_types: list of functions (as given by the parser) to be extracted: SUBJ, DO, IO, OBLC, PRD, OPRD, ADV, PP-LOC, PP-DIR
- pos_groups: define the PoS groups that will be used in the output. The following list contains the PoS groups used in UPF experiments:
  - np: nouns and pronouns
  - adj: adjectives
  - v: inflected verbs
  - vp: past participle verbs
  - vg: gerund verbs
  - vn: infinitive verbs
  - s: prepositions
  - c: conjunctions
  - rg: adverbs
- gr_types_to_child (output the children of these complements in order to study their content): OBLC, PRD, OPRD, PP-LOC, PP-DIR
- gr_types_to_deppos (add the PoS for these complements): DO, PRD, OPRD, SUBJ, COMP, OBLC, PP-LOC, PP-DIR
- gr_types_to_lex (add the lemma for these complements): OBLC, PRD, OPRD

With these parameters, the output of the web service would be of the kind (some examples):

- DO_np:SUBJ_np: SCF with a DO and a subject, both of them realized by an np.
- OBLC_s-de=>COMP_vn:SUBJ_np: SCF with a bounded-prepositional complement with preposition “de” realized by a verb, and a subject np.

3.1.2.3.2 Details on using tpc_subcat_inductive webservice for English

We used the tpc_subcat_inductive web service to produce an English SCF lexicon for each domain containing SCFs for the 28 or 29 verbs in each gold standard. Table 1 below reports the parameter setting for this experiment.

Parameter Name	**Setting**
Target Verb	Set as appropriate for the domain.
Threshold	Tested values from 0 through 0.04.
Parser Format	RASP.
Target GR Types	Direct object (dobj), prepositional object (iobj), second object of ditransitive (obj2), finite clausal complement without complementizer (ccomp_), finite clausal complement with "that" complementizer (ccompthat), non-finite clausal complement without complementizer (xcomp_), non-finite clausal complement with "to" (ccompto), prepositional complement (pcomp), particle (ncmodprt), finite clausal subject without complementizer (csubj_), finite clausal subject with "that" complementizer (csubjthat), non-finite clausal complement (xsubj). All modifier types are excluded.
Ignore Instance GR Types	Passive.
POS Groups	Groups are created for: Noun (N), Verb (V), Bare Verb (VBARE), Tensed Verb (VTENSED), Present Participle Verb (VING), Past Participle Verb (VEN), Wh-phrase (WH), Wh-complement (WHCOMP), Wh-adverb (WHADV), Adjective (ADJ), Adverb (ADV), Preposition (PREP)
GR Types to Dep POS	The GR types dobj, obj2, ccomp_, ccompthat, xcomp_, xcompto all have POS groups specified as part of the SCF. Specifically, these are: {"dobj":["N","WH","WHCOMP","WHADV"],"obj2" ["N","WH"],"ccomp_":["VBARE","VING","VTENSED", "VEN","WHCOMP","WHADV","I"],"ccompthat":["VBARE","VING","VTENSED","VEN","WHCOMP","WHADV","I"],"xcomp_" ["VBARE","VING","VTENSED","VEN","WHCOMP"," WHADV","ADJ","I"],"xcompto":["VBARE"],}
GR Types to Child	Null, for coarse-grained SCF inventory.
GR Types to Lex	Null, for coarse-grained SCF inventory.

Table 1: ***Parameter setting of the English inductive SCF acquisition***

We examined several filtering thresholds to determine the precision-recall tradeoff.

We then took the additional step of removing from the lexicon any SCFs containing an OTH part of speech tag. These SCFs typically represent parser errors, since they contain words with POS tags that are not considered likely parts of the SCF as defined in the GR Types to Dep POS parameter. In preliminary experiments we found that this resulting in much greater accuracy. It does also result in losing some correct examples, however, since e.g. coordinations and other structures may result in POS tags identified as OTH despite being legitimate.

3.1.3 SCF Induction with Parser Combination

In D6.1 a goal was set out of improving the hypothesis selection step in SCF acquisition by using a parser ensemble (Sagae and Lavie, 2006; Miyao et al., 2008). A version of the SCF classifier of Preiss et al. (2007) has been developed, using the output of the unlexicalized Stanford parser (Klein and Manning, 2003) whereas the original classifier accepted the RASP parser output format. A method has been implemented which selects only those SCFs agreed on by the RASP and Stanford parsers, and it was found that the parser ensemble method served as a filter on SCF hypotheses.

3.1.3.1 Related Papers

This SCF parser combination work was presented at the LREC Workshop on Language Resource Merging.

- Rimell, Laura; Poibeau, Thierry and Korhonen, Anna. (2012). **Merging Lexicons for Higher Precision Subcategorization Frame Acquisition**. In Proceedings of the LREC Workshop on Language Resource Merging, Istanbul, Turkey.

A number of filtering and smoothing techniques have been proposed in order to improve the precision of automatically acquired SCF lexicons. Filtering SCFs which are attested below a relative frequency threshold for any given verb, where the threshold is applied uniformly across the whole lexicon, has been shown to be effective (Korhonen, 2002; Messiant et al., 2008). However, this technique relies on empirical tuning of the threshold, necessitating a gold standard in the appropriate textual domain, and it is insensitive to the fact that some SCFs are inherently rare. The most successful methods of increasing accuracy in SCF lexicons rely on language- and domain-specific dictionaries to provide back-off distributions for smoothing (Korhonen, 2002). This paper presents a different approach to acquiring a higher precision SCF resource, namely the merging of two automatically acquired resources by retaining only the information that the two resources agree on. Previous work on language resource merging has generally focused on increasing coverage by adding information from one resource to another, e.g. (Crouch and King, 2005; Molinero et al., 2009), which focus on merging multiple levels of information from disparate resources. More closely related to our work, (Necsulescu et al., 2011; Bel et al., 2011; Padro´ et al., 2011) merge two manually built SCF lexicons, unifying SCF s when possible but with the goal of retaining informa-tion from both lexicons. Treating language resource merger as (roughly) a union operation is appropriate for manually developed resources, or when coverage is a priority. How-ever, when working with automatically acquired resources it may be worthwhile to adopt the approach of merger by intersection.

We focus here on the fact that the tagger and parser are one source of noise in automatic SCF acquisition, and combine two lexicons built with different parsers. This approach is similar in spirit to parser ensembles, which have been used successfully to improve parsing accuracy (Sagae and Lavie, 2006; Sagae and Tsujii, 2007). We build two SCF lexicons using the framework of (Korhonen, 2002; Preiss et al., 2007), which was designed to classify the output of the RASP parser (Briscoe et al., 2006), and which we extend to classify the output of the unlexicalized Stanford parser (Klein and Manning, 2003). We then build a combined lexicon that includes only SCFs that are agreed on by both parsers. Using this simple combination approach, we obtain a lexicon with higher precision than the lexicon built with either parser alone.

See D7.4 for the evaluation of the parser combination against a general language SCF gold standard.

3.1.4 SCF Induction in the Biomedical Domain

Exploration has been undertaken of SCF acquisition in the biomedical domain. Although not one of the PANACEA project domains, these experiments resulted in a better understanding of SCF domain and subdomain variation.

3.1.4.1 Related Papers

Exploration of the lexical characteristics of the biomedical domain can be found in a COLING 2010 paper.

- Korhonen, Anna; Lippincott, Tom; ó Séaghdha, Diarmuid; Sun, Lin. (2010). E**xploring variation across biomedical subdomains** . In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.

One of the most noticeable trends in the past decade of NLP research has been the deployment of language processing technology to meet the information retrieval and extraction needs of scientists in other disciplines. This meeting of fields has proven mutually beneficial: scientists increasingly rely on automated tools to help them cope with the exponentially expanding body of publications in their field, while NLP researchers have been spurred to address new conceptual problems in theirs. Among the fundamental advances from the NLP perspective has been the realisation that tools which perform well on textual data from one source may fail to do so on another unless they are tailored to the new source in some way. This has led to significant interest in the idea of contrasting *domains* and the concomitant problem of *domain adaptation*, as well as the production of manualloy annotation domain-specific corpora.

One definition of *domain variation* associates it with differences in the underlying probability distributions from which different sets of data are drawn (Daume III and Marcu, 2006). The concept also mirrors the notion of variation across thematic subjects and the corpus-linguistic notions of *register* and *genre* (Biber, 1988). In addition to the differences in vocabulary that one would expect to observe, domains can vary in many linguistic variables that affect NLP systems. The scientific domain which has received the most attention (and is the focus of this paper) is the biomedical domain. Notable examples of corpus construction projects for the biomedical domain are PennBioIE (Kulick et al., 2004) and GENIA (Kim et al., 2003). These corpora have been used to develop systems for a range of processing tasks, from entity recognition (Jin et al., 2006) to parsing (Hara et al., 2005) to coreference resolution (Nguyen and Kim, 2008).

An implicit assumption in much previous work on biomedical NLP has been that particular subdomains of biomedical literature – typically molecular biology – can be used as a model of biomedical language in general. For example, GENIA consists of abstracts dealing with a specific set of subjects in molecular biology, while PennBioIE covers abstracts in two specialised domains, cancer genomics and the behaviour of a particular class of enzymes. This assumption of representativeness is understandable because linguistic annotation is labour-intensive and it may not be worthwhile to produce annotated corpora for multiple subdomains within a single discipline if there is little task-relevant variation across these subdomains. However, such conclusions should not be made before studying the actual degree of difference between the subdomains of interest.

One of the principal goals of this paper is to map how the concept of "biomedical language", often construed as a monolithic entity, is composed of diverse patterns of behaviour at more fine-grained topical levels. Hence we study linguistic variatio in a broad biomedical corpus of abstractsand full papers, the PMC Open Access Subset. We select a range of lexical and structural phenomena for quantitative investigation. The results indicate that common subdomains for resource development are not representative of biomedical text in general and furthermore that different linguistic features often partition the subdomains in quite different ways.

Two journal articles related to SCF acquisition in the biomedical domain (joint with University of Colorado and National ICT Australia) have been submitted to the Journal of Biomedical Informatics. The two articles collectively include an exploration of subdomain variation, a new biomedical SCF gold standard, and an investigation of different definitions of subcatgorization used in biomedicine, where adjuncts are typically retained in SCFs. The submission also introduced a novel method of SCF-specific filtering. The first article, a methodological review, is forthcoming in the journal. The second article is under revision.

- Lippincott, Thomas; Rimell, Laura; Verspoor, Karin; Korhonen, Anna. **Approaches to Verb Subcategorization for Biomedicine**. Forthcoming, Journal of Biomedical Informatics.

Information about verb subcategorization frames (SCFs) is important to many tasks in natural language processing (NLP) and, in turn, text mining. Biomedicine has a need for high-quality SCF lexicons to support the extrac-tion of information from the biomedical literature, which helps biologists to take advantage of the latest biomedical knowledge despite the overwhelming growth of that literature. Unfortunately, techniques for creating such resources for biomedical text are relatively undeveloped compared to general language. This paper serves as an introduction to subcategorization and existing approaches to acquisition, and provides motivation for devel-oping techniques that address issues particularly important to biomedical

NLP. First, we give the traditional linguistic definition of subcategorization, along with several related concepts. Second, we describe approaches to learn-ing SCF lexicons from large data sets for general and biomedical domains. Third, we consider the crucial issue of linguistic variation between biomedi-cal fields (subdomain variation). We demonstrate significant variation among subdomains, and find the variation does not simply follow patterns of general lexical variation. Finally, we note several requirements for future research in biomedical SCF lexicon acquisition: a high-quality gold standard, investiga-tion of different definitions of subcategorization, and minimally-supervised methods that can learn subdomain-specific lexical usage without the need for extensive manual work

- Rimell, Laura; Lippincott, Thomas; Verspoor, Karin; Johnson, Helen L.; Korhonen, Anna. **Acquisition and Evaluation of Verb Subcategorization Resources for Biomedicine**. Under revision, Journal of Biomedical Informatics.

Biomedical natural language processing (NLP) applications that have access to detailed resources about the linguistic characteristics of biomedical language demonstrate improved performance on tasks such as relation extraction and syntactic or semantic parsing. Such applications are important for transforming the growing unstructured information buried in the biomedical literature into structured, actionable information. In this paper, we address the creation of linguistic resources that capture how individual biomedical verbs behave. We specifically consider verb subcategorization, or the tendency of verbs to “select” co-occurrence with particular phrase types, which influences the interpretation of verbs and identification of verbal arguments in context. There are currently a limited number of biomedical resources containing information about subcategorization frames (SCFs), and these are the result of either labor-intensive manual collation, or automatic methods that use tools adapted to a single biomedical subdomain. Either method may result in resources that lack coverage. Moreover, the quality of existing verb SCF resources for biomedicine is unknown, due to a lack of available gold standards for evaluation.

This paper presents three new resources related to verb subcategorization frames in biomedicine, and four experiments making use of the new resources. We present the first biomedical SCF gold standards, capturing two different but widely-used definitions of subcategorization, and a new SCF lexicon, BioCat, covering a large number of biomedical sub-domains. We evaluate the SCF acquisition methodologies for BioCat with respect to the gold standards, and compare the results with the accuracy of the only previously existing automatically-acquired SCF lexicon for biomedicine, the BioLexicon. Our results show that the BioLexicon has greater precision while BioCat has better coverage of SCFs. Finally, we explore the definition of subcategorization using these resources and its implications for biomedical NLP. All resources are made publicly available.

3.1.5 Unsupervised SCF Induction: Tensor Factorization

Two additional novel approaches to SCF acquisition have been pursued, both unsupervised methods that address hypothesis generation and selection. First, non-negative tensor factorization (NTF) has been used) to learn SCFs and SPs jointly. The method takes parser output and uses NTF, a dimensionality reduction technique, to find clusters of verbs with similar syntactic and semantic behavior. The accuracy is respectable for an unsupervised method.

3.1.5.1 Related Papers

This work was presented at COLING 2012.

- Van de Cruys, Tim; Rimell, Laura; Poibeau, Thierry; Korhonen, Anna (2012). **Multi-way Tensor Factorization for Unsupervised Lexical Acquisition**. In Proceedings of COLING, Mumbia, India.

This paper introduces a novel method for joint unsupervised aquisition of verb subcategorization frame (SCF) and selectional preference (SP) information. Treating SCF and SP induction as a multi-way co-occurrence problem, we use multi-way tensor factorization to cluster frequent verbs from a large corpus according to their syntactic and semantic behaviour. The method extends previous tensor factorization approaches by predicting whether a syntactic argument is likely to occur with a verb lemma (SCF) as well as which lexical items are likely to occur in the argument slot (SP), and integrates a variety of lexical and syntactic features, including co-occurrence information on grammatical relations not explicitly represented in the SCFs.

Our method uses a co-occurrence model augmented with a factorization algorithm to cluster verbs from a large corpus. Specifically, we use non-negative tensor factorization (NTF) (Shashua and Hazan, 2005), a generalization of matrix factorization that enables us to capture latent structure from multi-way co-occurrence frequencies. The factors that emerge represent clusters of verbs that share similar syntactic and semantic behaviour. To evaluate the performance on SCF acquisition, we identify the syntactic behaviour of each cluster. The SCF lexicon that emerges from the clusters achieves a promising F-score of 68.7 against a gold standard. We further introduce a novel SP evaluation in which we investigate the model's ability to induce preferences for the co-occurrence of a particular verb lemma and all of its arguments at the same time. The model achieves a high accuracy of 77.8 on this new evaluation. We also perform a qualitative evaluation which shows that the joint model is capable of learning rich lexical information about both syntactic and semantic aspects of verb behaviour in data.

See D7.4 for an evaluation of both SCF and SP acquisition against general language gold standards.

3.1.6 Unsupervised SCF Induction: Tensor Factorization

The second unsupervised method of SCF acquisition involved graphical models. The method takes either parsed or POS-tagged data and models subcategorization using a Bayesian network. The method is evaluated against a general language gold standard and in a verb clustering task, achieving state-of-the-art accuracy.

3.1.6.1 Related Papers

This work was presented at ACL 2012.

- Lippincott, Thomas; Korhonen, Anna and Ó Séaghdha, Diarmuid. (2012). **Learning Syntactic Verb Frames Using Graphical models**. In Proceedings of ACL, Jeju Island, Korea.

High quality SCF lexicons are difficult to build automatically. The argument-adjunct distinction is challenging even for humans, many SCFs have no reliable cues in data, and some SCFs (e.g. those involving control such as type raising) rely on semantic distinctions. As SCFs follow a Zipfian distribution (Korhonen et al., 2000), many genuine frames are also low in frequency. State-of-the-art methods for building data-driven SCF lexicons typically rely on parsed input (see section 2). However, the treebanks necessary for training a high-accuracy parsing model are expensive to build for new domains. Moreover, while parsing may aid the detection of some frames, many experiments have also reported SCF errors due to noise from parsing (Korhonen et al., 2006a; Preiss et al., 2007). Finally, many SCF acquisition methods operate with predefined SCF inventories. This subscribes to a single (often language or domain-specific) interpretation of subcategorization a priori, and ignores the ongoing debate on how this interpretation should be tailored to new domains and applications, such as the more prominent role of adjuncts in information extraction (Cohen and Hunter, 2006).

In this paper, we describe and evaluate a novel probabilistic data-driven method for SCF acquisition aimed at addressing some of the problems with current approaches. In our model, a Bayesian network describes how verbs choose their arguments in terms of a small number of frames, which are represented as distributions over syntactic relationships. First, we show that by allowing the inference process to automatically define a probabilistic SCF inventory, we outperform systems with hand-crafted rules and inventories, using identical syntactic features. Second, by replacing the syntactic features with an approximation based on POS tags, we achieve state-of-the-art performance without relying on error-prone unlexicalized or domain-specific lexicalized parsers. Third, we highlight a key advantage of our method compared to previous approaches: the ease of integrating and performing joint inference of additional syntactic and semantic information. We describe how we plan to exploit this in our future research.

We tested several feature sets either based on, or approximating, the concept of grammatical relation. Our method is agnostic regarding the exact definition of GR, and for example could use the Stanford inventory (De Marneffe et al., 2006) or even an entirely different lexico-syntactic formalism like CCG supertags (Curran et al., 2007). In this paper, we distinguish "true GRs" (tGRs), produced by a parser, and "pseudo GRs" (pGRs), a POS-based approximation, and employ subscripts to further specify the variations described below. Our input has been parsed into Rasp-style tGRs (Briscoe et al., 2006), which facilitates comparison with previous work based on the same data set.

Our graphical modeling approach uses a Bayesian network. Its generative story is as follows: when a verb is instantiated, an SCF is chosen according to a verb-specific multinomial. Then, the number and type of syntactic arguments (Grs) are chosen from two SCF-specific multinomials. These three multinomials are modeled with uniform Dirichlet priors and corresponding hyperparameters. The model is trained via collapsed Gibbs sampling.

3.1.7 Extrinsic Evaluations for SCF Acquisition

In D6.1 UCAM proposed to investigate extrinsic evaluations for SCF acquisition. UCAM has investigated the utility of the VALEX general language subcategorization lexicon to improve the accuracy of a lexicalized parser. This work resulted in an MPhil dissertation supervised by PANACEA members. The results were somewhat complex, showing improvement in parser accuracy on common verbs but less so on rarer verbs, due to interactions with the SCF selection mechanisms of the parser, but with further investigation could lead to a future publication.

3.1.7.1 Related Experiments

This work was part of an unpublished MPhil dissertation.

- Dong, Yizhen. **Using Lexical Resources to Improve Parsing Accuracy**. MPhil Dissertation, 2011, University of Cambridge.

This project uses VALEX, an automatically extracted wide-coverage subcategorization resource in an attempt to improve the C&C parser. The stages are: VALEX subcategorization frames are mapped to CCG categories used in the C&C parser through a common formalism called Grammatical Relations. The mapping scheme is used to convert the VALEX subcategorization lexicon to a tag dictionary which relates words to corresponding CCG categories. Experiments are conducted on combining the generated tag dictionary with the original parser's dictionary under various settings of word frequencies. Performance of the parser i evaluated on CCGbank and a Wikipedia dataset under different experimental settings. Finally, a pilot study investigates enhancing features in the original parsing model by training with artificially generated data.

Mapping is the most fundamental step in this project as it establishes ways how VALEX SCFs can relate to CCG categories. Before actual mapping, an initial investigation of the two formalisms reveals that the mapping is inevitably many to many.

Multiple SCFs are mapped to one category because VALEX and CCG have different approaches for modeling types of argument verbs can take. CCG only models shallow surface syntax while

VALEX has more fine-grained frames modeling underlying syntax like raising and control . For example, *sank* in *His reputation sank low* and *appears* in *He appears crazy* have the same category (S\NP)/(S[adj]\NP) but two different SCFs because the subject is raised in the second sentence. As a result the two SCFs are mapped to one category. The other direction of many-to-many mapping results from sentence features which, for example, can specify a generic (S\NP)/NP category into (S[dcl]\NP)/NP, (S[pt]\NP)/NP, (S[ng]\NP)/NP, (S[b]\NP)/NP and the passive voice category S[pss]\NP. Those categories are all mapped to one SCF for transitive verbs because those categories are related to tense and voice which VALEX does not distinguish.

3.1.8 Spanish Parser

Though the NLP tools such as taggers and parsers, which help to create the input for SCF acquisition, are mostly described under WP4, we want to mention here the development of a Spanish parser, as it was a significant portion of the development of the SCF system. UPF has developed a version of the Malt parser for Spanish, a crucial pre-requisite for SCF acquisition, which has been deployed as a web service. Output from this parser has been used as the input to the language-independent SCF classifier (tpc_subcat_inductive), for a set of experiments, attaining initial results of of around 50 F-score against domain gold standards. A paper using this SCF component will be submitted to ACL 2013.

3.1.8.1 Web services

3.1.8.1.1 malt_parser

This web service calls an instance of Malt parser (http://www.maltparser.org/) for Spanish trained with the Iula treebank (Marimon et al, 2012) developed in the Metanet4you project.

The input of this web service is plain text. The service performs PoS tagging with FreeLing and then performs the dependency parsing using Malt parser. The output follows Conll format.

3.1.9 Gold Standards

CNR, UPF, and UCAM have completed domain-specific SCF gold standards. All of the gold standards are available on the PANACEA web site. The gold standards were used for evaluation of SCF acquisition during WP7.

In D6.1 UCAM planned to measure human annotation time in order to estimate the benefits of automatic lexical acquisition. This was done informally (i.e., an annotator tracking their own work, but not using a formal timing device) and found to be approximately 100 sentences per hour. Therefore if data on e.g. 50 verbs from a new domain was required, with a minimum of 200 sentences per verb, automatic acquisition could save 100 person-hours for each domain.

The SCF gold standards are described in detail in D7.4.

3.2 Selectional Preferences (WP6.1)

Selectional preferences (SPs) describe the semantic restrictions imposed by a predicate on its arguments. The task of learning SPs is similar to the task of judging the plausibility of a predicate and argument occurring together. The 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, so that the likelihood of a verb appearing with an argument can be predicted even when the verb-argument pair is unseen.

As planned in D6.1, SP research in PANACEA was not integrated into the platform, due to the computational requirements of state-of-the-art methods which make them unsuitable for web services.

3.2.1 English Selectional Preferences using NTF

In previous work, SP induction has been limited to the relationship between verbs and direct objects, or, at most, verb-subject-object triples (Van de Cruys, 2010). UCAM has applied tensor factorization methods to model SPs in additional argument slots, including PP (prepositional phrase) arguments and clausal arguments. In this research, SPs were learned jointly with SCFs, in experiment described above under Subcategorization Frames. An evaluation of the method on SPs using pseudo-disambiguation was completed and is included in the COLING 2012 paper. See D7.4 for the evaluation.

3.2.1.1 Related Papers

This work was presented at COLING 2012.

- Van de Cruys, Tim; Rimell, Laura; Poibeau, Thierry; Korhonen, Anna (2012). **Multi-way Tensor Factorization for Unsupervised Lexical Acquisition**. In Proceedings of COLING, Mumbia, India.

(See summary of this paper above.)

3.2.2 Italian Selectional Preferences using NTF

CNR and UCAM have developed a SP model for Italian, where SPs have not previously been very much studied. For the Italian SP model, UCAM tools have been used with ILC-CNR data. A non-negative matrix factorization model has been prepared for Italian data, based on the Repubblica corpus, containing 3183 verbs and 9467 objects, and factored to 300 dimensions.

3.2.3 Selectional Preferences Using a Lexical Hierarchy

UCAM investigated Bayesian selectional preference models that incorporate knowledge from a lexical hierarchy such as WordNet. These approaches are based either on “cutting” the hierarchy at an appropriate level of generalization or on a “walking” model that selects a path from the root to a leaf. The models were evaluated against human plausibility judgements and were shown to improve estimation of plausibility for out-of-vocabulary items.

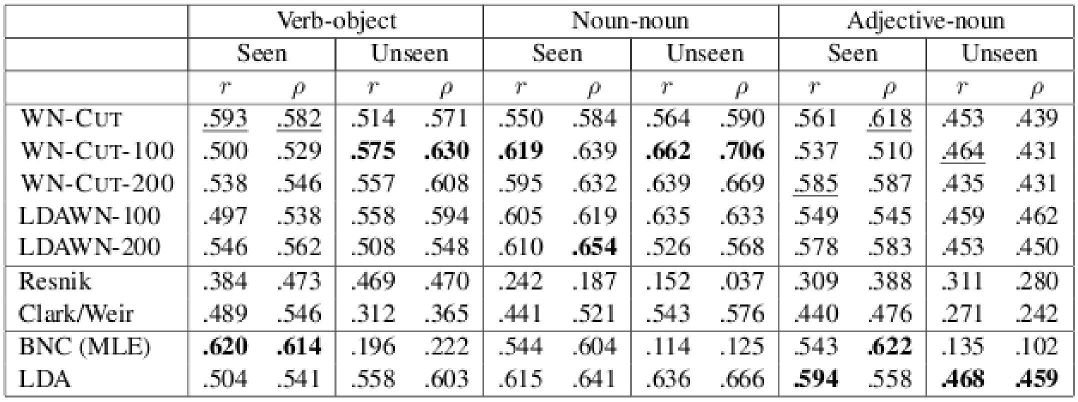
3.2.3.1 Related Papers

This work was presented at *SEM 2012.

- Ó Séaghdha, Diarmuid and Korhonen, Anna. (2012). **Modelling selectional preferences in a lexical hierarchy**. In Proceedings of *SEM, Montreal, Canada.

This paper describes Bayesian selectional preference models that incorporate knowledge from a lexical hierarchy such as WordNet. Inspired by previous work on modelling with WordNet, these approaches are based either on “cutting” the hierarchy at an appropriate level of generalisation or on a “walking” model that selects a path from the root to a leaf. In an evaluation comparing against human plausibility judgements, we show that the models presented here outperform previously proposed comparable WordNet-based models, are competitive with state-of-the-art selectional preference models and are particularly well- suited to estimating plausibility for items that were not seen in training.

Recent research has investigated the potential of Bayesian probabilistic models such as Latent Dirichlet Allocation (LDA) for modelling selectional preferences (O S´eaghdha, 2010; Ritter et al., 2010; Reisinger and Mooney, 2011). These models are flexible and robust, yielding superior performance compared to previous approaches. In this paper we present a preliminary study of analogous models that make use of a lexical hierarchy (in our case the WordNet hierarchy). We describe two broad classes of probabilistic models over WordNet and how they can be implemented in a Bayesian framework. The two main potential advantages of incorporating WordNet information are: (a) improved predictions about rare and out-of-vocabulary arguments; (b) the ability to perform syntactic word sense disambiguation with a principled probabilistic model ad without the need for an additional step that heuristically maps latent variables onto WordNet senses. Focussing here on (a), we demonstrate that our models attain better performance than previously-proposed WordNet-based methods on a plausibility estimation task and are particularly well-studied to estimating plausibility for arguments that were not seen in training and for which LDA cannot make useful predictions.



	Verb-object				Noun-noun				Adjective-noun			
	Seen		Unseen		Seen		Unseen		Seen		Unseen	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
WN-CUT	.593	.582	.514	.571	.550	.584	.564	.590	.561	.618	.453	.439
WN-CUT-100	.500	.529	**.575**	**.630**	**.619**	.639	**.662**	**.706**	.537	.510	.464	.431
WN-CUT-200	.538	.546	.557	.608	.595	.632	.639	.669	.585	.587	.435	.431
LDAWN-100	.497	.538	.558	.594	.605	.619	.635	.633	.549	.545	.459	.462
LDAWN-200	.546	.562	.508	.548	.610	**.654**	.526	.568	.578	.583	.453	.450
Resnik	.384	.473	.469	.470	.242	.187	.152	.037	.309	.388	.311	.280
Clark/Weir	.489	.546	.312	.365	.441	.521	.543	.576	.440	.476	.271	.242
BNC (MLE)	**.620**	**.614**	.196	.222	.544	.604	.114	.125	.543	**.622**	.135	.102
LDA	.504	.541	.558	.603	.615	.641	.636	.666	**.594**	.558	**.468**	**.459**

Table 3: Results (Pearson r and Spearman ρ correlations) on Keller and Lapata's (2003) plausibility data; underlining denotes the best-performing WordNet-based model, boldface denotes the overall best performance

3.2.4 Gold Standards

For Selectional preference modules, and although the plan for WP6.1 initially called for manual annotation of SP gold standards, due to the intensive nature of annotation required, it was decided instead to use the well-accepted methodology of pseudo-disambiguation for PANACEA experiments.

3.3 Multiword Expressions (WP6.2)

MWEs still nowadays pose problems to most language technology and applications. In particular, they impact greatly on the performance of Machine Translation systems and automatic dictionary compilation. If not recognised and handled properly, MWEs will result in mistranslations hampering the overall text readability (see e.g. Monti et al. 2011, Bilal et al. 2005).

Although the past decades have seen many experiments on methods for the automatic acquisition of MWEs, there are not very many readily available and possibly customizable tools, although some have recently been released on a free or open source basis.

3.3.1 Italian MWE Acquisition

CNR developed a tool for the automatic creation of MWE lexicons and tested it on Italian data (for which not many tools are readily available), although the tool is potentially language independent. The tool implements "light" statistical methods for the acquisition of MWEs and collocations in order to be robustly integrated in the distributed web service platform.. The purpose was not so much devising a new or innovative method, but to provide a free to use tool that creates a full lexical resource: The output of the tool in fact is a full MWE lexicon, in LMF-XML. The MWE lexicon building system works by: (1) extracting candidate collocation pairs with a desired POS-tag pair for first and last component of an MWE; (2) applying an initial frequency filter based on local maxima; (3) retrieving full collocation patterns; (4) using distributional evidence to filtered out irrelevant patterns; (5) further post-filtering to reduce noise, (6) building a lexicon in LMF format enriched with morphosyntactic and frequency information.. The multi-step nature of the module is designed to operate efficiently in a web service distributed environment. In PANACEA we decided to implement post-filters as separate pieces of software and services that users can combine with the core module to obtain the desired results.

3.3.1.1 Web Services

3.3.1.1.1 MutiwordExtractor (CNR-ILC)

CNR-ILC has delivered a MWE acquisition component which has been tested on Italian, but is potentially language independent (http://langtech3.ilc.cnr.it:8080/soaplab2-axis/#panacea.extractormwv7_row). The component produces an LMF lexicon where each multi-word expression or term is annotated also with respect to the POS pattern it instantiates, its frequency, association measures, and other information.

This related paper describes an experiment of MW extraction using the PANACEA MW-extraction service. The extraction is conducted first using the original PANACEA crawled corpus, then a "deduplicated" version of the same corpus. The goal is to evaluate the capacity of the tool to deal with noisy data, and in particular with texts containing a significant amount of duplicated paragraphs. The accuracy of the extraction of multi-word expressions from the original crawled corpus is compared to the accuracy of the extraction from a later "de-duplicated" version of the corpus. The paper shows how our method can extract with sufficiently good precision also from the original, noisy crawled data.

3.3.1.2 Related Papers

This work gave birth to two conference papers which provide more details about the implementation:

- Quochi V., Frontini F., Rubino F. (2012) "A MWE Acquisition and Lexicon Builder Web Service". *Proceedings of the COLING 2012*. Mumbai. India.

- Frontini F., Quochi V., Rubino F. (2012) "Automatic Creation of quality Multi-word Lexica from noisy text data" *Proceedings of the Sixth Workshop on Analytics for*

Noisy Unstructured Text Data. COLING 2012. Mumbai, India.

3.3.2 Arabic MWE Acquisition

DCU has investigated the automatic acquisition of Arabic MWEs. Three complementary approaches have been investigated to extract MWEs from available data resources. The first approach relies on correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. The second approach collects English MWEs from Princeton WordNet 3.0, translates the collection into Arabic using Google Translate, and utilizes different search engines to validate the output. The third approach uses lexical association measures to extract MWEs from a large unannotated corpus.

3.3.2.1 Related Papers

This work has been presented at COLING 2010.

- Attia, Mohammed; Toral, Antonio; Tounsi, Lamia; Pecina, Pavel; van Genabith, Josef. (2010). **Automatic Extraction of Arabic Multiword Expressions**. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), Beijing, China: Coling 2010.

In this paper we investigate the automatic acquisition of Arabic Multiword Expressions (MWE). We propose three complementary approaches to extract MWEs from available data resources. The first approach relies on the correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. The second approach collects English MWEs from Princeton WordNet 3.0 (PWN), translates the collection into Arabic using Google Translate, and utilizes different search engines to validate the output. The third uses lexical association measures to extract MWEs from a large unannotated corpus. We experimentally explore the feasibility of each approach and measure the quality and coverage of the output against gold standards.

We use three approaches to identify and extract MWEs: (a) crosslingual correspondence asymmetries, (b) translation-based extraction, and (c) corpus-based statistics. For each approach we use a number of linguistic and statistical validation techniques and both automatic and manual evaluation.

In the first approach we make use of the crosslingual corresopndence asymmetry, or many-to-one relations between the titles in the Arabic Wikipedia and the corresponding titles in other languages to harvest MWEs. In the second approach we assume that automatic translation of MWEs colelcted from PWN into Arabic are high likelihood MWE candidates that need to be automatically checked and validated. In the third approach we try to detect MWEs in a large raw corpus relying on statistical measures and POS-annotation filtering.

3.3.3 Gold Standards

CNR has compiled Italian MWE gold standards for the Environment and Labour Legislation domains by drawing from existing online domain dictionaries and glossaries. A full description will be provided in D7.4. Unfortunately, most of the resources are copyright protected. Thus, CNR will not be able to release them for public distribution as originally intended. However, if requested, they could be made accessible confidentially to reviewers.

3.4 Lexical-Semantic Classes (WP6.3)

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. 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).

PANACEA has included research on nominal, verbal, and adjectival semantic classes.

3.4.1 Nominal Semantic Classes Using Decision Trees and Bayesian Models

UPF has focused on nominal semantic classes, performing research on the use of Decision Trees and Bayesian Models to approach this task. See *Related Papers* section for details about used methods.

Nominal lexical semantic classes gather together properties that appear to be linguistically significant for a number of linguistic phenomena. Determiner selection, selectional restrictions and noun collocation have been described in terms of such groupings of properties. In addition, these classes are often used to generalize over particular senses of different words. For instance, Miller et al. (1990) used a number of lexical semantic classes as features that ordered the nominal meaning hierarchy in WordNet. Applications that use nouns annotated with lexical semantic classes include: machine translation, discrimination of referents in tasks such as event detection and tracking (Fillmore et al., 2006), question answering (Lee et al., 2001), entity typing in named entity recognition (Ciaramita & Altun, 2005; Fu, 2009), automatic building and extending of ontologies (Buitelaar et al., 2005), textual inference (de Marneffe et al., 2009), etc. Furthermore, nominal lexical semantic classes have also recently proved to be useful information for grammar induction (Agirre et al., 2011), where problems come from the need of generalizing over a high dimensional space.

Lexical semantic noun tagging in large lexica is still mostly done by hand, and the high cost of this exercise hinders the production of rich lexica for different languages. In addition, domain tuning of lexica is considered too expensive, and the use of an inadequate lexicon is one of the causes of poor performance of many applications. Thus, current research on automatic production of class-annotated lexica is expected to have a high impact on the performance of most NLP applications. Most critically, it will bring significant improvements in their coverage over different languages and domains. Thus, any reduction in the amount of human work required for the production of these resources can contribute to improve the current situation.

UPF has addressed cue-based noun classification in PANACEA WP6.2. Its main objective is to automatically acquire lexical semantic information by classifying nouns into previously known lexical classes. This is achieved by using particular aspects of linguistic contexts where the nouns occur as cues that represent distributional characteristics of a specific lexical class and which also support the building of specialized classifiers. Note that, although the particular case of lexical-semantic classes has been addressed, the methods can be used to classify words into any other linguistically motivated class.

In order to represent the linguistic contexts that are representative of the class, first of all it is necessary to define a set of *n* linguistically motivated cues that represent those contexts. Then, an *n*-dimensional vector containing the number of times each cue has been seen with the studied noun is built. This vector is stored following the Weka (Witten and Frank, 2005) file format: Attribute-Relation File Format (ARFF, http://weka.wikispaces.com/ARFF).

As for classification methods, UPF used two supervised approaches to perform noun classification: Decision Tree (DT) classifiers and Naive Bayes classifiers with Bayesian inference of the parameter. Both of them are trained using the cue vectors of a set of pre-classified nouns, i.e. the nouns in the gold standard.

Regarding the experiments with Decision Tree classifiers, UPF used a pruned Decision Tree classifier in the Weka (Witten and Frank, 2005) implementation of C4.5DT (Quinlan, 1993), which proved to be very effective in the lexical acquisition tasks of Merlo and Stevenson (2001).

3.4.1.1 Web Services

UPF delivers 7 classifiers for English and 9 for Spanish for the following classes: HUMAN, EVENT, LOCATION, SEMIOTIC (only Spanish), ABSTRACT, ARTIFACT, MATTER, SOCIAL_ENTITY. PROCESS (only Spanish). The classifiers are based on pre-trained Decision Trees, reaching an accuracy between 70 and 80%. The output of the classifier delivers a confidence measure that supports assessment on the quality of every decision in order to be able to separate the instances classified with high precision from those that may need manual revision.

UPF also delivers a web service implementing a Naive Bayes classifier and a service to learn the parameters for these classifiers using Bayesian inference and some useful tools to perform research on lexical classification among others: a corpus indexer and querier, a service to create ARFF files given a set of regular expressions, and a Maximum Likelihood estimator of the probability of each cue given each class given an ARFF file.

In what follows, we give a brief description of the deployed web services. For more details about usage and input/output formats, consult the Registry entry for each web service (http://registry.elda.org/) and the Common Interface and Travelling object definitions in this document.

3.4.1.1.1 dt_noun_classifier_[CLASS] (UPF)

Those are a set of web services that perform the classification of a given set of nouns into a lexical semantic class. There is one web service for each available class: dt_noun_classifier_abstract, dt_noun_classifier_artifact, dt_noun_classifier_eventive, dt_noun_classifier_human, dt_noun_classifier_location, dt_noun_classifier_matter, dt_noun_classifier_process, dt_noun_classifier_semiotic, dt_noun_classifier_social. The classification is performed from PoS annotated data using a pre-trained Decision Tree.

The mandatory input parameters of these web services are: a PoS tagged corpus, tab separated, following FreeLing tagset; the language (English or Spanish) and a label to identify the corpus we are using (e.g. the corpus domain). The input corpus can be introduced with a list of URLs pointing to different PoS files, using the optional parameter *inputIsURLlist*. Using optional parameters, we can choose to classify a list of given nouns or all nouns in corpus that appear more than a certain number of times.

These web services output a lexicon in LMF format containing the classifier predictions for each noun. See the Travelling Object section for details about the format of this lexica.

3.4.1.1.2 noun_classification_filter (UPF)

Given a LMF file with nouns classified with a score (e.g. the output of the dt_noun_classifier_[CLASS]), a threshold for the members of the class and a threshold for the non members of the class, separate those elements that are classified over the threshold from those that are not. There are three cases:

- class nouns (score > 0) over the class threshold: convert them to "class=yes".
- no-class nouns (score < 0) over the non-class threshold (in absolute value): convert them to "class=no".
- nouns under the threshold (in absolute value): convert them to "class=unknown"

3.4.1.1.3 naive_bayes_classifier (UPF)

This web service performs traditional Naive Bayes classification of instances given in a Weka (ARFF) file[3]. It outputs the predicted classification for each instance and some statistics about the performance of the classification. The parameters needed as input can be learnt using estimate_bayesian_parameters web service.

3.4.1.1.4 estimate_bayesian_parameters (UPF)

Given a training set encoded as vectors of cue occurrences, estimate the parameters $P(cue_i|class)$: the probability of seeing each cue as a member or non-member of the class. This estimation is performed using Bayesian inference, which combines prior knowledge with observed data. The parameters estimated with this web service can be used, for example, to classify new instances using a Naive Bayes classifier. The output format is the one needed as input for the naive_bayes_classifier webservice.

3.4.1.1.5 cqp_index (UPF)

This web service implements a corpus indexer based on the IMS Open Corpus Workbench (CWB)[4]. It takes as input a PoS tagged corpus in tabular format and the structure of the data as needed by CWB[5]. The optional parameter *inputIsURLlist* allows us to introduce the input corpus as a list of URLs instead of the direct text.

The output of the service is the ID of the corpus, which can be used to make queries with the cqp_query web service.

[3] http://weka.wikispaces.com/ARFF

[4] http://cwb.sourceforge.net/

[5] http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf

3.4.1.1.6 cqp_query (UPF)

This web service implements a corpus querier based on the IMS Open Corpus Workbench (CWB). The input is the ID of a corpus previously indexed with cqp_index web service and the query following CWB language[6]. The output is the query result.

3.4.1.1.7 create_weka_noun_signatures (UPF)

This web service creates the weka (ARFF) file given a set of regular expressions and a previously indexed corpus (using cqp_index). It will create the vectors for a given list of nouns or for all nouns in corpus over a given threshold.

The output of the service is the weka file and a list of lemmas that were not found in the corpus or that appeared less than the given threshold.

3.4.1.1.8 compute_p_cue_class_from_weka (UPF)

Given a weka file with feature vectors, this web servcie computes the observed frequency (with Maximum Likelihood estimator) of each feature given each class, useful to study cue distribution among classes. The input is a weka file, following the format given by create_weka_noun_signatures web service and the output is a comma separated file with the frequencies of each cue given each class.

3.4.1.2 Related Papers

The work using DTs has been presented at COLING 2010, on recognizing non-deverbal event nouns, and at LREC 2012, on recognizing a wide variety of noun classes.

- Bel, Núria; Romeo, Lauren and Padró, Muntsa. (2012). **Automatic Lexical Semantic Classification of Nouns**. In Proceedings of LREC 2012, Istanbul, Turkey.
- Bel, Núria; Coll, Maria; Resnik, Gabriela. (2010). **Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, China: Coling 2010 Organizing Committee. Pàg. 46-52. ISBN 978-7-900268-00-6.

UPF also investigated on using Bayesian inference (Griffiths et al., 2008; Mackay, 2003) for noun classification. A method to formally introduce linguistic knowledge as priors in a Bayesian framework was proposed. Priors are to compensate the lack of evidence that affects the correct significance assessment of some co-occurrence contexts. The results of the experiments show a significant improvement when learning from small samples with very sparse data. The proposed method for introducing linguistic priors can benefit the development of appropriate lexical resources in contexts where these conditions are met, for instance for less-resourced languages and for domain adaptation of lexical resources.

A paper has been prepared but has not yet been accepted as a conference/journal publication.

- Bel, Núria; Padró, Muntsa. Using linguistic priors in lexical semantic classification: a Bayesian approach. Prepared draft.

[6] http://cwb.sourceforge.net/files/CQP_Tutorial.pdf

3.4.2 Identification of Deverbal Nouns in Italian

CNR has also investigated the identification of deverbal nouns in Italian with event readings, using syntagmatic and collocational cues.

Deverbal nouns obtained through transpositional suffixes (such as -zione; -mento, -tura and -aggio) are commonly known as nouns of action, i.e. nouns which denote the process/action described by the corresponding verbs. However, this class of nouns is also known for a specific polysemous alternation: they may denote the result of the process/action of the corresponding verb. This paper describes a statistically based analysis that helps to develop a classifer for automatic identification of deverbal nouns denoting events in context by exploiting rules obtained from syntagmatic and collocational cues identied by linguists, proposing a methdology for event detection as a key task in order to access information through content.

3.4.2.1 Related Papers

This work was presented at CICLing 2011 and published in a proceedings volume.

- Russo, Irene; Caselli, Tommaso; Rubino Francesco. (2011). **Recognizing deverbal events in context**. *International Journal of Computational Linguistics and Applications –* IJCLA Vol.2 (1-2).

3.4.3 Verbal Semantic Classes Using Hierarchical Verb Clustering

Verbal semantic classes are the most studied of lexical-semantic classes, with a key resource being (Levin, 1993). Lexical-semantic classes for verbs are usually defined by their argument and alternation patterns, which generally correspond to meaning classes. 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.

UCAM has focused partly on verbal semantic classes. Most previous research on verb clustering has focused on acquiring flat classifications from corpus data, although many manually built classifications are taxonomic in nature. NLP applications can also benefit from taxonomic classifications because they vary in terms of the granularity they require from a classification.

3.4.3.1 Related Papers

UCAM has introduced a new clustering method called Hierarchical Graph Factorization Clustering (HGFC) and extended it so that it is appropriate for hierarchical verb clustering. This method outperforms agglomerative clustering when evaluated against a test set from VerbNet. This work has been presented at EMNLP 2011.

- Sun, Lin and Anna Korhonen. (2011). **Hierarchical Verb Clustering Using Graph Factorization.** 2011. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

A variety of verb classifications have been built to support NLP tasks. These include syntactic and semantic classifications, as well as ones which integrate aspects of both (Grishman et al., 1994; Miller, 1995; Baker et al., 1998; Palmer et al., 2005; Kipper, 2005; Hovy et al., 2006). Classifications which integrate a wide range of linguistic properties can be particularly useful for tasks suffering from data sparseness. One such classification is the taxonomy of English verbs proposed by Levin (1993) which is based on shared (morpho-)syntactic and semantic properties of verbs. Levin's taxonomy or its extended version in VerbNet (Kipper, 2005) has proved helpful for various NLP application tasks, including e.g. parsing, word sense disambiguation, semantic role labeling, information extraction, question-answering, and machine translation (Swier and Stevenson,2 004; Dang, 2004; Shi and Mihalcea, 2005; Zapirain et al., 2008).

Because verbs change their meaning and behaviour across domains, it is important to be able to tune existing classifications as well as to build novel ones in a cost-effective manner, when required. In recent years, a variety of approaches have been proposed for automatic induction of Levin style classes from corpus data which could be used for this purpose (Schulte im Walde, 2006; Joanis et al., 2008; Sun et al., 2008; O Seaghdha and Copestake, 2008; Vlachos et al., 2009). The best of such approaches have yielded promising results. However, they have mostly focussed on acquiring and evaluating flat classifications. Levin's classification is not flat, but taxonomic in nature, which is practical for NLP purposes since applications differ in terms of the granularity they require from a classification.

In this paper, we experiment with hierarchical Levin-style clustering. We adopt as our baseline method a well-known hierarchical method – agglomerative clustering (AGG) –which has been previously used to acquire flat Levin-style classifications (Stevenson and Joanis, 2003) as well as hierarchical verb classifications not based on Levin (Ferror,2004; Schulte im Walde, 2008). The method has also been popular in the related task of noun clustering (Ushioda, 1996; Matsuo et al., 2006; Bassiou and Ktropoulos, 2011).

We introduce then a new method called Hierarchical Graph Factorization Clustering (HGFC) (Yu et al., 2006). This graph-based, probabilistic clustering algorithm has some clear advantages over AGG (e.g. it delays the decision on a verb's cluster membership and any level until a full graph is available, minimising the problem of error propagation) and it has been shown to perform better than several other hierarchical clustering methods in recent comparisons (Yu et al.,2 006). The method has been applied to the identification of social network communities (Lin et al., 2008), but has not been used (to the best of our knowledge) in NLP before.

We modify HGFC with a new tree extraction algorithm which ensures a more consistent result, and we propose two novel extensions to it. The first is a method for automatically determining the tree structure (i.e. number of clusters to be produced for each level of the hierarchy). This avoids the need to pre-determine the number of clusters manually. The second is addition of soft constraints to guide the clustering performance (Vlachos et al., 2009). This is useful for situations where a partial (e.g. a flat) verb classification is available and the goal is to extend it.

3.4.4 Verb and Noun Clustering in Automatic Metaphor Identification

UCAM has used verb and noun clustering in the task of automatic metaphor identification. The work on metaphor has served to evaluate and demonstrate the practical usefulness of lexical acquisition, specifically SCF acquisition, SP acquisition, and verb clustering, as it makes use of these techniques in a task. Metaphor is a frequent phenomenon in language and task-based evaluation of lexical acquisition is important for demonstrating the success of the models. The following papers show that the PANACEA techniques can greatly aid metaphor identification. The methods are unsupervised save for an initial seed set.

3.4.4.1 Related Papers

This work was presented at COLING 2010.

- Shutova,Ekaterina; Sun, Lin; Korhonen, Anna. (2010). **Metaphor Identification Using Verb and Noun Clustering**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.

- Shutova, Ekaterina; Van de Cruys, Tim; Korhonen, Anna. (2012). **Unsupervised Metaphor Paraphrasing Using a Vector Space.** Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai, India.

- Shutova, Ekaterian; Teufel, Simone; Korhonen, Anna. (2012). **Statistical Metaphor Processing**. *Computational Linguistics* 39(2).

Besides enriching our thought and communication with novel imagery, the phenomenon of metaphor also plays a crucial structural role in our use of language. Metaphors arise when one concept is viewed in terms of the properties of the other. Below are some examples of metaphor.

(1) How can I *kill* a process? (Martin, 1988)

(2) Inflation has *eaten up* all my savings. (Lakoff and Johnson, 1980)

(3) He *shot down* all of my arguments. (Lakoff and Johnson, 1980)

We present a novel approach to automatic metaphor identification in unrestricted text. Starting from a small seed set of manually annotated metaphorical expressions, the system is capable of harvesting a large number of metaphors of similar syntactic structure from a corpus. Our method is distinguished from previous work in that it does not employ any hand-crafted knowledge, other than the initial seed set, but, in contrast, captures metaphoricity by means of verb and noun clustering. Being the first to employ unsupervised methods for metaphor identification, our system operates with precision of 0.79.

The motivation behind the use of clustering methods for metaphor identification task lies in the nature of metaphorical reasoning based on association. Compare, for example, the target concepts of *marriage* and *political regime*. Having quite distinct meanings, both of them are cognitively mapped to the source domain of *mechanism*, which shows itself in the following examples:

(4) Our relationship is not really *working*.

(5) Diana and Charles did not succeed in *mending* their marriage.

(6) The *wheels* of Stalin's regime were *well oiled* and already *turning*.

We expect that such relatedness of distinct target concepts should manifest itself int he examples of language use, i.e. target concepts that are associated with the same source concept should appear in similar lexico-syntactic environments. Thus, clustering concepts using grammatical relations (GRs) and lexical features would allow us to capture their relatedness **by association** and harvest a large number of metaphorical expressions beyond our seed set. For example, the sentence in (4) being part of the seed set should enable the system to identify metaphors in both (5) and (6).

The system was evaluated against a manually annotated dataset. The precision of the system against the manually annotated dataset was 0.79, against a WordNet-based baseline precision of 0.44.

3.4.5 Verb Classes in French

UCAM has also applied spectral clustering to French verb classes, using lexical, syntactic and semantic features.

3.4.5.1 Related Papers

This work was presented at COLING 2010.

- Sun, Lin; Poibeau, Thierry; Korhonen, Anna; Messiant, Cedric. (2010). **Investigating the cross-linguistic potential of VerbNet-style classification**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.).Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.

Verb class which integrate a wide range of linguistic properties (Levin, 1993) have proved useful for natural language processing applications. However, the real-world use of these classes has been limited because for most languages, no resources similar to VerbNet (Kipper-Schuler, 2005) are available. We apply a verb clustering approached developed for English to French – a language for which no such experiment has been conducted yet. Our investigation shows that not only the general methodology but also the best performing features are transferable between the languages, making it possible to learn useful VerbNet style classes for French automatically without language-specific tuning.

3.4.6 Review of Lexical Classification

A review of lexical classification methods and challenges has been prepared.

3.4.6.1 Related Papers

This work appears in the following paper in the *Philosophical Transactions of the Royal Society.*

- Korhonen, Anna. (2010). **Automatic Lexical Classification - Bridging Research and Practice.** Philoshophical Transactions A of the Royal Society. 368: 3621-3632.

Natural language processing (NLP)—the automatic analysis, understanding and generation of human language by computers—is vitally dependent on accurate knowledge about words.

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

3.4.7 Adjective Classes

Adjective lexical-semantic classes, like nominal classes, are less studied than verbal ones, although there is some recent work on adjective classes. Hatzivassiloglou and McKeown (1993) identified adjective scales, e.g. *hot-warm-cold*, by using cues for scalar adjectives and then clustering adjectives, and evaluated against human-created clusters. Tomuro et al. (2007) presented a clustering algorithm based on word sense induction to cluster adjectives in Japanese and English, evaluated against lexical resources such as WordNet. Navarretta (2000) clustered Danish adjectives based on predicative patterns, followed by manual editing of the clusters. Boleda and Alemany (2003) performed unsupervised acquisition of Catalan adjective classes, with the resulting clusters evaluated by human judges. Boleda (2004) focused on unsupervised clustering of Catalan adjective semantic classes by exploiting a range of shallow distributional linguistc features.

In D6.1 UCAM intended to work on noun classes. However, an opportunity arose to research classification of adjectives rather than nouns. As a result, an MPhil dissertation supervised by PANACEA members has been completed on unsupervised adjective clustering. This work uses spectral clustering to group adjectives using syntactic features, specifically subcategorization patterns, along with co-occurrence and semantic features. Evaluation is against a novel gold standard based on Dixon (1991) and the F-score of the best feature set is 58.

3.4.7.1 Related Papers

- Vo, Quang Phu. **Unsupervised acquisition of adjective classes**. MPhil Dissertation, 2012, University of Cambridge.

Previous researchers have explored the task of clustering nouns and verbs according to their semantic and syntactic behaviour. Not many works have focused on adjectives, although they are equally important for many useful natural language applications. In this work, we investigate a novel task of clustering adjectives into syntactic-semantic classes. A wide range syntactic, semantic and lexical features were extracted from the GigaWord corpus and we experimented using of three clustering algorithms. We evaluate the results using F-measure against the gold-standard created based on the Dixon’s classification for adjectives. We show that our features appear to be useful for the task and yield promising results although the evaluation is particularly challenging. In addition, we report the performance of the adopted machine learning methods using different features and feature representations.

- Vo, Quang Phu; Rimell, Laura; Sun, Lin; Korhonen, Anna. **Unsupervised acquisition of adjective classes**. Draft submission being prepared.

3.4.8 Gold Standards

UPF has developed gold standards for noun classification. A total of 9 different test sets were compiled for English and Spanish nouns for HUMAN, EVENT, LOCATION, SEMIOTIC, ABSTRACT, ARTIFACT, MATTER, SOCIAL_ENTITY, PROCESS. (http://www.panacea-lr.eu/en/info-for-researchers/gold-standards/nominal-classification). All will be available on the PANACEA website.

UCAM has compiled a new gold standard for adjective classes, which will be made available at the PANACEA web site.

3.5 Scientific Articles

Papers describing the different experiments are annexed to this document (Annex A). The complete list of publications is as follows; this is a collection of the papers listed in the individual sections above. Note that papers on merging are included in D6.4.

- Attia, Mohammed; Toral, Antonio; Tounsi, Lamia; Pecina, Pavel; van Genabith, Josef. (2010). **Automatic Extraction of Arabic Multiword Expressions**. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), Beijing, China: Coling 2010.
- Bel, Nuria; Coll, Maria; Resnik, Gabriela. (2010). **Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, China: Coling 2010 Organizing Committee. Pàg. 46-52. ISBN 978-7-900268-00-6.
- Bel, Núria; Romeo, Lauren and Padró, Muntsa. (2012). **Automatic Lexical Semantic Classification of Nouns**. In Proceedings of LREC 2012, Istanbul, Turkey.
- Caselli, Tommaso; Rubino, Francesco; Frontini, Francesca; Russo, Irene; Quochi, Valeria. (2012). **Flexible Acquisition of Verb Subcategorization Frames in Italian.** In Proceedings of LREC2012, Istanbul, Turkey.
- Frontini F., Quochi V., Rubino F. (2012) “Automatic Creation of quality Multi-word Lexica from noisy text data” *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data.* COLING 2012. Mumbai, India.
- Korhonen, Anna. (2010). **Automatic Lexical Classification - Bridging Research and Practice.** Philoshophical Transactions A of the Royal Society. 368: 3621-3632.
- Korhonen, Anna; Lippincott, Tom; ó Séaghdha, Diarmuid; Sun, Lin. (2010). E**xploring variation across biomedical subdomains** . In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.
- Lippincott, Thomas; Korhonen, Anna and Ó Séaghdha, Diarmuid. (2012). **Learning Syntactic Verb Frames Using Graphical models**. In Proceedings of ACL, Jeju Island, Korea.

- Lippincott, Thomas; Rimell, Laura; Verspoor, Karin; Johnson, Helen L.; Korhonen, Anna. **Approaches to Verb Subcategorization for Biomedicine**. Forthcoming, Journal of Biomedical Informatics.

- Ó Séaghdha, Diarmuid and Korhonen, Anna. (2012). **Modelling selectional preferences in a lexical hierarchy**. In Proceedings of *SEM, Montreal, Canada.

- Quochi, Valeria, Frontini, Francesca and Rubino Francesco (2012). **A MWE Acquisition and Lexicon Builder Web Service**. *Proceedings of the COLING 2012*. Mumbay. India

- Rimell, Laura; Poibeau, Thierry and Korhonen, Anna. (2012). **Merging Lexicons for Higher Precision Subcategorization Frame Acquisition**. In Proceedings of the LREC Workshop on Language Resource Merging, Istanbul, Turkey.

- Russo, Irene; Caselli, Tommaso; Rubino Francesco. (2011). R**ecognizing deverbal events in context**. *International Journal of Computational Linguistics and Applications* – IJCLA Vol.2 (1-2).

- Shutova,Ekaterina; Sun, Lin; Korhonen, Anna. (2010). M**etaphor Identification Using Verb and Noun Clustering**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.). Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.

- Shutova, Ekaterian; Teufel, Simone; Korhonen, Anna. (2012). Statistical Metaphor Processing. *Computational Linguistics* 39(2).

- Shutova, Ekaterina; Van de Cruys, Tim; Korhonen, Anna. (2012). **Unsupervised Metaphor Paraphrasing Using a Vector Space.** Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai, India.

- Sun, Lin; Poibeau, Thierry; Korhonen, Anna; Messiant, Cedric. (2010). I**nvestigating the cross-linguistic potential of VerbNet-style classification**. In Huang; Chu-Ren and Jurafsky, Dan (Eds.).Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010.

- Sun, Lin and Anna Korhonen. (2011). **Hierarchical Verb Clustering Using Graph Factorization.** 2011. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

- Van de Cruys, Tim; Thierry Poibeau and Anna Korhonen. (2011). **Latent Vector Weighting for Word Meaning in Context**. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK

- Van de Cruys, Tim; Rimell, Laura; Poibeau, Thierry; Korhonen, Anna. (2012). **Multi-way Tensor Factorization for Unsupervised Lexical Acquisition**. Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai, India: Coling

The following papers are in preparation, under revision, or are unpublished reports.

- Bel, Núria; Padró, Muntsa. **Using linguistic priors in lexical semantic classification: a Bayesian approach**. Prepared draft.

- Dong, Yizhen. **Using Lexical Resources to Improve Parsing Accuracy**. MPhil Dissertation, 2011, University of Cambridge.

- Rimell, Laura; Lippincott, Thomas; Verspoor, Karin; Johnson, Helen L.; Korhonen, Anna. **Acquisition and Evaluation of Verb Subcategorization Resources for Biomedicine**. Under revision, Journal of Biomedical Informatics.
- Vo, Quang Phu. **Unsupervised acquisition of adjective classes**. MPhil Dissertation, 2012, University of Cambridge.

PANACEA project members have been involved in organising a number of workshops to encourage the development of techniques required for real-world lexical acquisition. A list of the workshops appears here (note an additional workshop on Lexical Merger appears in D6.4).

- Omri Abend, Anna Korhonen, Ari Rappoport and Roi Reichart. 2011. Proceedings of the EMNLP Workshop on Unsupervised Learning in NLP.
- Omri Abend, Chris Biemann, Anna Korhonen, Ari Rappoport, Roi Reichart and Anders Sogaard. 2012. Proceedings of the EACL Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP.

4 Intregration of lexical acquisition tools into PANACEA platform.

In this section, we present the list of web services integrated into the platform, their common interfaces and the produced travelling objects. Also, we report on the developed workflows.

4.1 Lexical Acquisition Web Services

The consolidated list of deployed web services for lexical acquisition is presented here. See the following sections of this document for additional description of the individual web services, broken down by lexical acquisition task.

We omit here the supporting services such as tokenizers, lemmatizers, and taggers, which are generally pre-requisites for lexical acquisition to take place, and which are listed in D3.4. Here we focus on the high level lexical acquisition components. We also include four parser web services which have been newly introduced in version 3 of the Platform, since parsers are mid-level tools which are crucial for much lexical acquisition. (Abbreviations: LC = Lexical Classification, SCF = Subcategorization Frames, MWE = Multiword Expressions.)

Name	**Task**	**Category**	**Language**	**Provider**	**Registry Number**
noun_classification_filter	LC	Lexicon/Terminology Extraction	Language-independent	UPF	246
dt_noun_classifier_abstract	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	264
dt_noun_classifier_artifact	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	265
dt_noun_classifier_eventive	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	227
dt_noun_classifier_	LC	Lexicon/Terminol	Spanish, English	UPF	243

human		ogy Extraction			
dt_noun_classifier_location	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	244
dt_noun_classifier_matter	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	266
dt_noun_classifier_process	LC	Lexicon/Terminology Extraction	Spanish	UPF	267
dt_noun_classifier_semiotic	LC	Lexicon/Terminology Extraction	Spanish	UPF	268
dt_noun_classifier_social	LC	Lexicon/Terminology Extraction	Spanish, English	UPF	269
naive_bayes_classifier	LC	Lexicon/Terminology Extraction	Language-independent	UPF	229
estimate_bayesian_parameters	LC	Lexicon/Terminology Extraction	Language-independent	UPF	228
create_weka_noun_signatures	LC	Lexicon/Terminology Extraction	Language-independent	UPF	226
compute_p_cue_classes_from_weka	LC	Statistics Analaysis	Language-independent	UPF	225
SubcategorizationFramesExtractor_IT	SCF	Lexicon/Terminology Extraction	Italian	CNR-ILC	212
estrattore_scf_lang_indip	SCF	Lexicon/Terminology Extraction	Language-independent	CNR-ILC	250
tpc_subcat_inductive	SCF	Lexicon/Terminology Extraction	Language-independent	UCAM	223
MultiwordExtractor_IT	MWE	Lexicon/Terminology Extraction	Italian	CNR-ILC	211
TPC_Desr_dependencyparser_it	Parsing	Syntactic Tagging	Italian	CNR-ILC	210
malt_parser	Parsing	Syntactic Tagging	Spanish	UPF	249
freeling3_dependency	Parsing	Syntactic Tagging	English, Catalan, Spanish, Asturian, Galician	UPF	240
tpc_rasp	Parsing	Syntactic Tagging	English	UCAM	222

4.2 Common Interfaces for Lexical Acquisition Components

All PANACEA components deployed as web services make use of Common Interfaces as a way to ensure interoperability. Common Interfaces provide users and Web Service Providers with a reference showing which mandatory parameters must be set for each functionality (PoS tagging, tokenization, sentence alignment, subcategorization frame induction, lexical class induction, etc.).

The rationale and background for the Common Interfaces has been explained in detail in deliverable D3.1; see also D3.3, D3.4. Each release of the Platform has required that a Common Interface be defined for the tools released. Common Interfaces have therefore been designed for each WP6 component type included in the 3rd release of the Platform.

This section sets out the Common Interfaces for all the lexical acquisition component types in the WP6 modules. Each component type (except a few pipeline-internal components such as the weka creator) has mandatory input and output parameters which each web service must conform to, as well as a number of optional inputs and outputs which may be used by different tools instantiating the same lexical acquisition task. Some of the lexical acquisition component types have relatively complex interface specifications, especially a wide variety of optional parameters, since they are highly configurable. Note that common interfaces are common for each component type, for example, the CI for SCF extractor is common for the extractor developed by UCAM and ILC.

4.2.1 Verb SCF Extractor

- **Inputs**
  - Mandatory
    - input (text) (Description: parsed corpus)
    - corpus_structure (pick list) (Description: choose from the list of parser output formats that the component accepts as input)
  - Optional
    - verb_tags (text) (Description: list of verb POS tags to allow identification of the verbs in the corpus regardless of tagset; can be a regular expression)
    - lemmas (text) (Description: a list of verb lemmas for which to extract SCFs)
    - target_dependency_types (text) (Description: list of dependency labels which should be considered as part of SCFs)
    - ignore_dependency_types (text) (Description: list of dependency labels which indicate a verb instance should be ignored, e.g. if passives should be ignored)
    - pos_groups (text) (Description: how to group POS tags together for SCFs which include information about POS tag of dependent, e.g. a single grouping containing noun and pronoun pos tags)
    - dependency_types_to_deppos (text) (Description: list of dependency types where the POS tag group of the dependent is part of the SCF)
    - dependency_types_to_lexicalize (text) (Description: list of dependency types where the lexical value of the dependent is part of the SCF)
    - dependency_types_lex_groups (text) (Description: groups of lexical items for lexicalized dependencies, e.g. grouping all directional prepositions together)
    - dependency_types_to_extend (text) (Description: list of dependency types where the dependent's dependencies are part of the SCF)
    - filtering_type (pick list): which type of filtering to use, e.g. relative frequency thresholding

    - filtering_parameter (text): a user-defined parameter for the filtering, e.g. if user wishes to specify a frequency threshold
    - confidence (H|M|L) (Description: level of confidence desired for the output, high/medium/low, default H)
    - configuration_properties (text): (Description: a file to customise various properties of the extractors)
- **Outputs**
  - lexicon (Description: SCF lexicon)

4.2.2 CQP indexer

- **Inputs**
  - Mandatory
    - corpus (text): PoS tagged corpus, tab separated (e.g. in the form: word lemma PoS).
    - structure (text): structure of the corpus, this is, order of components. For example: -P lemma -P pos -P token is the structure for FreeLing.
  - Optional
    - charset (pick list): encoding of the corpus (may be iso or utf-8)
    - inputIsURLlist (boolean): whether the input is a list of urls containing PoS tagged files or a whole corpus
- **Outputs**
  - corpusId: ID of indexed corpus

4.2.3 CQP querier

- **Inputs**
  - Mandatory
    - query (text): CQP query, e.g. [lemma="be"]; cat;
    - corpusId (string): ID of the CQP indexed corpus.
- **Outputs**
  - output: Result of the CQP query

4.2.4 DT noun classifier

- **Inputs**
  - Mandatory
    - input (text): PoS tagged corpus, tab separated, in the form: word lemma PoS. The tagset must follow FreeLing tagset
    - language (pick list): language
    - label (string): label to indentify the corpus from which we are extracting noun occurrences to classify. For example, it may indicate the domain or the origin of the corpus. This label will be used in the output LMF.
    - Optional
    - inputIsURLlist (boolean): whether the input is a list of urls containing PoS tagged files or a whole corpus
    - lemmas (text): a list of lemmas to be classified. If no list is given, all nouns in corpus will be classified
    - minOccurrences (number): minimal number of occurrences in corpus to take the noun into account and classify it. All nouns occurring less than this number won't be classified.
    - output_type (pick list): scored or filtered. Indicates whether a filter to select nouns classified with high precision should be used or not.

- **Outputs**
  - LMFoutput: lexicon in LMF containing the classifier predictions for the selected nouns
  - weka: weka file with the signatures used by the DT. Useful for debugging purposes
  - notFoundLemmas: list of lemmas that were not in the corpus or appeared less than the selected number of occurrences.

4.2.5 noun_classification_filter

Filters the output of the noun classifier given two thresholds

- **Inputs**
  - Mandatory
    - input (text): LMF file with scored classification of nouns
    - class_threshold (number): Threshold for members of the class (only nouns classified with a highest score than this threshold will be considered as class members)
    - no_class_threshold (number): Threshold for non-members of the class (only nouns classified with a negative score that is highest (in absolute value) than this threshold will be considered as non-class members)
  - Optional
    - class_name: class to be filtered (in case there are several classes in input LMF file). If empty, the same filter will be applied to all classes.
    - corpusLabel: corpusLabel to be filtered (in case there are several corpusLabels in input LMF file). If empty, the same filter will be applied to all corpusLabels.
- **Outputs**
  - filtered_LMF: LMF with yes/no/unknown information instead of scored information.

4.3 MWE Extractor

- **Inputs**
  - Mandatory
    - input (text) (Description: POS-tagged, possibly parsed corpus, CoNNL format)
  - Optional
    - tag1 (text) (Description: POS tag or class for first word of extracted MWEs)
    - tag2 (text) (Description: POS tag or class for second word of extracted MWEs)
    - extraction_type (postag|deprel) (Description: whether to extract MWEs using POS tags alone or also dependency relations)
    - window (number) (Description: size of window to search for MWE pairs, only relevant if extraction_type=postag)
    - depth_rel (number) (Description: how many dependencies the extractor looks for, only relevant if extraction_type=deprel)
    - stop_words (url) (Description: file containing a list of stop words)
    - bad_words (url) (Description: file containing a list of words to be discarded)
    - bad_multiwords (url) (Description: file containing a list of multiwords to be discarded)
    - prefilter (text) (Description: choice of pre-filter from tool documentation)
    - order_by (freq|ll|pmi) (Description: how to order pairs in the output file)
- **Outputs**
  - Mandatory
    - mwe lexicon (url|file): (Description: output lexicon as either URL to the actual resource, or a file)
  - Optional

- outputType(predefined params): (Description: the type of output format for the lexicon. E.g. XML-LMF, tabbed,...)

4.4 Travelling Object definitions for Acquired Lexica

In this section, the format of the lexica acquired and delivered by PANACEA platform is specified. The targeted standard format chosen for these travelling objects is basic Lexical Markup Framework, LMF (Francopoulo et al. 2008).

In this document, the format of the lexica acquired and delivered by PANACEA platform is specified. The targeted standard format chosen for these travelling objects is basic Lexical Markup Framework, LMF (Francopoulo et al. 2008).

Below we provide LMF examples for the levels of concern in PANACEA, all of them based on LMF using DTD in revision 16.

4.4.1 General issues about LMF

(extracted from LMF specifications revision 16[7])

"The LMF core package describes the basic hierarchy of information of a lexical entry, including information on the form. The core package is supplemented by various resources that are part of the definition of LMF. These resources include:

- Specific data categories used by the variety of resource types associated with LMF, both those data categories relevant to the metamodel itself, and those associated with the extensions to the core package (for data categories here we understand the names of the XML elements, that correspond to the main building blocks of a lexical resource (e.g. LexicalEntry, Lemma, Sense etc), and of the mandatory attributes (e.g. id, entry, targets…)).
- The constraints governing the relationship of these data categories to the meta-model and to its extensions;
- Standard procedures for expressing these categories and thus for anchoring them on the structural skeleton of LMF and relating them to the respective extension models;
- The vocabularies used by LMF to express related informational objects for describing how to extend LMF through linkage to a variety of specific resources (extensions) and methods for analyzing and designing such linked systems.

[...]

LMF extensions are expressed in a framework that describes the reuse of the LMF core components (such as structures, data categories, and vocabularies) in conjunction with the additional components required for a specific resource.

[...]

LMF provides general structures and mechanisms for analyzing and designing new electronic lexical resources, but LMF does not specify the structures, data constraints, and vocabularies to be used in the design of specific electronic lexical resources. LMF also provides mechanisms for analyzing and describing existing resources using a common descriptive framework. For the purpose of both designing new lexical resources and describing existing lexical resources, LMF

[7] http://www.tagmatica.fr/lmf/iso_tc37_sc4_n453_rev16_FDIS_24613_LMF.pdf

defines the conditions that allow the data expressed in any one lexical resource to be mapped to the LMF framework, and thus provides an intermediate format for lexical data exchange."

4.4.2 General information about the resource/lexicon

LMF requires two high level, general elements: Lexical Resource and Lexicon. They are supposed to be used to encode general and administrative information about the resource and the lexicons included in it such as size, date of creation, authors, availability, and so on. Here below we list the features that are/can be added automatically by the service and those that may be added manually if the final user wants to publish/distribute the resource.

The used features will be presented in tables, containing the name of the attribute, the kind of values it can have and whether it is mandatory or optional. The proposed features are already compliant to Metashare. Furthermore, most of them are traceable in IsoCat. In the tables below we give the link to these correspondences when available.

4.4.2.1 Features added automatically by the lexicon acquisition component(s)

4.4.2.1.1 GlobalInformation

In the table below we list the features that can be found under `<GlobalInformation>` in PANACEA generated lexica.

Attribute	Value	Status	IsoCat
resourceType	"lexicalConceptualResource"	mandatory	http://www.isocat.org/rest/dc/3806
lexicalConceptualResourceType	"lexicon"	mandatory	http://www.isocat.org/datcat/DC-2487
conformanceToStandardsBestPractices	"LMF"	mandatory	
mediaType	"text"	mandatory	http://www.isocat.org/datcat/DC-2490
mimeType	"text/xml"	mandatory	http://www.isocat.org/datcat/DC-2571
characterEncoding	"UTF-8"	mandatory	http://www.isocat.org/datcat/DC-2564
resourceName	open	optional	
modalityType	"written language"		http://www.isocat.org/datcat/DC-2490

4.4.2.1.2 Lexicon

Those are the features that are related to `<Lexicon>` entry. Nevertheless, some PANACEA tools (such as the mergers) deliver this information in `<GlobalInformation>` in order to ease the extraction of the data to be used in the Metashare Metadata creator. The idea is to put all metadata in `<GlobalInformation>` and then convert automatically the LMF data inside this element to Metadata compliant with Metashare.

Attribute	Value	Status	IsoCat
originalSource	open (e.g. corpu name)*	optional	http://www.isocat.org/datcat/DC-2534
domain	open *	optional	http://www.isocat.org/datcat/DC-2467
size	open, value type: number.	mandatory	http://www.isocat.org/datcat/DC-2580
sizeUnit	open suggested: "entry")	mandatory	http://www.isocat.org/datcat/DC-2583
creationMode	open (suggested value list: "automatic|manual|mixed")	mandatory	http://www.isocat.org/datcat/DC-2516
creationModeDetails	open	optional	http://www.isocat.org/datcat/DC-2511
creationTool	open	optional	
encodingLevel	open	optional	
lingualityType	list: "monolingual|bilingual|multilingual"	mandatory	http://www.isocat.org/datcat/DC-2491
languageID	open (the feat can be repeated)	mandatory	http://www.isocat.org/datcat/DC-2482
languageName	open (the feat can be repeated)	mandatory	http://www.isocat.org/rest/dc/2484
resourceName	open	optional	http://www.isocat.org/datcat/DC-2545

* Value should be passed to a wrapper from the Xces/Graf header.

4.4.2.2 Features to be added manually by the lexicon "curator"/publisher

Here we include only those recommended features, compliant to Metashare that may be added manually to those indicated above.

4.4.2.2.1 GlobalInformation

Attribute	Value	Status	IsoCat
description	free text	optional	http://www.isocat.org/datcat/DC-2520
availability	"available-restrictedUse"	optional	Similar to: http://www.isocat.org/datcat/DC-2453
license	open (recommended "CC-BY-3.0")	optional	http://www.isocat.org/datcat/DC-2457

licenseurl	open	optional	http://www.isocat.org/datcat/DC-2457
restrictionsOfUse	open (recommended “attribution”)	optional	
attributionText	open, free text	optional	
distributionAccessMedium	“downloadable”	optional	http://www.isocat.org/datcat/DC-2458
foreseenUse	“nlpApplication”	optional	
owner	open, free text	optional	
email	email	recommended	http://www.isocat.org/datcat/DC-2521
organisationName	open, free text	optional	http://www.isocat.org/datcat/DC-2459
organisationShortName	open, achronym	optional	
departmentName	open, free text	optional	
projectName	open, free text	optional	http://www.isocat.org/datcat/DC-2537
projectShortName	open	optional	http://www.isocat.org/datcat/DC-2536
projectID	open	optional	http://www.isocat.org/datcat/DC-2535
fundingType	open (see Metashare list)	optional	

4.4.2.2.2 Lexicon

Attribute	Value	Status	IsoCat
version	number	optional	http://www.isocat.org/datcat/DC-2547
description	open, free text	optional	http://www.isocat.org/datcat/DC-2520

4.4.2.3 Example

```
<?xml version=“1.0” encoding=“UTF-8”?>
<!DOCTYPE LexicalResource SYSTEM
“http://www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd” >
<LexicalResource dtdVersion=“16”>
    <!-- metadata for GlobalInfo are as far as possible compliant to MetaShare
    metadata and/or ISO Cat. Penny Labropoulou, Elina Desipri (eds)
    Documentation and User Manual of the META-SHARE Metadata Model. Date:
```

```
06/03/2012.  http://www.meta-net.eu/meta-share/metadata-schema/.  However,
metadata are given as flat lists; not organised into components
Some of these metadata have been manually added for distributions. Other
are automatically added by the acquisition tool.-->


<GlobalInformation>
    <feat att=“resourceType” val=“lexicalConceptualResource” />
    <feat att=“lexicalConceptualResourceType” val=“lexicon” />
    <feat att=“resourceName” val=“PANACEA_SCF_IT_ENV” />
    <feat att=“description” val=“This is the PANACEA acquired SCF lexica
    for Italian and Environment domain” />
    <feat att=“conformanceToStandardsBestPractices” val=“LMF” />
    <feat att=“mediaType” val=“text” />
    <feat att=“modalityType” val=“writtenLanguage” />
    <feat att=“characterEncoding” val=“UTF-8” />
    <feat att=“availability” val=“available-restrictedUse” />
    <feat att=“license” val=“CC-BY-3.0” />
    <feat att=“licenseurl” val=http://creativecommons.org/licenses/by/3.0/ />
    <feat att=“restrictionsOfUse” val=“attribution” />
    <feat att=“attributionText” val=“The Language Resource Group. CNR-ILC.
    Caselli et al.(2012)” />
    <feat att=“distributionAccessMedium” val=“downloadable” />
    <feat att=“foreseenUse” val=“nlpApplication” />
    <feat att=“owner” val=“The Language Resources Group” />
    <feat att=“email” val=“risorse@ilc.cnr.it” />
    <feat att=“organisationName” val=“Consorzio Nazionale delle Ricerche”/>
    <feat att=“organisationShortName” val=“CNR-ILC” />
    <feat att=“departmentName” val=“Istituto di Linguistica Computazionale
    A. Zampolli” />
    <feat  att=“projectName”  val=“Platform  for  Automatic,  Normalised
    Annotation  and  Cost-Effective  Acquisition  of  Language  Resources  for
    Human Language Technologies” />
    <feat att=“projectShortName” val=“PANACEA” />
    <feat att=“projectID” val=“FP7-ICT-2009-4-248064” />
    <feat att=“fundingType” val=“euFunds” />
    <feat  att=“description”  val=“This  is  an  automatically  acquired  and
    created lexicon for verb subcategorisation frames for the Environment
    domain.” />
</GlobalInformation>
<Lexicon>
    <feat att=“domain” val=“Environment” />
    <feat att=“encodingLevel” val=“syntax” />
    <feat att=“linguisticInformation” val=“syntax-SubcatFrame” />
    <feat att=“creationMode” val=“automatic” />
    <feat att=“creationModeDetails” val=“induction” />
    <feat  att=“creationTool” val=“SCF_Extractor_IT” />
    <feat att=“creationDate” val=“20120715” />
    <feat att=“originalSource” val=“PANACEA_MCv2_ENV_IT” />
    <feat att=“version” val=“1.0” />
```

```
        <feat att="lingualityType" val="monolingual" />
        <feat att="languageID" val="it" />
        <feat att="languageName" val="Italian" />
        <feat att="size" val="370" />
        <feat att="sizeUnit" val="SyntacticBehaviour" />
        <LexicalEntry id="le_1">
            (...)
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

4.4.3 PANACEA SubCat lexicon format

We present and discuss an example with one lexical entry for the verb *accusare*. Two syntactic frames for this verb are described here as an example:

@SUBJ@OBJ (i.e. a syntactic frame with two arguments/complements: a subject and a direct object)

@SUBJ@OBJ@COMP-DI (i.e. a syntactic frame with three arguments/complements: a subject, a direct object, and a prepositional phrase complement introduced by the preposition *di*)

Some comments are contained in the appropriate comment field. <!-- -->

```
<Lexicon>
  <LexicalResource>
    <!-- LexicalEntry represents the verb main entry -->
    <LexicalEntry id="le_1">
      <feat att="partOfSpeech" val="V" /> <!-- it is recommended to use feat
      "partOfSpeech" to set the partOfSpeech of the entry -->
      <!--Lemma is obligatory in LMF and should be used to
      encode the morphosyntactic information applicable to the
      whole lemma -->
      <Lemma>
          <feat att="writtenForm" val="accusare"/>
      </Lemma>
      <!-- SyntacticBehaviour contains the link btw the verb and the
      subcategorisation frame(s) relevant for the verb at hand; auxiliary and frequency
      information can also be encoded here.  Because of the nature of the extracted data and
      especially because of frequency information that usually refers to the verb-subcat pair ,
      we would recommend to have each syntactic behaviour point to only one
      subcategorisation frame. But this is not constrained by LMF, and infact may not be
      true nor a good practice for other types of lexicons -->
      <SyntacticBehaviour id="sb_1" subcategorizationFrames="scf_11">
        <feat att="aux" val="avere"/>
        <feat att="freq" val="0.3"/>
        <!--"domain" is a label used to identify the domain in the case of manually
        developed lexica or the corpus from where the lexicon has been extracted, in case it
        has been automatically acquiered -->
```

```
     <feat att=“domain” val=“general”/>
 </SyntacticBehaviour>

 <SyntacticBehaviour id=“sb_2” subcategorizationFrames=“scf_22”>
  <feat att=“aux” val=“avere”/>
  <feat att=“freq” val=“0.04”/>
  <feat att=“domain” val=“general”/>
 </SyntacticBehaviour>
</LexicalEntry>

<!-- SubcategorizationFrame contains the description of the syntactic structures in
terms of syntactic arguments -->
<SubcategorizationFrame id=“scf_11”>
  <feat att=“scf-type” val=“@obj”/> <!-- the attribute scf-type here is used
  simply to assign a lable to the whole SCF, which may be useful for evaluation purposes
  -->
  <!-- SyntacticArgument specifies the properties of each single argument: e.g.
  information about position, function, its optionality, syntactic realization, etc may be
  expressed here. In the PANACEA TO it is recommended/mandatory to use function
  (to express the grammatical function of the argument) and/or realisation (to express
  somehow the surface realization of the argument) as the key/obligatory features for
  syntctic arguments. They will be used e.g. for merging purposes. Optionality and
  position are optional.-->
  <SyntacticArgument>
    <feat att=“position” val=“0”/>
    <feat att=“optionality” val=“yes”/>
    <feat att=“function” val=“subj”/>
    <feat att=“realization” val=“NP”/>
  </SyntacticArgument>

  <SyntacticArgument>
    <feat att=“position” val=“1”/>
    <feat att=“optionality” val=“no”/>
    <feat att=“function” val=“obj”/>
    <feat att=“realization” val=“NP”/>
  </SyntacticArgument>
</SubcategorizationFrame>

<SubcategorizationFrame id=“scf_22”>
   <feat att=“scf-type” val=“@obj@comp-di”/>
   <SyntacticArgument>
    <feat att=“position” val=“0”/>
    <feat att=“optionality” val=“yes”/>
    <feat att=“function” val=“subj”/>
    <feat att=“realization” val=“NP”/>
   </SyntacticArgument>
```

```
       <SyntacticArgument>
         <feat att=“position” val=“1”/>
         <feat att=“optionality” val=“no”/>
         <feat att=“function” val=“obj”/>
         <feat att=“realization” val=“NP”/>
       </SyntacticArgument>

       <SyntacticArgument>
         <feat att=“position” val=“2”/>
         <feat att=“optionality” val=“no”/>
         <feat att=“function” val=“comp”/>
         <feat att=“realization” val=“PP_di”/>
       </SyntacticArgument>
      </SubcategorizationFrame>
  </Lexicon>
</LexicalResource>
```

4.4.3.1 Spanish SCF

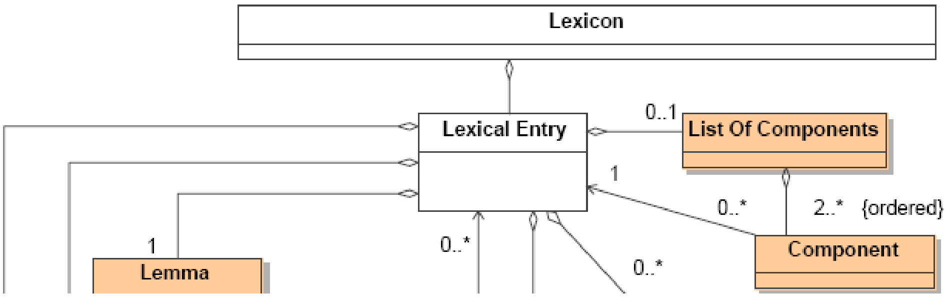
The general LMF structure used in the PANACEA lexica is common for all languages. Nevertheless, there is some information that can change in the different languages depending on the kind of information available in each case. Regarding SCFs, the concrete realization of the `SyntacticArgument` for Spanish is different than Italian. Thus, here we present some examples of Spanish `SyntacticArgument` and their contents:

```
  <SyntacticArgument id=“syn_arg_43_1”>
    <feat att=“position” val=“1”/>
    <!-- for Spanish there is only one kind of complement, named “comp”. The realization
    states the different kind of complements>
    <feat att=“function” val=“comp”/>
    <!-- np: noun phrase -->
    <feat att=“realization” val=“np”/>
  </SyntacticArgument>

  <SyntacticArgument id=“syn_arg_47_1”>
    <feat att=“position” val=“1”/>
    <feat att=“function” val=“comp”/>
    <!-- ppa: indirect object -->
    <feat att=“realization” val=“ppa”/>
  </SyntacticArgument>

  <SyntacticArgument id=“syn_arg_43_2”>
    <feat att=“position” val=“2”/>
    <feat att=“function” val=“comp”/>
    <!-- cp: object is a clause phrase, state also the kind of clause.-->
    <feat att=“realization” val=“cp”/>
    <!-- inf: infinitive clause -->
    <feat att="type" val="inf"/>
  </SyntacticArgument>

  <SyntacticArgument id="syn_arg_43_2">
    <feat att="position" val="2"/>
    <feat att="function" val="comp"/>
    <!-- cp: object is a clause phrase, state also the kind of clause -->
    <feat att="realization" val="cp"/>
    <!-- fin: finite clause -->
    <feat att="type+cl_type" val="fin"/>
  </SyntacticArgument>

  <SyntacticArgument id="syn_arg_43_3">
    <feat att="position" val="3"/>
    <feat att="function" val="comp"/>
    <!-- pp: prepositional complement, state which kind of object it has -->
    <feat att="realization" val="pp"/>
    <!-- concrete preposition that introduces the pp: -->
    <feat att="prep" val="a"/>
    <!-- the object of the pp can be "np" or "cp". If it is a "cp" the type of the "cp" is also
    stated -->
    <feat att="pp_object" val="cp"/>
    <feat att="pp_object+type" val="inf"/>
  </SyntacticArgument>

  <SyntacticArgument id="syn_arg_44_0">
    <!-- the subject is essentitally equal to Italian, the realization can be "np" or "cp". If it
    is "cp", it can have the same options than complement "cp" -->
    <feat att="position" val="0"/>
    <feat att="function" val="subj"/>
    <feat att="optionality" val="yes"/>
    <feat att="realization" val="np"/>
  </SyntacticArgument>
```

4.4.4 PANACEA Multiword lexicon format

For the TO for MW lexica we propose the simplest representational means offered by LMF.

LMF has 3 possible extensions for representing MWE: the “Morphology”, the “Morphological patterns” and the “NLP multiword expression patterns” extensions, but the main components for their representations, available in all extensions, are List Of Components (aggregated to Lexical Entry) and Component (aggregated to List Of Components and pointing to Lexical Entry).



Already with these representational objects we may be able to describe also the internal composition and properties of MWE in a relatively simple way.

Lexical Entry may contain data categories that specify that the entry is a multiword and a data category specifying the POS (or MWE) pattern it instantiates.

See an example below:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LexicalResource SYSTEM
"http://www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
  <-- metadata for genaral info, to make compliant to MetaShare
  metadata -->
  <GlobalInformation>
    <feat att=“originalSource” val=“panacea_corpus_20111023”/>
        <feat att=“crawlDate” val=“2011”/>
        <feat att=“size” val=“20”/>
        <feat att=“sizeUnit” val=“words”/>
        <feat att=“sizeUnitMultiplier” val=“million”/>
        <feat att=“author” val=“CNR”/>
        <feat att=“creationMode” val=“automatic”/>
        <feat att=“creationModeDetails” val=“acquisition”/>
  </GlobalInformation>

  <Lexicon>
    <feat att="type" val="Panacea_MWE_Lexicon"/>
    <feat att="language" val="Italian"/>
    <--here follows a list of single words that are used in the Multiword lexicon -->
    <LexicalEntry id="le_ea38d68660cd14356bbc858586790d1e">
      <feat att="entryType" val="Singleword"/>
      <feat att="absoluteFrequency" val="33675"/>
        <feat att="pos" val="s"/>
      <Lemma>
        <feat att="writtenForm" val="datore"/>
      </Lemma>
    </LexicalEntry>
    <LexicalEntry id="le_fbc2154ed38299eea3458847abababafe3">
      <feat att="entryType" val="Singleword"/>
      <feat att="absoluteFrequency" val="295032"/>
        <feat att="pos" val="s"/>
      <Lemma>
        <feat att="writtenForm" val="lavoro"/>
      </Lemma>
    </LexicalEntry>
    <LexicalEntry id="le_ad72734656bb0f51bdd5dfcfcb35607f">
      <feat att="entryType" val="Singleword"/>
        <feat att="pos" val="e"/>
      <Lemma>
        <feat att="writtenForm" val="di"/>
      </Lemma>
    </LexicalEntry>
```

<--here is the list of actual MWEs, with their features and list of components. Each component points to the single Lexical Entry as referred above; the MWEs contain the feature Domain to mark the fact that they belong to a special domain. -->

```
<LexicalEntry id="le_254b8f8a92b5d4efdd22e057edae1874">
  <feat att="entryType" val="Multiword"/>
  <feat att="MWEPattern" val="s+e+s"/>
  <feat att="absoluteFrequency" val="32149"/>
  <feat att="logLikelihood" val="0.002367902295912881"/>
  <feat att="writtenform" val="datore di lavoro"/>
  <feat att="lemmaPair" val="datore-lavoro"/>
  <feat att="domain" val="labour"/>
  <Lemma></Lemma>
  <ListOfComponents>
    <Component entry="le_ea38d68660cd14356bbc858586790d1e">
      <feat att="rank" val="0"/>
      <feat att="pos" val="s"/>
      <feat att="lemma" val="datore"/>
      <feat att="writtenform" val="datore"/>
      <feat att="function" val="head"/>
    </Component>
    <Component entry="le_ad72734656bb0f51bdd5dfcfcb35607f">
      <feat att="rank" val="1"/>
      <feat att="pos" val="e"/>
      <feat att="lemma" val="di"/>
      <feat att="writtenform" val="di"/>
    </Component>
    <Component entry="le_fbc2154ed38299eea3458847ababafe3">
      <feat att="rank" val="2"/>
      <feat att="pos" val="s"/>
```

```
          <feat att="lemma" val="lavoro"/>
          <feat att="writtenform" val="lavoro"/>
        </Component>
      </ListOfComponents>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

4.4.5 PANACEA Lexical Classes lexicon format

We present two LMF examples for lexical semantic classes. Our proposal is to include the information regarding the semantic class under `<Sense>` entry. For the given examples, we will assume that the nouns are classified in three classes: eventive, human and location. The two different LMF samples that we present differ only on how the information of belonging or not belonging to the class is encoded. This depends on how the classifier is used:

- **Scored LMF**: each noun in the lexicon receives a score (between -1 and 1) for each class indicating the confidence of the classifier. If the score is higher than 0, the noun is considered a member of the class, scores close to 1 indicate high confidence of the classifier. If the score is below zero, it is considered a non-member of the class (with more confidence as closer to -1 is the score).
- **Filtered LMF:** instead of giving a score for the classification, the nouns receive a ternary classification: yes/no/unknown. The unknown elements are those that have been classified with small confidence by the classifier.

4.4.5.1 Scored LMF example:

```
<Lexicon>
  <LexicalEntry id="le_1">
    <!--Lemma is obligatory in LMF and should be used to encode the morphosyntactic information applicable to the whole lemma -->
    <Lemma>
        <feat att="writtenForm" val="boy"/>
    </Lemma>
    <!--use feat "partOfSpeech" to set the PoS of the entry -->
    <feat att="partOfSpeech" val="noun"/>
    <Sense>
        <!-- add one feat for each class and its assigned score -->
        <!-- "boy" belongs to the class human but not to location or eventive -->
      <feat att="event" val="-0.85"/>
      <feat att="hum" val="0.95"/>
      <feat att="loc" val="-0.75"/>

      <!-- "domain" is a label used to identify the corpus from where the lexicon has been extracted, in case it has been automatically acquiered -->
      <feat att="domain" val="labour"/>
    </Sense>
  </LexicalEntry>
  <LexicalEntry id=“le_2”>
    <feat att=“partOfSpeech” val=“noun”/>
    <Lemma>
        <feat att=“writtenForm” val=“car”/>
    </Lemma>
    <Sense>
        <feat att=“domain” val=“labour”/>
      <!-- “car” does not belong to the class human nor eventive, but it has small score to
      belong to class location -->
      <feat att=“event” val=“-0.65”/>
      <feat att=“hum” val=“-0.75”/>
      <feat att=“loc” val=“0.25”/>
    </Sense>
  </LexicalEntry>
  <LexicalEntry id=“le_3”>
    <feat att=“partOfSpeech” val=“noun”/>
    <Lemma>
      <feat att=“writtenForm” val=“storm”/>
    </Lemma>
    <Sense>
      <feat att=“domain” val=“labour”/>
        <!-- “storm” belongs to the class eventive but not to location or human -->
      <feat att=“event” val=“0.95”/>
      <feat att=“hum” val=“-0.75”/>
      <feat att=“loc” val=“-0.80”/>
    </Sense>
  </LexicalEntry>
</Lexicon>
```

4.4.5.2 Filtered LMF example:

```
<Lexicon>
  <LexicalEntry id=“le_1”>
    <feat att=“partOfSpeech” val=“noun”/>
    <!--Lemma is obligatory in LMF and should be used to encode the morphosyntactic
    information applicable to the whole lemma -->
    <Lemma>
        <feat att=“writtenForm” val=“boy”/>
    </Lemma>
    <!--use feat “partOfSpeech” to set the PoS of the entry -->
    <feat att=“partOfSpeech” val=“noun”/>
    <Sense>
```

```
        <!-- add one feat for each class and its assigned score -->
        <!-- “boy” belongs to the class human but not to location or eventive -->
      <feat att=“event” val=“no”/>
      <feat att=“hum” val=“yes”/>
      <feat att=“loc” val=“no”/>
      <!-- “domain” is a label used to identify the corpus from where the lexicon has
      been extracted, in case it has been automatically acquiered -->
      <feat att=“domain” val=“labour”/>
    </Sense>
  </LexicalEntry>
  <LexicalEntry id=“le_2”>
    <feat att=“partOfSpeech” val=“noun”/>
    <Lemma>
        <feat att=“writtenForm” val=“car”/>
    </Lemma>
    <Sense>
        <feat att=“domain” val=“labour”/>
      <!-- “car” does not belong to the class human nor eventive, but the classifier is not
      sure about class location -->
      <feat att=“event” val=“no”/>
      <feat att=“hum” val=“no”/>
      <feat att=“loc” val=“unknown”/>
    </Sense>
  </LexicalEntry>
  <LexicalEntry id=“le_3”>
    <feat att=“partOfSpeech” val=“noun”/>
    <Lemma>
      <feat att=“writtenForm” val=“storm”/>
    </Lemma>
    <Sense>
      <feat att=“domain” val=“labour”/>
        <!-- “storm” belongs to the class eventive but not to location or human -->
      <feat att=“event” val=“yes”/>
      <feat att=“hum” val=“no”/>
      <feat att=“loc” val=“no”/>
    </Sense>
  </LexicalEntry>
</Lexicon>
```

4.5 Workflows

Within the PANACEA platform, Taverna workflows are designed to provide a seamless end-to-end solution for lexical acquisition. In the WP6 workflows, tools for creating lexica are chained together, demonstrating the ability to perform automatic lexical acquisition. The PANACEA lexical acquisition components were particularly designed to use technologies that are scalable and implementable in a distributed environment.

The following workflows relevant to WP6 lexical acquisition components have been developed. Workflows are also documented at http://myexperiment.elda.org. (Workflows related to lexical merger are presented in D6.4.). For each workflow listed here, the components included are also listed. In some cases the workflows start from raw text and in others from parsed data, where parsed corpora already exist or where parsing is relatively time-consuming and is assumed to have taken place offline or in a separate workflow. As an example, the diagram is given for some of the workflows.

4.5.1 SCF workflows

4.5.1.1 Dependency parsed 2 Italian SCF acquisition

http://myexperiment.elda.org/workflows/50

This workflow contains the Subcategorisation frames acquisition service for the Italian language. The service takes in input dependency parsed text. The output is a lexicon encoded according to LMF and serialized in an XML conformant to the LMF DTD rev.16.

Components: estrattore_scf

Provider: CNR

4.5.1.2 Dependency parsed 2 Italian MWE and SCF acquisition

http://myexperiment.elda.org/workflows/48

This workflow combines together two services running in parallel: the MWE acquisition and the Subcategorisation frames acquisition services, both for the Italian language. The two services take in input dependency parsed text. The output is two separate lexicons encoded according to LMF and serialized in an XML comformant to the LMF DTD rev.16.

Components: estrattore_mw_2, estrattore_scf, SCF

Provider: CNR

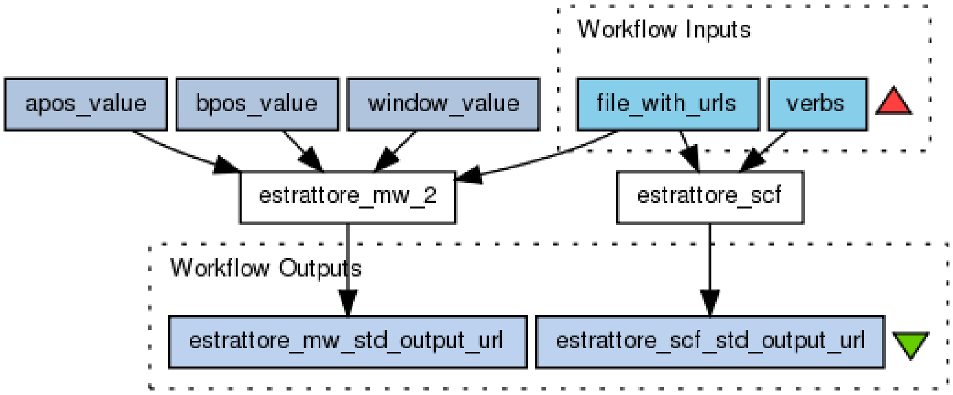


Figure 1: Dependency parsed 2 Italian MWE and SCF acquisition workflow

4.5.1.3 Rasp parsed to English SCF

http://myexperiment.elda.org/workflows/70

This workflow takes English data parsed with the RASP system and outputs a SCF lexicon. The default settings for the workflow are those used in PANACEA experiments, but all settings are customizable by the user.

Components: tpc_subcat_inductive

Provider: UCAM

4.5.1.4 Spanish SCF extractor from parsed corpus for a given list of verbs

http://myexperiment.elda.org/workflows/80

This workflow uses UCAM SCF extractor to extract the SCF for a set of given verbs. The input corpus must be already parsed. The parameters used in this workflow assume that the corpus has been tagged (or follows the format) whit Spanish malt parser webservice as deployed by Panacea

Components: tpc_subcat_inductive

Provider: UPF

4.5.1.5 Spanish SCF extractor from parsed corpus for all verbs appearing more than a given number of times in the corpus

http://myexperiment.elda.org/workflows/81

This workflow uses UCAM SCF extractor to extract the SCF in Spanish for all verbs appearing in the corpus more than a given number of times. The input corpus must be already parsed. The parameters used in this workflow assume that the corpus has been tagged (or follows the format) whith Spanish malt parser webservice as deployed by Panacea. This corpus is indexed with cqp to get all verbs appearing more than the given number of times in the corpus and then the list of these verbs is used to call the SCF extractor.

Components: cqp_index , cqp_query_all_verbs, columns_selector, tpc_subcat_inductive

Provider: UPF

4.5.2 MWE acquisition workflows

4.5.2.1 MWE lexicon extractor from text

http://myexperiment.elda.org/workflows/58

This is a workflow for acquiring a Lexicon of Multiwords for Italian from a corpus of web pages (in basicxces format). In PANACEA this corpus is the output of crawling workflows. NB. This wf assumed the crawling phase has already been performed and the result saved on a server.

Components: converter_to_plain, freeling_it, fc_freeling_text_2_conll_it, desr, estrattore_mw

Provider: CNR

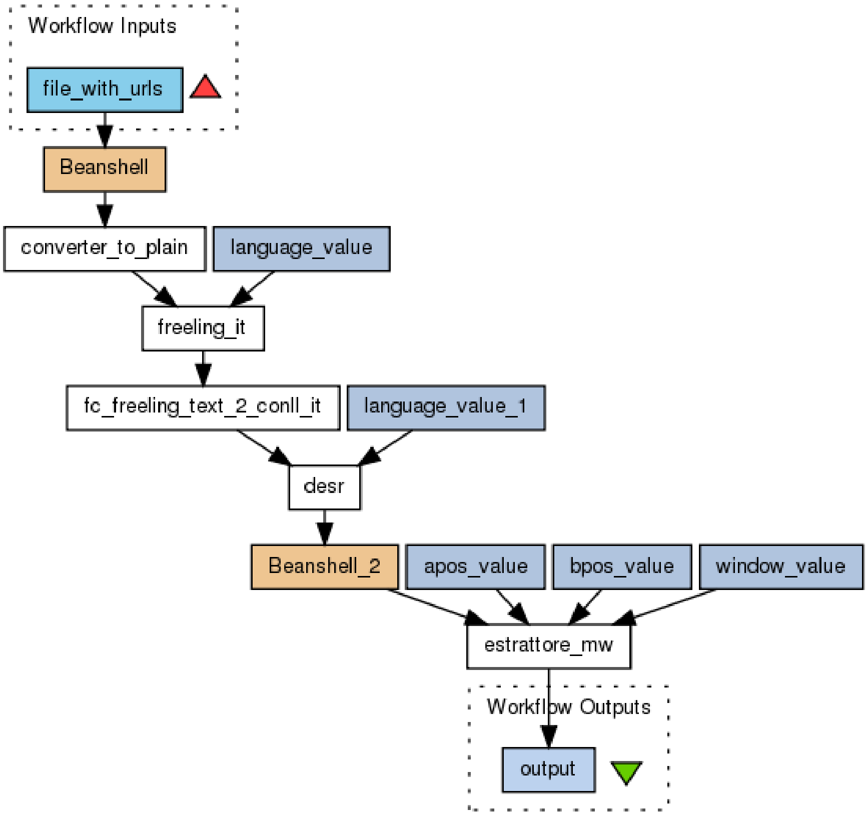


Figure 2: MWE lexicon extractor from text

4.5.2.2 Dependency parsed 2 Italian MWE and SCF acquisition

http://myexperiment.elda.org/workflows/48

See section 4.4.1.2

4.5.2.3 Multiword acquisition with post-filtering

http://myexperiment.elda.org/workflows/94

This workflow combines the language independent Multiword Extractor and 4 post-filters. The input is a pos-tagged corpus formatted in CoNLL (10 columns format). The output is a multiword form lexicon optionally represented in LMF-XML or in tabbed format.

4.5.3 Workflows for Lexical Semantic Classes Acquisition

4.5.3.1 Classification of nouns found in crawled data into lexical classes

http://myexperiment.elda.org/workflows/63

This workflow annotates with FreeLing the input crawled data (in TO1 format) and sends it to three different noun classifiers: event, location and human nouns. Each classifier produces a LMF output. The three obtained LMF files are merged into a single LMF lexicon containing information for all classes using the merging web service. This workflow works for English and Spanish, since those are the languages for which there are noun classifiers available.

Components: url_list_split, panacea_conversor, freeling3_tagging, dt_noun_classifier_human, dt_noun_classifier_location, dt_noun_classifier_enventive, merge_lmf_loc_event, merge_lmf_loc_event_hum

Provider: UPF

4.5.3.2 Classification of nouns in PoS tagged data for English and 7 available classes

http://myexperiment.elda.org/workflows/84

This workflow uses FreeLing annotated data to classify the given list of nouns with the different available noun classifiers for English. The LMF outputs of each classifier are merged into a single LMF lexicon containing information for all classes.

Components: dt_noun_classifier_eventive, dt_noun_classifier_human, dt_noun_classifier_location, dt_noun_classifier_abstract, dt_noun_classifier_artifact, dt_noun_classifier_matter, dt_noun_classifier_social, merge_list_of_lmf_files

Provider: UPF

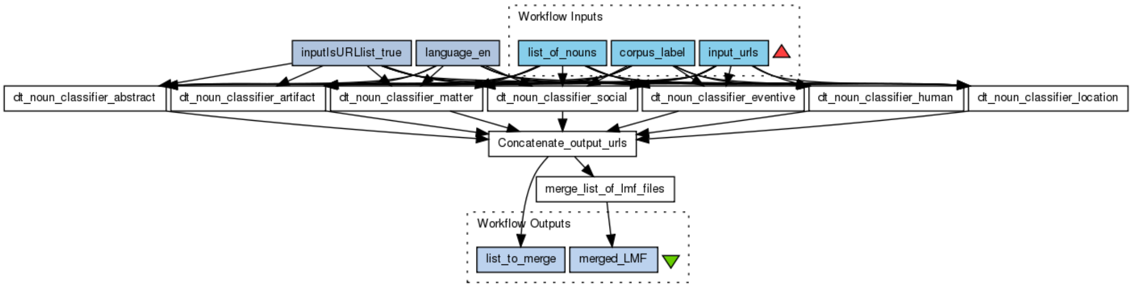


Figure 3: Classification of nouns in PoS tagged data for English and 7 available classes

4.5.3.3 Classification of nouns in PoS tagged data for Spanish and 9 available classes

http://myexperiment.elda.org/workflows/85

This workflow uses FreeLing annotated data to classify the given list of nouns with the different available noun classifiers for Spanish (9 classes). The LMF ouputs of each classifier are merged into a single LMF lexicon containing information for all classes.

Components: dt_noun_classifier_eventive, dt_noun_classifier_human, dt_noun_classifier_location, dt_noun_classifier_abstract, dt_noun_classifier_artifact, dt_noun_classifier_matter, dt_noun_classifier_process, dt_noun_classifier_semiotic, dt_noun_classifier_social, merge_list_of_lmf_files

Provider: UPF

4.5.3.4 Freeling tagging, weka creation and model training from crawled data

http://myexperiment.elda.org/workflows/67

This workflow annotates with FreeLing PoS tagger the input crawled data and creates a weka file using given regular expressions and gold standard. This weka file is used to estimate the bayesian parameters using the given priors. The output model can be used to classify new instances with a Naive Bayes classifier.

Components: url_list_split, panacea_conversor, freeling3_tagging, cqp_index, create_weka_noun_signatures

Provider: UPF

4.5.3.5 Freeling tagging, weka creation and classification for crawled data

http://myexperiment.elda.org/workflows/68

This workflow annotates with FreeLing PoS tagger the input crawled data and creates a weka file using given regular expressions and gold standard. This weka file is then classified using Naive Bayes classifier and the given model.

Components: url_list_split, panacea_conversor, freeling3_tagging, cqp_index, create_weka_noun_signatures, naive_bayes_classifier

Provider: UPF

5 Deliveries

The monolingual lexica built using the presented workflows are delivered in D6.3. See that document for details.

We also deliver (attached to this document) the scientific articles related to the work presented here and the Regular Expression sets to develop cue vectors for noun-lexical acquisition.

References

E. Agirre, K. Bengoetxea, K. Gojenola and J. Nivre. 2011. Improving Dependency Parsing with Semantic Classes. *In Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics*, ACL-HLT 2011 Short Paper, Portland, Oregon.

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.

P. Buitelaar, P. Cimiano, and B. Magnini. 2005. *Ontology learning from text: Methods, evaluation and applications*. Amsterdam: IOS Press.

M. Ciaramita, and Altun, Y. 2005. Named-Entity Recognition in Novel Domains with External Lexical Knowledge. In Workshop on Advances in Structured Learning for Text and Speech Processing (NIPS 2005).

G. Chrupala. 2003. Acquiring Verb Subcategorization from Spanish Corpora. DEA Thesis, University of Barcelona.

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

Ch. Fillmore, Narayanan, J. and Baker, C. 2006. What Can Linguistics Contribute to Event Extraction? In Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis, pp. 18-23.

G. Fu, G. 2009. Chinese Named Entity Recognition Using a Morpheme-Based Chunking Tagger. In Proceedings of the 2009 International Conference on Asian Language Processing, pp. 289-292.

Gil Francopoulo, Núria Bel, Monte George, Nicoletta Calzolari, Mandy Pet, and Claudia Soria. 2008. Multilingual resources for NLP in the lexical markup framework (LMF). *Journal of Language Resources and Evaluation*, 43 (1).

T. L. Griffiths, C. Kemp, and J.B. Tenenbaum. 2008. Bayesian models of cognition. In Ron Sun (ed.), *Cambridge Handbook of Computational Cognitive Modeling*.Cambridge University Press.

R. Grishman, C. Macleod, and A. Meyers. 1994. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of COLING 94*, 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.

G. Lee, G., Seo, J., Lee, S., Jung, H., Cho, B.H., Lee, C., Kwak, B.K., Cha, J., Kim, D., An, J-H., Kim, H. and Kim, K. 2001. SiteQ: Engineering High Performance QA System Using LexicoSemantic Pattern Matching and Shallow NLP. In *Proceedings of the TREC-10 Conference*. NIST, Gaithersburg, MD, pp. 437–446.

B. Levin. 1993. English verb classes and alternations: A preliminary investigation. Chicago: University of Chicago Press.

D. J. C. MacKay. 2003. Information Theory, Inference, and Learning Algorithms. *Cambridge University Press*, 2003. ISBN 0-521-64298-1

M. Marimon, B. Fisas, N. Bel, M. Villegas, J. Vivaldi, S. Torner and M. Lorente. 2012. The IULA Treebank. In *Proceedings of LREC 2012*, Istanbul, Turkey.

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

M. de Marneffe, Padó S. and Manning, C. 2009. Multi-word expressions in textual inference: Much ado about nothing?. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, Suntec, Singapore, pp. 1-9.

P. Merlo and S. Stevenson. 2001. Automatic Verb Classification based on Statistical Distribution of Argument Structure.*Computational Linguistics*, 27:3.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*. 3, 235-244.

S. Necsulescu, N. Bel, M. Padró, M. Marimon and E. Revilla: Towards the Automatic Merging of Language Resources. In P*roceedings of WoLeR 2011*. Ljubljana, Slovenia.

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

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 *Proceedings of ACL 2007*.

R. J. Quinlan. 1993. C4.5: Programs for Machine Learning. *Series in Machine Learning*.Morgan Kaufman, San Mateo, CA.

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

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

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

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

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