ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation
www.accurat-project.eu
Project no. 248347

Deliverable D2.3
Report on information extraction from comparable corpora

Version No. 1.0
September 01, 2011

Document Information

Deliverable number:	D2.3
Deliverable title:	Report on information extraction from comparable corpora
Due date of deliverable:	31/08/2011
Actual submission date of deliverable:	01/09/2011
Main Author(s):	Ahmet Aker, Robert Gaizauskas, Evangelos Kanoulas, Monica Paramita, Emma Barker, Paul Clough, Marcis Pinnis, Tatiana Gornostay, Radu Ion, Dan Stefanescu, Elena Irimia, Nikos Glaros, Marko Tadic, Dan Tufis
Participants:	USFD, Tilde, ILSP, FFZG, RACAI, LT
Internal reviewer:	RACAI
Workpackage:	WP2
Workpackage title:	Multi-level Alignment Methods and Information Extraction from Comparable Corpora
Workpackage leader:	RACAI
Dissemination Level:	PU, Public
Version:	V1.0
Keywords:	Named Entity and Terminology Extraction and Mapping, Machine Translation, Comparable corpora, Evaluation, Lexical Dictionaries

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
0.1	24/06/2011	Draft	USFD	Draft	Skeleton
0.2	25/07/2011	Draft	All partners	Contribution	For Internal Review
0.2	01/08/2011	Review	RACAI	Comments	Internal Review
1.0	31/08/2011	V1.0	USFD	Final Version	Submitted

Executive Summary

In this deliverable we report different tools for named entity and terminology extraction. For each of these tools we report its underlying algorithms and performance(s) on unseen data. Each tool is dedicated to a specific Accurat language. We also present tools we have developed for mapping between such linguistic terms (named entities, terminology) in pairs of texts from different Accurat languages. This mapping information is of use for identifying candidate translation units in comparable corpora which in turn can be used to improve statistical or rule-based MT systems. We run the mapper on texts written in two different languages and test its performance using human evaluation. Finally, we discuss how lexical dictionaries can be obtained from comparable corpora to improve the extraction and mapping of entities and technical terms.

Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Term/Definition
Accurat	Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation
CLIR	Cross Language Information Retrieval
EL	Greek
EN	English
FWL	Frequent Word List
GPE	Geopolitical Entities
idf	inverse document frequency
LCS	Longest Common Subsequence
LCSR	Longest Common Subsequence Ratio
LCST	Longest Common Substring
LL	Log Likelihood
ME	Maximum Entropy
MLE	maximum likelihood estimates
MT	Machine Translation
MUC	Message Understanding Conference
NDFA	non-deterministic finite state automaton
NE	Named Entity
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification
NLP	Natural Language Processing
POS	Part of Speech
SMT	Statistical Machine Translation
TE	Terminology Extraction
tf	term frequency

1 Introduction

ACCURAT has assembled various types of comparable corpora with a view to mining them for translation units which can be used to improve machine translation (MT) systems. Two types of such corpora are: (1) news reports on the same news event written in different languages and (2) technical texts in different languages within the same narrow domain.

One hypothesis of the ACCURAT project is that named entities within news reports and technical terms within narrow domain corpora can be identified and aligned and that such alignments will be useful for MT by: (1) supplying term translations that can supplement to any bilingual lexical resources, and (2) signalling that the larger textual units within which the aligned terms are embedded may be translations.

Many news events are reported in multiple languages. Such reports will vary in detail and their content is likely to be tailored to a specific reader group and may be influenced by the perspective the writer takes on the event. However, what will appear in all the different reports are many of the same named entities designating the key roles players in the event and the event's location. These named entities can be persons, locations or organizations, though the term is usually interpreted broadly enough to include many other entity types such as times, dates and monetary amounts.

Reports in different languages about the same event can be regarded as comparable because, while not direct translations of each other, they are likely to say some of the same things and hence are likely to share some textual units which are translations of each other. Named entities (NE) can play an important role in finding such textual units. For instance, if two sentences in reports about the same news event in different languages share several named entities then it is possible that these sentences are saying the same or similar things and hence contain some translation units. Sentences in different languages which do not share named entities are less likely to have such units. Furthermore, regardless of whether the named entities signal larger translation units, discovering translations of named entities is of value for MT in its own right, since they are an open class of term unlikely to be found in dictionaries and changing rapidly over time. Thus, identifying corresponding named entities across pairs of news texts about the same news event in different languages holds significant promise for improving MT. However, identifying named entities in texts of different languages which correspond to each other is a challenging task. First, it requires the **identification** of named entities in each text separately and then their correct **alignment** or **mapping**.

As with news texts written in different languages but reporting the same news event, technical texts written in different languages but within the same narrow domain are also likely to say some of the same things. Within narrow domains technical terms play an analogous role to named entities within news texts. That is, the same terms occurring in separate sentences from the two languages suggest that these sentences may contain some translation units, while those that do not contain such terms are less likely to contain such units. And again, regardless of whether shared technical terms signal larger translation units, identifying translations of terms is of value for MT in its own right, since similar to NE they are an open class unlikely to found in dictionaries and changing rapidly over time.

In this deliverable we report the tools developed to perform named entity and term identification and mapping. The named entity identification tools are described in Chapter 2. Tools for term extraction are reported in Chapter 3. We describe the mapping of such linguistic entities across languages in comparable texts in Chapter 4. We also extract lexical dictionaries from comparable corpora to improve mapping and information extraction (Chapter 5).

2 Named Entity Extraction (NER)

In the context of the ACCURAT project, Named Entity Recognition (NER) is useful at various levels of alignment of comparable corpora and in creating bilingual named entity lists that can enhance MT system performance. In addition, NER can be exploited in narrow domain comparable corpora collection tasks, especially those relying on focused crawling. In such systems NER (as well as term-extractors) can be used to semi-automatically represent a given narrow domain by means of its underlying terms and named entities both prior crawling (seed term/entity list creation) and while crawling (seed term/entity list expansion).

2.1 Related Work

Named Entity Recognition (NER), sometimes called Named Entity Recognition and Classification (NERC), consists in the identification and classification of phrases denoting entities of given categories (locations, persons, organizations, etc.) that are of importance to a range of language processing applications.

NER has been well studied since it emerged as an area of research in its own right in the mid-1990s. Initial approaches typically involved manually authored rule-based recognizers. However, the focus soon shifted to supervised learning techniques and these now dominate the field. Such methods were particularly well explored in addresing the shared tasks in two of the Computational Natural Language Learning (CoNLL) conferences, CoNLL-2002[1] and CoNLL-2003[2]. Both CoNLL-2002 and ConLL-2003 encouraged language-independent approaches to NER, yet each focused on two languages only: CoNLL-2002 focussed on Spanish and Dutch, while participants of ConLL-2003 were offered training and test data for English and German. The most common method used by participants in the CoNLL-2003 shared task was the Maximum Entropy (ME) Model approach. Bender et al. [2003], Chieu and Ng [2003] and Curran and Clark [2003] used the Maximum Entropy Model on its own, whereas, others [Florian et al., 2003; Klein et al., 2003] used ME methods in combination with other techniques.

For a detailed survey about different approaches in NER see Nadeau and Sekine [2007].

ACCURAT is not interested in NER as a research problem in its own right. Rather it is interested in NER because of its potential to assist in mining information from comparable corpora to assist MT. Therefore, existing NER techniques or even implementations have been used where possible to minimize development work in this area. Nevertheless, as NER tools were not available for several of the languages addressed in ACCURAT, NER systems have had to be developed in these cases. The remainder of this section details the approach taken to NER in each of the ACCURAT languages.

2.2 Greek NER

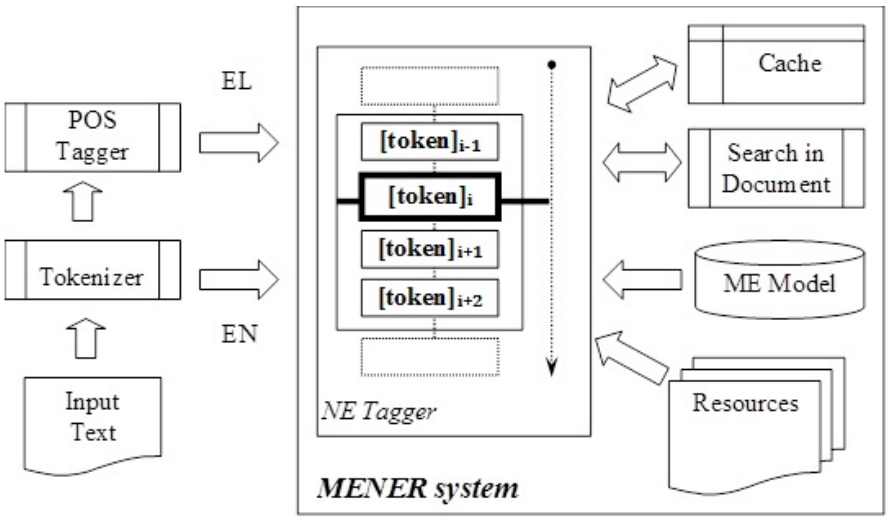
The Greek named entity (NE) recognition system (MENER) is a highly modified version of the system developed by Chieu and Ng [2003] that was the best-scoring system in the CoNLL-2003 shared task . MENER is a single-level maximum entropy approach that makes use of a broad range of features extending from conjunctive ones, that lend the system limited pattern recognition abilities, to individual ones, that indicate statistically important evidence extracted automatically from the training data. The main idea behind the system can be summarized as to maximize the probability $p(N|S,Doc)$, where N is the sequence of NE tags assigned to each

[1] http://www.cnts.ua.ac.be/conll2002/
[2] http://www.cnts.ua.ac.be/conll2003/

word in sentence S, and Doc is the relevant information that can be extracted from the whole document containing S. To achieve this goal, MENER combines sentence-based local evidence about words, with global evidence, which is a collection of features drawn from other occurrences of those words within the same document (global features). Other machine learning-based NER systems usually try to maximize the probability $p(N|S)$ only, or, often, make use of global data by incorporating a second-level classifier that tries to improve the output of its sentence-based predecessor.

The MENER system is language and domain independent (provided that adequate data are available), yet it was extensively trained and evaluated on Greek business news. A simplified architecture of the system and its main components is depicted in the block diagram in Figure 2.1.

Figure 2.1: **MENER architecture.**



Initially, using a sliding window of four tokens (comprising $[\text{token}]_i$, or the focus token, the preceding ($[\text{token}]_{i-1}$), and the next two tokens, if any), the NE-Tagger scans the entire input XML file (output of POS tagger). It collects all contextual features applicable to the focus token by consulting search and cache units appropriately along with the system's resources and submits the resulting set of features to the Maximum Entropy (ME) model for evaluation. The response of the ME model is a probability distribution over the classes it has been trained on. Then NE-Tagger selects the class with the maximum probability to be assigned to the focus word and restarts the processing cycle by sliding the window to the next token.

Additionally, by injecting its previous decisions, as an additional feature, into contexts that will be evaluated in the future, the system cache unit accumulates already recognized NE tokens along with the tags assigned, in an effort to circumvent the lack of memory, that is inherent in every ME model. The cache unit is emptied only when processing reaches the end of each tagged document.

The search unit is actually the provider of global features. Using document-wide searches, it examines, for each token, the context of its other occurrences looking for trigger words, contained in system's resources, and other clues (e.g. capitalisation information). Extracted features from each occurrence are united to form a set, which, as it grows, keeps only the high ranking features, in order to eliminate weak evidence.

Linguistic resources of the system are automatically extracted from the training data during the pre-training stage, using a separate tool. For each name class the following set of feature lists are compiled (Table 2.1):

Table 2.1: **Features extracted for each name class.**

feature	description
unigrams	single words that precede the name class
bigram	bigrams of words that precede the name class. In order to keep the strongest evidence only, this list includes bigrams with higher probability of appearing before the name class than the contained unigram itself (for example, "city of" vs. "of", the first one appears more often before locations than the other)
post unigrams	words that succeed the name class
suffixes	three letter suffixes of words pertaining to the name class
prefixes	three letter prefixes of words pertaining to the name class
terminal tokens	tokens that terminate the name class, for example the organization class often terminates with tokens such as '1Inc." and "Corp."
functional words	lower case words or punctuation symbols that occur within the name class, for example "van der", "of", etc.

Apart from name classes, the system also consults a Frequent Word List (FWL) that consists of words occurring in the training data over a given threshold. This is set to 4 for the EN data and 3 for the EL model, and is used to determine the rareness of a word, a fact that is reflected as an extra feature. Additionally, MENER may, optionally, make use of an external knowledge base in the form of lists with line-delimited records of known names (per name class), compiled from a variety of sources (Internet, CoNLL 2003 shared task data, annotated data, etc.).

In addition to FWL and name lists, all other lists are sorted according to the ascending order of the correlation coefficient [Ng et al., 1997] C of an item w in relation to a name class NC, which is defined as:

$$C = \frac{(N_{r+} * N_{n-} - N_{r-} * N_{n+}) * \sqrt{N}}{\sqrt{(N_{r+} + N_{r-}) * (N_{n+} + N_{n-}) * (N_{n+} + N_{r+}) * (N_{n-} + N_{r-})}} \tag{2.1}$$

where N is the total number of sentences in training data, N_{n-} (N_{n+}) is the number of non-relevant sentences in which the item w does not occur and which contain no (at least one) token of NC class; N_{r-} , N_{r+} refer to the number of relevant sentences which do include item w either under right conditions that meet its meaning or not, respectively. For example, in case of a unigram or bigram item w, N_{r+} refers to the number of sentences in which w actually precedes an instance of NC class. Correlation coefficient is a variant of the χ^2 metric and can be characterized as a "one-sided" χ^2 metric. It selects exactly those items that are highly indicative of membership in a category, whereas χ^2 will also pick items that are indicative of non-membership in the category.

2.2.1 NE Annotation Scheme

The classification schema chosen is compatible with the ACE[3] (Automatic Content Extraction) schema, which supports the recognition and classification of the following types of NEs: person (*PER*), organisation (*ORG*), location (*LOC*) and geopolitical entity (*GPE*). Moreover, NE's of the type LOC were also assigned a subtype value, namely: location (LOC), geographical region (GEO) and facility (FAC). Our classification schema retains most of the types and

[3] Annotation Guidelines for Entity Detection and Tracking (EDT) Version 4.2.6 20040401.

subtypes provided for by ACE, yet, it attempts to disambiguate between LOC and GPE usage of names. At the same time an extended annotation schema with subtypes for further classifying the spotted NE's has been introduced; however, only subtypes of LOC entities have been used. Moreover, to sustain consistency in annotation, a condensed classification schema has been finally used. To this end, mappings of the extended classification schema have been performed: GPE has been mapped on LOC, and LOC sub-classification has been dropped. A short description of our NE extended annotation schema and relevant guidelines is provided below:

PERSON: Names of individuals, family names and widely used aliases or nicknames of people are marked as NE's of the type PERSON. Similarly, proper names that refer to saints or dead people are also marked as PERSON NE's unless they are used to name other entities (i.e., ships, churches, locations, prizes or awards, etc.). Within the current schema, occupations, titles, honorific expressions that usually precede a name are not considered as part of the markable NE. Examples:

```
President [person Borrell Fontelles /person]

The annotation schema has also provided for the following
subtypes of the type PER:

PER.human: Names of people, either dead or alive are
further classified as human: Mr Ortuondo Larrea

PER.animal: Names of animals fall into this subtype:
Morris the cat.

PER.fictional: Names of fictional characters are tagged
as PER.fictional: Spiderman is children's hero.

PER.other:  All other animate entities that do not fall into
the above subtypes are to be tagged as PER.other.
```

ORGANISATION: Companies, enterprises, organisations or groups of people with an organisational status fall within this category and are marked as NEs of the type ORGANISATION.

```
the [org Iraqi government /org]

On behalf of the Group of the [org European People's Party /org]
and [org European Democrats /org]

The annotation schema has also provided for the following
subtypes of the type ORG though they have not been applied
to the data:

ORG.commercial: A commercial organization is focused primarily
upon providing Ideas, products, or services for profit, such as
industries, industrial sectors, etc.

ORG.educational: Institutions focused primarily upon the
```

```
furthering or promulgation of learning fall into this sub-class.

ORG.other: All other organizations that do not fall into the
above subclasses.
```

LOCATION: Proper names that designate landmarks are marked as being of the type LOCATION.

```
The following subclasses have been used at the annotation:

LOC.geo: Geographical entities, that have been created
naturally upon or above the surface of the earth, such as
mountains, masses of water, etc.

LOC.loc: Geographical regions that do not pertain to
the above class. Contextual information is used to
distinguish a LOC.loc from a GPE entity.

LOC.fac: Large functional man-made constructions are
facilities, that is artifacts that fall under the domain
of architecture and civil engineering. Contextual
information is used to distinguish a LOC.fac from an ORG entity.

LOC.other: Other entities that are used to designate a
space fall into this class, such as stars, planets, etc.
```

GEOPOLITICAL ENTITY: Geopolitical Entities (GPE) are geographical regions also defined by political and/or social groups. Following ACE specifications "A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people". In our schema, however, context has guided disambiguation between LOC and GPE uses of names:

- I especially want to welcome the arrival of [gpe Cyprus /gpe]
- She visited [loc Cyprus /loc]

GPE entities are further classified with the following subtypes of the type GPE:

- GPE.continent
- GPE.nation
- GPE.province
- GPE.city
- GPE.other

It should be noted that markable entities appear in the text with their full-name, an abbreviated/reduced form of this name, or a word/phrase - usually a metonymy - consistently used to describe it, and all these alternative mentions are tagged. However, simple pronominal or nominal references to NEs are not marked.

```
[org Athens Stock Exchange /org] - [org ASE /org]
President [person Borrell Fontelles /person] - the
president said (it is not marked)
```

Furthermore, words that usually precede a name (articles, modifiers, etc) are not to be included within the markable NE:

```
the [org Iraqi government /org]
```

NEs that are connected through part-whole and possessor-possessed relations are not marked as a single entity:

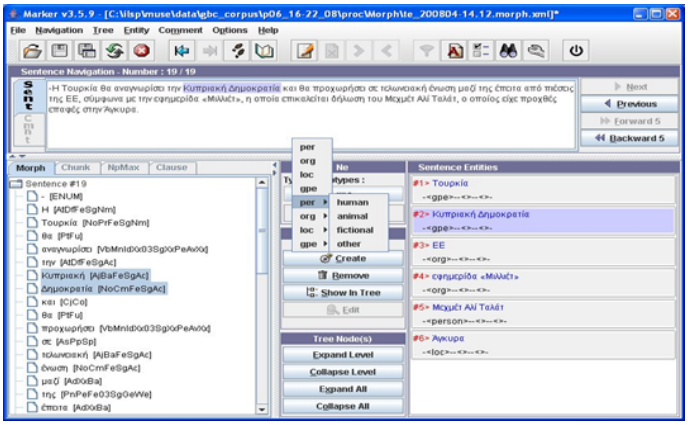
```
The [org Research Department /org] of [org Egnatia Securities /org]
```

2.2.2 Evaluation

Training and evaluation of ILSP's NERC system has been performed on EL and EN textual data pertaining to the domains of Politics and Travel. Newswire texts from various resources have also been included in the training and testing material. However, training of the models for the aforementioned languages and domains has not been solely based on the above data. The training material used for the English news model was also coupled with the English data provided at the CoNLL-2003 shared task. The material is part of the Reuters corpus. It consists of three parts: the core data (204k tokens), a development set (51k) and a test set (46k), all provided in a line-delimited textual format that we converted into an equivalent XML representation. During system development and feature selection, only core data were used for training, whereas, for the final model, training was performed on the entire corpus (total of 301k). Additionally, the training corpus of the Greek news model is an agglomeration of mainly news documents collected from the following sources: various Internet news sites (9k tokens), the Greek Business Channel (GBC), a media provider in placecountry-regionGreece (107k) and data that originate from the European Parliament (EP) web site (54k). GBC data consists of short daily news bulletin files, covering economical, domestic and world news, which briefly mention 6-8 events per file. Despite their size, they are full of named entities and hence are valuable for the NERC task. EP files are normalized transcripts of European Parliament sessions.

A 63K tokens EN corpus and a 16k tokens EL corpus in the News – Politics domain, along with two other coprora in the Travel domain (7k for EL and 4k for EN) have been hand annotated by using *Marker*, an ILSP's GUI for multi-level corpus annotation (see Figure 2.2 for a typical screenshot of *Marker*).

Figure 2.2: **NE Annotation session with Marker on Greek data.**



The above textual data were enriched by other sources as well. To estimate NERC system performance, we have used the evaluation software provided for the CoNLL2003 shared task available from their web page (http://cnts.uia.ac.be/conll2003/ner/).

System performance has been estimated on the basis of two metrics that are widely employed in Information Retrieval, namely *Precision* and *Recall* measures defined as follows:

$$Precision = \frac{CorrectIdentifiedInstances}{AllIdentifiedInstances} \tag{2.2}$$

$$Recall = \frac{CorrectIdentifiedInstances}{AllInstances} \tag{2.3}$$

Overall performance of the system is given by the *F-measure metric,* which combines Recall and Precision in a single efficiency measure being the *harmonic mean* of Precision and Recall:

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{2.4}$$

Evaluation Results: For evaluation purposes, we have taken CoNLL2003 best performing system reported in Chieu and Ng [2003] as a baseline. System performance for the English dataset is presented in Table 2.2.

Table 2.2: **CoNLL2003 Baseline system evaluation.**

Conll'03 Evaluation	**NE**	**Precision**	**Recall**	**F-measure**
Train : Core (203.621) + dev_set (51.362)	LOC	90.29%	90.89%	90.59
Test : test_set (46.435)	MISC	77.04%	76.50%	76.77
Iterations: 500	ORG	81.53%	86.09%	83.75
Cut-off: 1	PERSON	94.32%	95.42%	94.87
	Overall	**87.17%**	**88.99%**	**88.07**

ILSP's NERC system outperformed the system developed by Chieu and Ng for ConLL2003 in the Greek Politics and News Politics data rendering an F-measure of 94.87 (Table 2.4). However, it performed rather poorly in the English data that pertain to the same domain and text type (Table 2.3). This is attributed to the fact that the English model was extensively trained on the Reuters data, whereas evaluation was performed on data of rather different structure. Indeed, the latter data consist of data collected from web resources (these are similar in structure to the Reuters data set), as well as of texts originating from the Europarliament (EP), which exhibit peculiarities (i.e., text and period structure, headlines and capitalisation conventions, etc.) that were not taken into account at the training phase. In the following examples, the system has erroneously recognised the words *Members* and *Rules* as being a NE of the type ORG:

```
... being very similar to the vote of the <org Members /org> of the
<org European Parliament /org>
... pursuant to <org Rules /org> 130 and 131 of the <org Rules /org>
of Procedure
```

In respect to the Travel data, ILSP's NERC system performs well in PER and LOC classes, but behaves poorly in class ORG (Tables 2.4 & 2.5). This is attributed to the fact that the texts at hand allow for a large degree of ambiguity between ORG-PER and ORG-LOC that cannot effectively be resolved solely on the basis of statistical information, unless further contextual knowledge is taken into account:

```
Vasilis and Eliza Goulandris Foundation organizes every summer at the
<loc Museum of Modern Art /loc> exhibitions
The <org Museum of Modern Art /org> in Andros organized the exhibition.
```

Quantitative results are given in the Tables 2.3 to 2.6:

Table 2.3: **Results for the English politics and news politics data.**

Politics_EN Evaluation	**NE**	**Precision**	**Recall**	**F-measure**
Train: Core(203.621) + dev_set(51.362) + test_set(46.435)	LOC	69.15%	84.75%	76.16
Test: Web(9.401) + EuroParl(53.736)	ORG	47.14%	87.54%	61.28
Iterations / Cut-off: 400 / 1	PERSON	81.37%	73.38%	77.17
	Overall	**59.17%**	**82.70%**	**68.98**

Table 2.4: **Results for the Greek politics and news politics data.**

News_EL Evaluation	**NE**	**Precision**	**Recall**	**F-measure**
Train: EuroParl(49.289) + EL Business(97.775) + Web(8.350)	LOC	95.88%	93.40%	94.62
Train: EuroParl(49.289) + EL Business(97.775) + Web (8.350)	ORG	90.65%	94.61%	92.59
Iterations / Cut-off: 400 / 1	PERSON	99.02%	98.54%	98.78
	Overall	**94.82%**	**94.92%**	**94.87**

Table 2.5: **Results for the English Travel data.**

Travel_EN Evaluation	**NE**	**Precision**	**Recall**	**F-measure**
Train : 42.313 tokens	loc	69.68%	78.97%	74.04
Test : 4.295 tokens	org	20.00%	14.29%	16.67
Iterations / Cut-off : 350 / 1	per	63.33%	65.52%	64.41
	Overall	**67.97%**	**75.32%**	**71.46**

Table 2.6: **Results for the Greek Travel data.**

Travel_EL Evaluation	**NE**	**Precision**	**Recall**	**F-measure**
Train: 64.632 tokens	loc	71.97%	87.37%	78.93
Test: 7.148 tokens	org	35.00%	25.00%	29.17
Iterations / Cut-off: 400 / 1	per	66.67%	43.14%	52.38
	Overall	**70.23%**	**78.80%**	**74.27**

2.3 Romanian NER

The Romanian name entity extractor, also called NERA, works on Romanian and English language texts and allows as input either a raw bi-text or a preprocessed bi-text. For the former, the NERA has access to a preprocessing web service which can pos-tag and lemmatize the bi-text. According to the morpho-syntactic descriptors, the lemmas and a lexicon, the extractor determines the type of the lexical units.

NERA distinguishes between three main classes of named entities: PERSON, ORGANIZATION and LOCATION. Moreover, it can further refine the LOCATION class with subclasses

like: COUNTRY, CITY or any other user-defined subclass, if corresponding gazetteers exist for the languages in discussion. The assignment of a certain class to a named entity is performed by employing a combined approach of rule-base and statistical methods. The statistical method is a Maximum Entropy Classifier that takes into account context features when classifying the named entities. It is used only if none of the rules can be applied based on the idea that there is no need to guess when one is certain.

We also evaluated the performance of NERA using manual annotated news articles. We used 100 news articles in Romanian and 82 articles in English and manually annotated them with the PERSON, ORGANIZATION and LOCATION types. We run NERA on the original articles and compared the tags produced by it with the manual ones. The results of the comparison are shown in Table 2.7.

Table 2.7: **NERA performance on Romanian and English texts.**

	Romanian	English
Boundary Identification Precision	69.46%	51.06%
Boundary Identification Recall	57.53%	52.96%
Boundary Identification F-measure	62.94%	52.00%
Type Identification Precision once the boundary has been correctly identified	92.34%	87.73%

2.4 Latvian, Lithuanian NER

Currently the dominant approach to developing named entity recognition systems is supervised learning. As under-resourced languages, such as Latvian and Lithuanian, for instance, do not have large annotated corpora available to train supervised systems, a semi-supervised approach is used, which requires an annotated seed list and a large un-annotated data corpus. The system's performance is lower than can be achieved by a supervised system, but for under-resourced languages the semi-supervised learning approach is the best current choice.

The Latvian-Lithuanian named entity recognizer is based on the Stanford NER Conditional Random Field named entity recognition system [Finkel et al., 2005]. As Stanford NER is a supervised learning system, a bootstrapping system and data pre-processing and post-processing system has been developed around the classifier. The system design is shown in Figure 2.3.

For both languages seed list, development data and test data has been annotated. The corpora for both languages consists of IT news, general news and Wikipedia articles (in equal proportions). The annotated corpora statistics for both languages is shown in Table 2.8. The annotation for each corpus has been done by two annotators and a judge (third annotator). The third annotators task was to disambiguate ambiguous cases (where both previous annotators disagreed) and correct obvious (to the human annotator) mistakes. For faster annotation a tool called *NESimpleAnnotator* (runs only on the Windows operating system) has been developed (for a user manual and annotation guidelines refer to the documentation of deliverable D2.6).

Once the annotated corpora was created, the seed list and test data had to be pre-processed (POS-tagged and lemmatized) as the bootstrapping system requires pre-processed data in a tab separated format. The workflow contains scripts that allow the pre-processing of annotated texts using Tilde's POS-tagging web services for Latvian and Lithuanian.

Figure 2.3: **Named Entity Recognition and Classification System's bootstrapping workflow for Latvian and Lithuanian languages.**

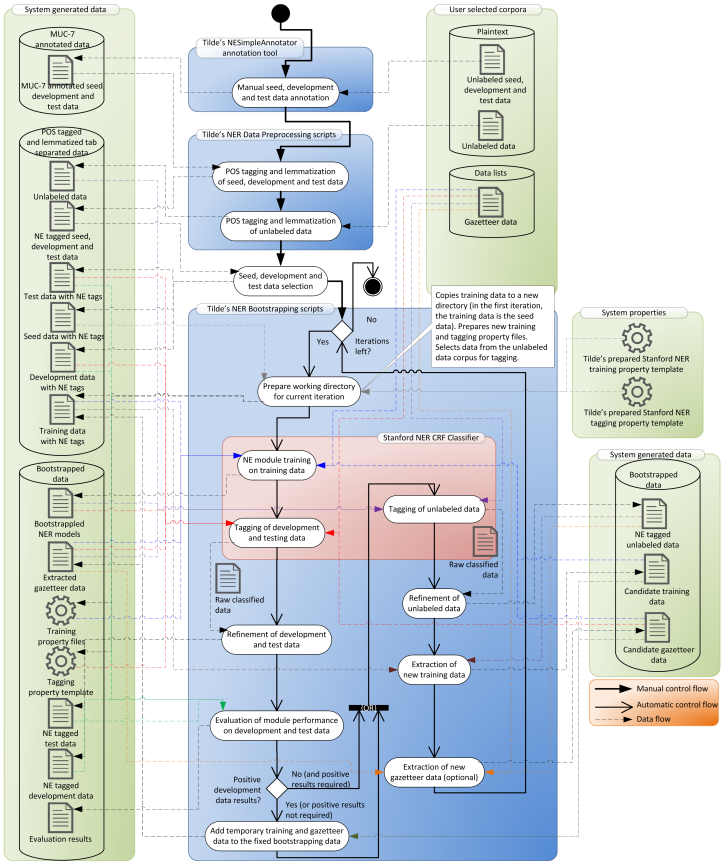


Table 2.8: **Latvian and Lithuanian corpus and annotation statistics.**

	Latvian	Lithuanian
Document count		
Seed list	40	37
Development data	25	33
Test data	66	55
Total	131	125
Word count		
Seed list	20959	18852
Development data	10053	17827
Test data	41208	36239
Total	72220	72918
Named entities		
Organization	1649	1118
Person	1040	975
Location	2614	2119
Product	866	873
Date	1590	1556
Time	353	233
Money	289	610
Total	8401	7484

The bootstrapping system iteratively (and automatically):

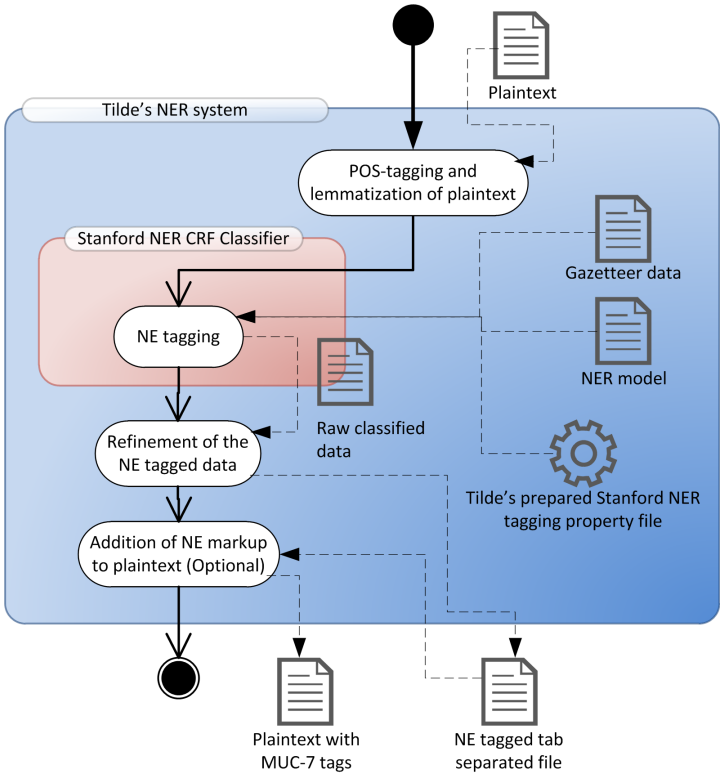
1. trains a NER model using training data (in the first iteration the training data is the seed list).
2. evaluates the trained model on development and test data. This step requires development and testing data to be tagged using the new NER model and applying refinements of the classified data to improve classification results. If the configuration parameters require, the system also assesses, whether the current iteration increases development data tagging results over the best previous iteration. It is possible to define, which of the evaluation results (precision, recall, accuracy or f-measure) is used in order to assess whether an iteration gives an increase.
3. tags the unlabelled data. If the configuration requires, the NER model of the best iteration is used in tagging unlabelled data. In the case of a negative iteration, the NER model of the previous positive iteration is used.
4. extracts new training data from the newly tagged unlabelled data.
5. extracts new gazetteer data from the newly tagged unlabelled data (this step is optional and can be skipped).

The NER system's workflow allows applying refinements to Stanford NER classified data. The refinements increase either recall or precision; therefore, it is possible to tune the system if it is necessary to cover more entities or it is necessary to extract more reliable entities (for instance, in gazetteer extraction). The available refinements are:

1. removal of unlikely tagged entities (low probability entities).
2. consolidation of equal lemma sequences (one sense per discourse; misclassified entities are re-classified). For instance, if a token sequence "Latvijas Republika" in two instances within a document has been tagged as "ORGANIZATION" and in one instance as "LOCATION", the refinement method tries to guess, which NE category out of the two is the dominant, and re-classifies other entities with the dominant entity's category). The algorithm is statistical and takes into account the total number of the token sequences (the NE, which is in the process of disambiguation), the number of token sequences for each individual NE category as well as the classifier's assigned NE probabilities for each token sequence.
3. forcing equal lemma sequences to be tagged (non-entities are classified as entities). For instance, if a token sequence is tagged as an "ORGANIZATION" with a high average probability (above a specified threshold) and the token sequence is not tagged with a different NE category elsewhere in the document, the method tries to find equal non-tagged token sequences and re-classifies them as "ORGANIZATION".
4. named entity border correction for entities, which contain an uneven number of quotation marks or brackets.
5. removal of full web addresses (containing protocol symbols "://") from named entities.
6. removal of named entities, which contain more than an allowed number of predefined strings (for instance, a person NE is not allowed to contain more than one "/" symbol).

In normal setting, however, users won't have to train their own NER models; therefore, the workflow contains scripts for plaintext or pre-processed data named entity tagging. The plaintext tagging workflow is shown in Figure 2.4.

Figure 2.4: **Named Entity Recognition and Classification System's plaintext tagging workflow for Latvian and Lithuanian languages.**



The provided scripts allow the results (NE tagged data) to be saved in a MUC-7 style[4] annotated plaintext or a tab-separated document containing POS-tags, lemmas and NE tag probabilities assigned by the Stanford NER CRF classifier. If the user provides a tab separated file for NE tagging, the pre-processing stage can be skipped.

2.4.1 Performance

The NER system has been designed to recognize seven named entity types - *organization, person, location, product, date, time* and *money*. It is possible to add more categories by updating the workflow and data pre and post-processing scripts (see deliverable D2.6 documentation for more details). The test data results of the trained models for Latvian and Lithuanian using bootstrapping are shown in Table 2.9.

The baseline system is a supervised system that is trained only on the seed list data. The next system is the baseline system, but tuned (using refinements) for the highest precision. The third system is the baseline system, but tuned for the highest F-measure. The tuning has been done using development data, but the results shown are on test data. The last two systems are the final bootstrapped systems. The first one is the bootstrapped system for better precision, which means that the system has been bootstrapped with new training data refined with precision increasing refinements. We present this system only for Latvian as for Lithuanian this approach did not increase precision over the baseline refined system. The second is the bootstrapped system for better F-measure (using F-measure increasing refinements). We present systems tuned for different performance scores, because in single document NE tagging tasks usually better precision is preferred, but for NE mapping better recall is of a greater importance.

[4]See http://www-nlpir.nist.gov/ for documentation of the MUC-7 named entity markup

Table 2.9: **Latvian and Lithuanian NER results.**

System	Precision	Recall	Accuracy	F-Measure
Latvian baseline				
Token	75.08	56.65	91.12	64.58
Full NE	62.88	48.23	-	54.59
Latvian baseline tuned for precision				
Token	83.54	46.94	89.71	60.11
Full NE	73.03	41.04	-	52.55
Latvian baseline tuned for F-measure				
Token	77.35	56.35	91.12	65.20
Full NE	65.64	48.94	-	56.07
Latvian bootstrapped for better precision				
Token	83.83	48.34	90.11	61.32
Full NE	73.78	43.78	-	54.95
Latvian bootstrapped for better F-measure				
Token	76.10	57.52	91.24	65.52
Full NE	64.89	50.92	-	57.06
Lithuanian baseline				
Token	74.44	63.54	92.30	68.56
Full NE	67.42	58.60	-	62.70
Lithuanian baseline tuned for precision				
Token	84.04	53.74	91.53	65.56
Full NE	77.01	49.63	-	60.36
Lithuanian baseline tuned for F-measure				
Token	76.31	63.50	92.47	69.32
Full NE	68.57	59.39	-	63.65
Lithuanian bootstrapped for better F-measure				
Token	76.90	63.77	92.42	69.72
Full NE	71.32	59.91	-	65.12

The results were calculated using the following formulas (accuracy is calculated only on the token level):

$$Precision = \frac{RelevantRetrieved}{AllRetrieved} \tag{2.5}$$

$$Recall = \frac{RelevantRetrieved}{AllRelevant} \tag{2.6}$$

$$Accuracy = \frac{RelevantRetrieved + Non - relevantNon - retrieved}{AllTokens} \tag{2.7}$$

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{2.8}$$

In the evaluation a token is considered *relevant retrieved* only if the retrieved token's NE category is equal to the gold-annotated token's NE category. *All retrieved* tokens are only those, which are classified as NEs in the retrieved data, and *all relevant* tokens are those, which are classified as NEs in the gold-annotated data. A token is considered *non-relevant non-retrieved* if the retrieved token is a non-entity and also the gold-annotated token is a non-entity. For

full NEs a NE is considered *relevant retrieved* only if the retrieved NE's borders (first and last token) match with the gold-annotated NE's borders and retrieved NE's all token categories are equal to gold-annotated NE's all token categories. A full NE is a multiple token sequence where the first token is classified with a *B-* category, for instance *B-ORG*, *B-PERS*, etc. and the following tokens are classified with an *I-* category, for instance, *I-ORG*, *I-PERS*, etc. This means that for full NE's border mismatching of even a single token (or NE category mismatching) is considered as a negative result. The token level accuracy is given to show the entire system's reliability not only on NE's, but also on other tokens (non-entities) and, therefore, tells us how many of all tokens have been correctly classified. For systems like NE-mapping systems higher accuracy might be more relevant than high precision, therefore this measure has been included as well.

2.4.2 Adding Support for Other Languages

The Latvian and Lithuanian named entity recognition and classification system has been developed specifically for Latvian, but also trained for Lithuanian. As the system provides a bootstrapping workflow, training models and adding new languages to the supported language list can be easily done using only one script. For new languages to be added, the system requires seed, development, testing and unlabeled data corpora. The corpora amount depends on the target language characteristics and the desired/required system performance that has to be achieved.

The workflow also provides a data pre-processing script (providing the named entities in the plaintext are marked following MUC-7 guidelines) that can POS-tag and lemmatize the plaintext. An interface with TreeTagger is also provided and it is possible to add integration with other POS-taggers following the given guidelines. For more information on how to train a new system using Tilde's NER workflow refer to the deliverable D2.6 of the Accurat project.

2.5 Croatian NER

The Named Entity Recognition and Classification tool in the Croatian language is presented in Bekavac and Tadić [2007]. It is a rule-based system that utilises a local grammars approach which is composed of a module for sentence segmentation, an inflectional lexicon of common words, an inflectional lexicon of names and regular local grammars for automatic recognition of numerical and temporal expressions. After the first step (sentence segmentation), the system attaches to each token its full morphosyntactic description and appropriate lemma and additional tags for potential categories for names without disambiguation. The third step (the core of the system) is the application of a set of rules for recognition and classification of named entities in already annotated texts. Rules based on described strategies (like internal and external evidence) are applied in cascade of transducers in defined order using the Intex development environment. Although there are other classification systems for NEs, the results of our system are annotated NEs which follow the MUC-7 specification. The system is applied to Croatian informative and noninformative texts and results are compared. The F-measure of the system applied to informative texts is over 90%.

3 Terminology Extraction (TE)

Terminology extraction is a subtask of Information Extraction which refers to extracting terms from a given corpus, relevant to the genre / domain of the corpus. Automated terminology extraction methods could be proved helpful in various ways for ACCURAT objectives, especially in the alignment and collection of narrow domain comparable corpora tasks. Alignment benefits occur, e.g. when trying to identify similar documents/paragraphs/sentences/phrases out of collected comparable corpora, by comparing their terminological content based on term-extractor tools and pre-existing bilingual terminological resources. Conversely, such bilingual term lists could be semi-automatically enriched by taking advantage of the output of generic alignment and parallel information extraction algorithms of WP2 being applied to comparable corpora. Automatic term extraction is also useful when semi-automatically defining the topic for narrow domain comparable corpora collection. In the sections that immediately follow, both state-of-the-art notes for the general issue of terminology extraction and a description of different TE systems developed for Accurat are given.

3.1 Related Work

We can view terms as the linguistic realisation of a domain specific concept, usually lexicalised in the form of a noun phrase. Term grammars and statistical tools are often used in systems for TE. A term grammar (usually a context free grammar that is applied to text with morphological and shallow syntax annotations) extracts all recognized phrases as term candidates [Bourigault, 1992]. Statistical tools used are similar to the ones developed in the field of information retrieval and text indexing. These tools include frequency counting, formulas from information theory, formulas that take into account the context of words, etc. [Church and Hanks, 1990; Frantzi and Ananiadou, 1997].

There are important differences between these two lines of action. A term grammar describes the syntactic structure that a valid term must satisfy, but it is possible that phrases recognised by the grammar are not valid terms. The weakness of a grammar is attributed to the fact that its rules, although a subset of NP (noun phrase) rules, are general enough to generate a large number of potential terms. Furthermore, a grammar cannot locate single word terms since such a term does not have any syntactic structure except part-of-speech information. In general, a term grammar can only produce a set of potential terms that remain to be validated by an expert or a module of different nature.

The statistical approach is based on the assumption that words and phrases indicative of the domain of a document tend to appear frequently (the same applies for phrases consisting of words that appear frequently together). Frequency can have two different interpretations: (1) a phrase is more frequent in the current text than in a representative collection of texts belonging to its domain and (2) a phrase is more frequent than others in the same text. Based on this "competitive" conception of frequency, each phrase is assigned a score representing its significance, (not taking into account functional words). Phrases at the top of this ranking have the highest probability of being valid terms. This method can extract single-word terms as well as multi-word terms. On the other hand, it cannot locate terms which do not satisfy the statistical criteria, i.e. they are not frequent enough. This is partly due to the fact that it is difficult to draw the line between middle frequency and high frequency. Finally, the selected statistical formula can affect the performance of extraction in the same way that the selected rules of the grammar, i.e. its syntactical coverage, affect the performance of the grammatical method. Statistical pro-

cessing is often combined with linguistic modelling in hybrid methodologies [Dunning, 1993a; Dagan and Church, 1994; Daille, 1996, 1994; Georgantopoulos and Piperidis, 2000b]. These systems initially construct a candidate term list using a term grammar and then filter this set through statistical techniques in order to remove syntactically acceptable phrases that are not "frequent" enough to be assigned valid termhood.

In their C/NC-value approach, Frantzi et al. [2000] combine pattern matching based on POS information, with a stop word list to extract (possibly nested) term candidates. These candidates are then filtered according to their C-value, a measure that combines statistical characteristics of the candidate string, that is, for a candidate string (cs) a) the total frequency of occurrences of cs in the corpus, b) the frequency of cs as part of other longer candidates, c) the number of these longer candidates, and d) the length of cs in words. NC-value, an extension of the basic method takes into consideration context via ranked lists of automatically extracted context words (adjectives, nouns, and verbs) of small sets of manually identified terms. Authors evaluated their method on an English medical corpus of eye pathology reports (800000 words). According to the authors there is a lack of precise criteria on termhood. Moreover, since it is often hard to find experts for term annotation, they decided to calculate "relevant" rather that "absolute" precision and recall, i.e. they compared the results of the C-value method and the results of extracting terms by frequency of occurrence. They report a 97.47 relative recall and a 31% precision for both methods, since the NC method just re-ranks the list of candidate terms. Nevertheless, for the top-40 of the terms in their ranked lists, they report an improved precision of 75.70% for the NC-value, over a precision of 70.84% for the same interval, when not taking into account context words.

Another recent interesting statistics-based approach focuses on the comparison of the distributional properties of terms across corpora of different domains. Basili et al. [2001] propose a measure called contrastive weight so that identification of relevant term candidates is carried out through analyzing their in-domain and out-of-domain distributions. Evaluation on the Italian Civil Code corpus and a collection of 6000 news showed that contrastive method outperforms pure frequency methods. The syntactic structure of the term, i.e. its head is also used to compute the contrastive weights of multiword terms, the method referred to as Contrastive Selection via Heads. This choice is justified by the typically low frequencies of multiword terms. Bonin et al. [2010] address this data sparseness by following a two-phased approach: (1) extract a shortlist of well-formed and relevant candidate multi-word terms using the C-NC value methodology and (2) evaluate them by applying a contrastive ranking method either by filtering noisy general terms or by discerning between semantically different types of terms within heterogeneous terminologies. Their evaluation was carried out in the History of Arts and the Legal domain, improving significantly multiword term extraction results. Especially for legal domains this method is of great help towards building legal ontologies.

Wong et al. [2007] further develop the contrastive weight formula with finer measures: domain prevalence indicating in-domain usage of a term and domain tendency that expresses the extent to which a term is used towards the target domain. In order to quantify linguistic evidences in the form of candidates, modifiers and context words four additional measures were derived from these base measures: discriminative weight, modifier factor, average contextual discriminative weight, and adjusted contextual contribution. Discriminative weight in particular is the product of domain prevalence and domain tendency, therefore both of these factors should be of a high value for establishing termhood status. Base and derived measures contribute to the computation of a final selecting/ranking weight known as Termhood (TH). Experiments were carried out in two text sources: a domain corpus containing 24 documents (51K words) in "liver cancer" from BioMedCentral.com, and a contrastive corpora consisting of 11115 news articles

(4.3M words) in various domains such as “technology”, “business”, “politics” and “sports”. Results were compared to NC-Value and Contrastive Weight methodologies. Termhood was found to downgrade candidates with a low domain tendency while moving high frequency candidates higher in the ranking. Termhood has also a smaller standard deviation than NC-Value and Contrastive Weight.

TE systems may include modules responsible for the recognition and grouping of term variants. Variation can be orthographical, morphological, lexical and structural, or it may be due to the use of acronyms and abbreviations. Approaches for variation handling include string matching techniques, or stemming and grouping of terms sharing similar stems. Nenadié et al. [2004] propose a method in which they combine conflation of different surface realizations for a candidate term and termhood estimation (via C-value) for whole synterms (groups of candidates that share canonical representations) rather than individual candidates. Evaluation was carried out using the GENIA corpus[1], which includes 76K manually annotated terms (30K after conflation of term variants) in 2000 abstracts of the biomedical domain. Authors report significant improvements in recall and precision , the latter being explained by consideration of joint frequencies of occurrence for all terms of candidate synterns.

In Wermter and Hahn [2005], termhood is decided via limited paradigmatic modifiability, a measure that, given a multi-word terminological unit, reflects the probability of not letting other words appear in the specific token slots of the unit. Authors give examples of terms that score lower than others by frequency of occurrence, and yet appear higher in a list ranked by the P-Mod method. Evaluation was carried out on a corpus of medical abstracts, while termhood was tested against a knowledge source of controlled vocabularies from the biomedical domain.[2] For the 30% percent of their ranked list of term candidates, authors report 0.37, 0.24, and 0.18 precision scores for bigrams, trigrams and quadgrams, respectively. In order to get a 0.5 recall for bigram terms, the P-Mod method needs to winnow 29% of its ranked list of candidates, compared to 35% and 37% for lists extracted by the authors by applying t-test and C-value respectively.

3.2 Greek TE

The Greek TE module is used to annotate text for term information. It is a hybrid system comprising a term pattern grammar based on finite-state technology, and a statistical filter, used for the removal of grammar-extracted terms lacking statistical evidence [Georgantopoulos and Piperidis, 2000a].

The pattern grammar used is a subset of pattern rules converted to a non-deterministic finite state automaton (NDFA). The grammar consists of a set of rules recognising single and multi-word candidate terms, based on POS tags assigned to each word by language-specific taggers. Statistical filtering is performed via a variation of the tf-idf measure. The architecture of the Greek TE system is shown in Figure 3.1.

The Term Extraction module has been tested on two different corpora, one per language (EN, EL). For the Greek language, a manually annotated corpus of approximately 65k words was used. The English corpus comprises of several manually annotated texts, which are grouped into three datasets (named below as core1, core2 and travel datasets) and are approximately 63k words in size. Gold annotations on these data were prepared using the ILSP’s *Marker* software (Figure 3.2).

[1] See GENIA project home page: http://www-tsujii.is.s.u-tokyo.ac.jp/ genia

[2] See UMLS Metathesaurus Fact Sheet: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Figure 3.1: **Architecture of Greek Term Extraction Module.**

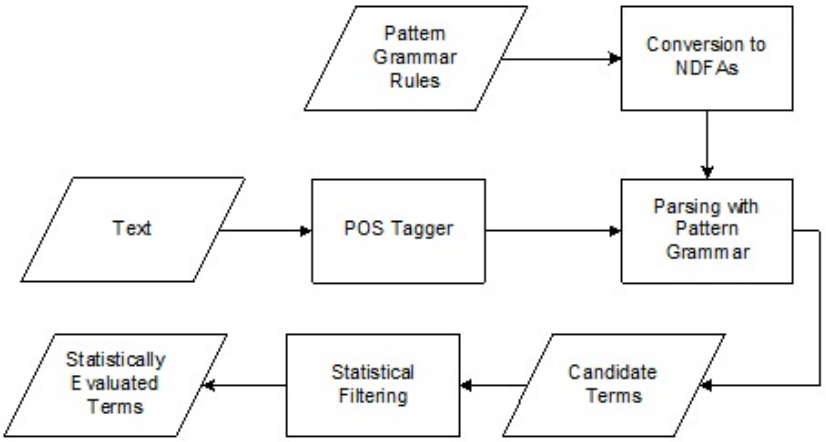
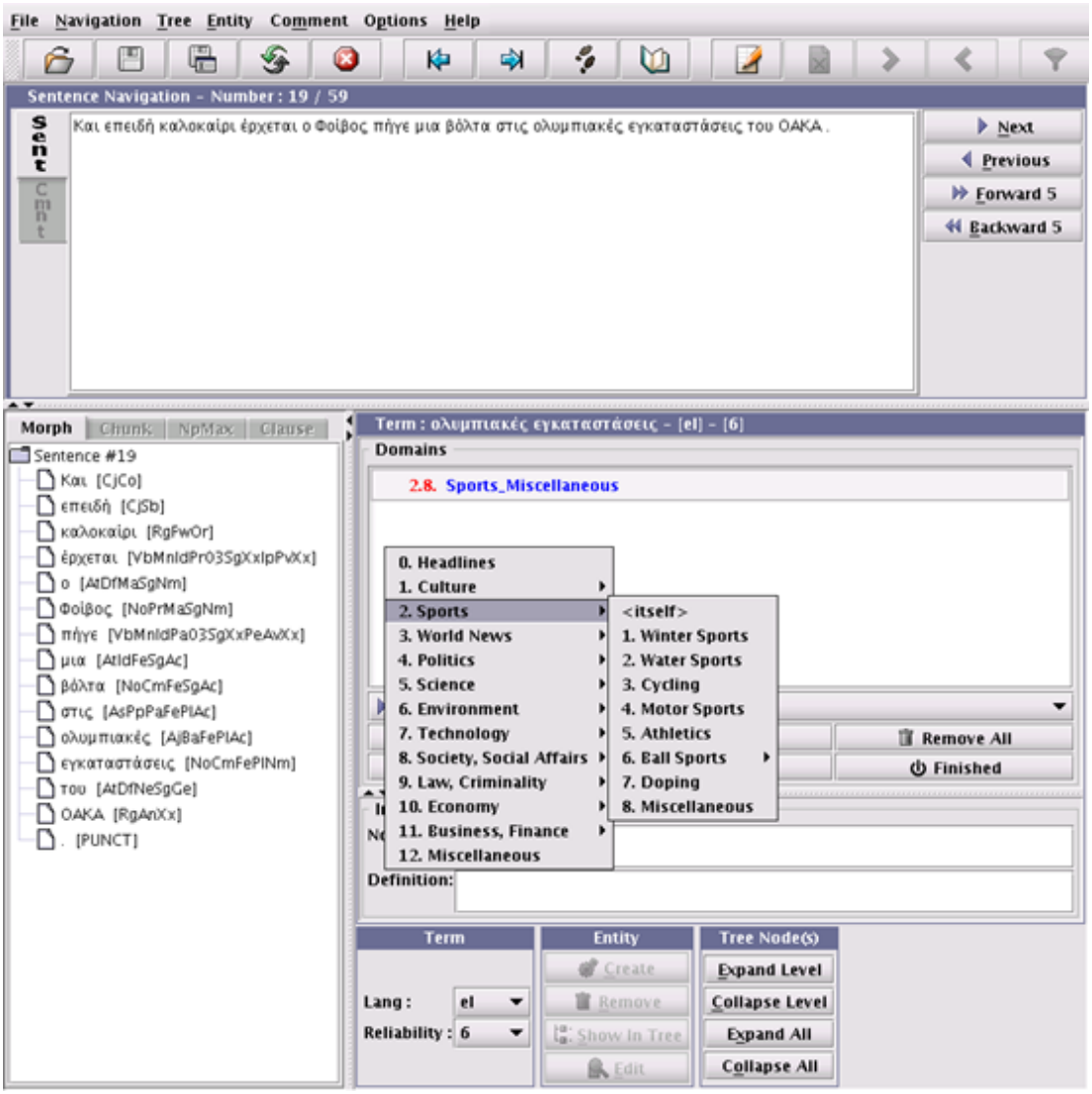


Figure 3.2: **Term and term topic annotation.**



Evaluation of the Term Extraction module was measured with the standard recall and precision figures. Precision reflects the correctness of the terms automatically identified in the test set. It is the ratio of relevant ("true positives") terms in the set of all terms ("true positives" and "false positives") automatically identified by the system. On the other hand, recall is measured as the ratio of "true positives" in the set of all terms manually annotated in the test set. F-measure, a single score used as a combination of precision and recall, is defined as their harmonic mean: *2*P*R/(P+R)*. Table 3.1 summarizes the evaluation results.

Table 3.1: **Term Extraction module evaluation results.**

Corpus	**Precision**	**Recall**	**F-measure**
Core 1-2	18.73%	62.09%	28.77%
Travel	20.01%	57.64%	29.71%
Greek-corpus	19.97%	64.09%	30.46%
Overall	**19.57%**	**61.27%**	**29.64%**

The system's recall is considered satisfactory. On the other hand, the precision figure is considerably, but not unexpectedly, low since term grammars inherently produce numerous terms because of the generality of the rules. Study of the terms that were not recognised by the module revealed three factors responsible for this:

1. incorrect part-of-speech identification (e.g. adverb instead of adjective or verb instead of noun),
2. many human-produced terms are included in a bigger machine-generated term since the grammar pattern can sometimes capture very lengthy candidate terms (e.g. human spotted "beach" and the term extraction module spotted the greater "most famous beach", human spotted "island" and the module spotted "rocky island soil"),
3. inadequate frequency statistics. Many human-crafted terms have either low frequency in the news texts because of small or zero frequency in the reference corpus, resulting in a very small tf-idf performance.

3.3 Romanian TE

In Romanian TE is made a clear distinction between single-word and multi-word terms, since their identification and extraction is usually performed by using different techniques. The current objective is to identify relevant terminological terms into comparable corpora and then, having all these terms extracted, to find translation equivalents among them. The following paragraphs describe the terminology extraction techniques.

3.3.1 Single-word terminology extraction

We decided to approach the task of single-word terminology extraction by using a modified Damerau method [Damerau, 1993] as it has been reported to yield very good results [Schutze, 1998; Paukkeri et al., 2008]. Damerau's approach compares the relative frequency in the documents of interest (user's corpus - CU) to the relative frequency in a reference collection (reference corpus - CR). The original formula for computing the score of a word is:

$$score(word) = \frac{\frac{f(word,C_U)}{\|C_U\|}}{\frac{f(word,C_R)}{\|C_R\|}} \tag{3.1}$$

where $f(word_i, C_j)$ is the frequency of the word i in corpus j, and $\|C_j\|$ is the total number of words. One can immediately notice that the score for a word is calculated according to the likelihood ratios of occurring in both corpora (that of the user and the reference). The main idea is to compare the maximum likelihood estimates (MLE) computed on the user corpus to the ones on the reference corpus. Consequently the reference corpus should be a large, balanced and representative corpus for the language of interest. Essentially, the MLE on such a corpus is equivalent with a unigram language model:

$$P_{MLE}(word) = \frac{f(word, C_R)}{\|C_R\|} \tag{3.2}$$

and, in practice, such models are usually used in information retrieval to determine the topic of documents. Thus, the Damerau formula works on comparing two unigram language models.

It has been proven however, that due to data sparseness, using that unigrams language models constructed only by the means of MLE behaves poorly and that a proper smoothing should be performed [Chen and Goodman, 1996]. In order to do this, we employ a variant of Good-Turing estimator smoothing of Kochanski [1995]:

$$P_{GT}(word) = \frac{f(word, C_R) + 1}{\|C_R\| + \|V_R\|} * \frac{E(f(word, C_R) + 1)}{E(f(word, C_R)} \tag{3.3}$$

where V_R is the vocabulary (the unique words in C_R) and $E(n)$ is an estimate of how many different words were observed exactly n times.

Let us consider a slightly modified example from Kochanski [1995]: let's say we have a (reference) corpus with 40,000 English words which contains only one instance of the word "unusualness": $f(word, C_R) = 1$. Let's also say that the corpus contains 10,000 different words that appear once and so, $E(1) = 10,000$, and that we have 5,500 words that appear twice, giving $E(2) = 5,500$. Let us also consider that the total number of the unique words in the corpus is 15,000 ($|V_R| = 15,000$). The Good-Turing estimate of the probability of "unusualness" is:

$$P_{GT}(unusualness) = \frac{1+1}{40,000+15,000} * \frac{5,500}{10,000} = \frac{1}{50,000} \tag{3.4}$$

Using MLE, we would have had a larger value:

$$P_{MLE}(unusualness) = \frac{1}{40,000} \tag{3.5}$$

Because the sum of the probabilities must be 1, we have a remaining probability mass to be assigned to the unseen words (U). Consequently, the probability of an unseen word depends on the estimated number of unseen words [Kochanski, 1995]:

$$P_{GT}(unseen) = \frac{E(1)}{(\|C_R\| + \|V_R\|) * \|U\|} = \frac{10,000}{55,000 * \|U\|} \tag{3.6}$$

Going back to the Damerau formula, we have now that:

$$score(word) = \frac{\frac{f(word, C_U)}{(\|C_U\|)}}{P_{GT}(wordinC_R)} \tag{3.7}$$

We may consider that the first n words having the highest scores are terminological terms.

In case CU is a large corpus, we can also compute Good Turing estimators for the numerator. For small corpora, this is however impractical since one cannot compute the estimates $E(n)$ with high enough confidence.

This approach can be improved by additional preprocessing of the corpora involved. First, for better capturing the real word distribution, it is better to use word lemmas (or stems) instead the occurrence forms. Second, the vast majority of the single terminological terms are nouns and so, one should apply a pos filtering in order to disregard the other grammatical categories. Both can be resolved by employing stand-alone applications that can POS-tag and lemmatize the considered texts. As our research and development is mainly focused on English and Romanian, we usually make use of the TTL preprocessing Web Service [Ion, 2007; Tufiş et al., 2008] when dealing with these languages. TTL is publicly available[3] and it can be used for: sentence splitting, tokenization, POS-tagging, lemmatization and chunking; its tag-set is Multext-East[4] compliant and its output conforms to the Corpus Encoding Standard for XML - XCES.[5]

As reference corpora, one can use the Agenda corpus [Tufiş and Irimia, 2006] and Wikipedia[6] for Romanian and Wikipedia[7] for English.

The method presented above can be reinforced with the well-known TF-IDF (term frequency - inverse document frequency) approach [Jones, 1972], provided that the corpus of interest is partitioned into many documents, as in the case of JRC Acquis [Steinberger et al., 2006], or that this partitioning can be automatically performed. This approach does not need the additional reference corpus and works only on the corpus of interest. TF-IDF is a statistical measure used for evaluating the importance of a word in a document, given a collection of documents. The importance is directly proportionate with the frequency of the word in the document and inverse proportionate with the number of documents that contain it. Therefore, it is a good measure that can be used for extracting terminology. TF stands for Term Frequency and it is computed as the normalized frequency of the word in the document, exactly like the MLE:

$$TF(word) = \frac{f(word, document)}{\|document\|} \tag{3.8}$$

IDF stands for Inverse Document Frequency and it is a measure of the general use of a word:

$$IDF(word) = log \frac{number of documents in the corpus}{number of documents in the corpus containing the word} \tag{3.9}$$

It is obvious that for a word W and a given document D, the TF-IDF value is high if W has a high frequency in D and it appears in very few other documents. This approach can also be improved by using lemmas instead of word frequencies and POS-filtering.

3.3.2 Multiple-word terminology extraction

Terminology extraction does not limit to the single-word terms and so, one must be able to extract multi-word terminology, too. Smadja [1993] advocated first that low variance in relative position is a strong indicator for multi-word terminological expressions, which can be found among the collocations of a corpus. These are expressions which cannot be translated word-by-word using only a simple dictionary and a language model, because they are characterized by limited compositionality - the meaning of the expression is more than the sum of the meaning of the words composing the collocation. Many definitions have been given for the term collocation. These are some of those frequently used in NLP [Stefanescu, 2010]:

[3] TTL is implemented both as SOAP (http://ws.racai.ro/ttlws.wsdl) and REST Web Services.
[4] http://nl.ijs.si/ME/
[5] http://www.xces.org/
[6] http://ro.wikipedia.org/wiki/
[7] http://en.wikipedia.org/wiki/

- "Collocations of a given word are statements of habitual or customary places of that word" [Firth and Palmer, 1968]
- "A sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" [Choueka, 1988]
- "A recurrent combination of words that co-occur more often than expected by chance and that correspond to arbitrary word usages" [Smadja, 1993]
- "A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things" [Manning et al., 1999]
- "Two or more words which occur significantly often together within a predefined window in a given large corpus" (Wolff and Quasthoff, 2002)

Collocations can be found as noun phrases (*red wine, weapons of mass destruction*), phrasal verbs (*pull a face, strike a bargain, make up, ro: a aduce atingere, a intra n vigoare*) and others (*young and restless, rich and powerful*).

The linguists usually employ the following criteria when dealing with collocations [Manning et al., 1999]:

- Non-compositionality - the meaning of the whole is more than the sum of the meanings of the parts
- Non-substitutability - the words composing the collocation cannot be substituted with synonyms
- Non-modifiability - many collocations cannot be modified with additional lexical material or through grammatical transformations

A collocation for which all three conditions described above hold is practically an idiom. Furthermore, the nature of a collocation can be morpho-syntactic or semantic; it can be general or domain specific. Different methods have been proposed for finding collocations. Justeson and Katz [1995] counted the occurrences of bigrams and then used a part-of-speech filter in order to rule out those bigrams which cannot be phrases. Smadja [1993] employed a method based on the mean and the variance of the distances between pairs of words, while others [Church et al., 1989] used t Test, chi square Test, $Log-Likelihood$ or $MutualInformation$ for finding pairs of words which appear together in the text more often than expected by chance.

Our solution for the identification and extraction of collocations is based on two of the methods enumerated above. A pair of words is considered a collocation if:

- the distance between them is relatively constant [Smadja, 1993];
- they appear together more often than expected by chance: Log-Likelihood [Church et al., 1989].

For collocation identification Smadja uses mean and variance computed for the distances between pairs of words in the corpus in order to identify those words which appear together in a somewhat fixed relation. The mean is simply the average of the distances, while the variance measures the deviations of the distances with respect to the mean already computed. The variance is calculated using the following formula:

$$\sigma^2 = \frac{\sum_{1}^{n}(d_i - \mu)^2}{n-1} \tag{3.10}$$

where n is the total number of distances, d_i are the distances and μ is the mean.

σ^2 is the variance, while σ is the standard deviation. If two words appear together always at the same distance, the variance is equal to 0. If the distribution of the distances is random (the case of those words which appear together by chance), then the variance has high values. Smadja [1993] shows that collocations can be found by looking for pairs of words for which the standard deviations of distances[8] are small.

In order to find terminological expressions, we employ a POS-filtering, computing the standard deviation for all *noun-noun* and *noun-adjective* pairs within a window of 11 non-functional words length, and we keep all the pairs for which standard deviation is smaller than 1.5 - a reasonable value according to Manning et al. [1999]. This method allows us to find good candidates for collocations but not good enough. We want to further filter out some of the pairs so that we keep only those composed by words which appear together more often than expected by chance. This can be done using Log-Likelihood (LL). The idea behind the LL is finding the hypothesis which describes better the data obtained by analyzing a text. The two hypotheses we consider are:

- the null hypothesis - independence:

$$H_0 = P(w_2 \| w_1) = p = P(w_2 \| \neg w_1) \tag{3.11}$$

- non-independence hypothesis - the words co-occur more than they would do by chance:

$$H_1 = P(w_2 \| w_1) = p_1 \neq P(w_2 \| \neg w_1) \tag{3.12}$$

The LL formula is:

$$LL = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_j log \frac{n_{ij*n_{**}}}{n_{i*} * n_{*j}} \tag{3.13}$$

If the score obtained is higher than a certain threshold, the null hypothesis (H_0) is rejected with the price of an error depending on that threshold. We compute the LL values for all the pairs obtained using Smadja's method. In calculating the LL value for a certain pair, we used only the parts of speech associated with the words forming that pair. We kept in a final list, the pairs for which the LL values were higher than 9, as for this threshold the probability of error is less than 0.004 according to the *chisquare* tables.

We keep as terminological expressions only those for which at least one of the terms forming them can be found between the single-word terminological terms, disregarding their context.

3.3.3 Evaluation

For the evaluation of RACAI's terminology extraction tool we used the JRC-Acquis and the Eurovoc thesaurus in the absence of a manually annotated gold standard. For English and Romanian we considered the newest 500 documents from 2006 and then applied the tool for extracting terms. In order to compute the accuracy figures for Eurovoc, we also generated the list of all Eurovoc terms that appeared only in these 500 documents for each language and counted how many of the recognized terms were found in the corresponding restricted list of Eurovoc terms. Regarding this evaluation methodology, we must observe the following:

- the list of Eurovoc terms is not exhaustive nor definitive and as such, there may be terms that the application discovers that are not in Eurovoc. Examples for English include

[8]The distance is negative when the pair is formed by the center-word and a word in front of it.

"Basel convention", "standards on aviation", "Strasbourg", "national safety standards", etc.

- we evaluate the recall of the application because we the Eurovoc terms exist only in a thesaurus fixed form, e.g. for the Eurovoc term "reduced price", the text might also contain the realization "reduced prices".

We postulated the fact that terminology extraction will do better in recognizing more frequent Eurovoc terms and thus, we wanted to evaluate the recall of the application on term occurrence frequency bands. Table 3.2 shows the recall of the application on four term occurrence frequency bands: all terms (frequency at least 1), all terms more frequent than 10, all terms more frequent than 100 and all terms more frequent than 500. It clearly shows that that the algorithm is, in fact, not able to increase its recall due to the fact that more frequent terms are not likely to be discovered by our term extraction procedure which is based on comparing the term's distribution in the current corpus with the distribution of the term in a balanced corpus.

Table 3.2: **Evaluation recall scores for the Romanian TE.**

	English	Romanian
Number of documents	500 /7972 (06.27%)	500 / 5792 (08.63%)
Size of preprocessed collection	33.8 MB	4.0 MB
Eurovoc terms identified out of those found in the collection having at least 1 occurence	816 / 2140 (38.13%)	117 / 483 (24.22%)
10 occurences	490 / 1523 (32.17%)	75 / 324 (23.15%)
100 occurences	269 / 1039 (25.90%)	44 / 185 (23.79%)
500 occurences	79 / 525 (15.04%)	8 / 51 (15.68%)

3.3.4 Further Work

We have presented different techniques currently used by our tools for extracting both single and multi-word terminology from texts and we are now applying them for English and Romanian on the comparable collection of documents of the ACCURAT project.

After extracting the terminological terms from each pair of comparable (or parallel) documents we employ a simple method for finding translation equivalence among these terms, mainly based on a Giza++ like translation model and the lexical similarities between terms. The best matches are considered translation equivalents if their associated score, according to the translation model, is higher than a certain threshold that can be set by the user. We make a clear distinction between *identifying term translation equivalents* and *aligning terms*.

The former is concerned with cross-lingual meaning equivalence irrespective of actual occurrence positions and can be used to build terminological translation equivalence tables that can be added as resources to alignment applications. Such applications can then use them as strong indicators / clues for fragments / paragraphs that might be closely related and further subject to usual / regular word / phrase alignment processing. The latter identifies the corresponding positions of a term and its actual translation in text. It essentially implies the existence of a general alignment application that can align words and expressions in a comparable (parallel) bi-text. Consequently, such an application would also align the terminological term, if such terms would have previously been identified.

To better emphasize the distinction, let us consider the terminological multi-word terms t_{ro}: *Uniunea Europeana* (ro) and t_{en}: *European Union* (en), from the JRC Acquis. Identifying terminological translation equivalents refers only to finding that they are translation equivalents, while aligning terminological terms refers to finding their corresponding occurrences in a bi-text. For example, if t_{ro} has 2 occurrences and t_{en} has 3 occurrences, the aligning process should find something like the set $(t_{ro}-1 \longleftrightarrow t_{en}-1; \emptyset \longleftrightarrow t_{en}-2; t_{ro}-2 \longleftrightarrow t_{en}-3)$.

As we make the above distinction, it is clear that in the context of Terminology Extraction, the terminology alignment expression refers to identifying terminological translation equivalents, since no alignment application is involved in the process at this point. The results of the terminology alignment would consequently improve the alignment tools designed for the explicit alignment of bi-texts, in various languages.

3.4 Latvian, Lithuanian TE

Both languages, Latvian and Lithuanian, belong to the Baltic group of the Indo-European family of languages. From a linguistic typology point of view, Latvian and Lithuanian languages are synthetic languages with rich inflection and a flexible word order (SVO - a direct word order, however, remains the basic one). For example, Lithuanian nouns, adjectives, participles, and numerals typically have 7 cases in singular and 7 ones in plural that makes 14 different wordforms of a single word [Grigonyte et al., 2011]. In Latvian, for example, nouns have 29 graphically different endings, adjective - 24 and verbs - 28, and only half of the endings are unambiguous [Skadina et al., 2011]. In addition, some Lithuanian nouns, adjectives, participles, and numerals have a grammatical category of gender (masculine, feminine and neutral) that contributes to a variety of wordforms [Zabarskaite and Vaisniene, 2011]. The majority of Latvian and Lithuanian wordforms are constructed with affixes (inflectional suffixes and endings), and the latter are the principal means of making syntagmatic relations between words in a sentence and/or relations among wordforms in a paradigm in both languages [Zabarskaite and Vaisniene, 2011; Skadina et al., 2011]. All these characteristics of Latvian and Lithuanian makes it difficult to apply pure knowledge-poor or statistical methods to language processing, TE in our case. Moreover, corpora resources of a significant size are extremely important for statistical modeling that is complicated for narrow and/or emerging domains which lack terminological data.

In the 1990s, a number of TE tools were developed, mainly for English and French languages, however, for Central and East European languages TE tools appeared later since at that time there was no satisfactory method of morphosyntactic analysis for most of these languages, and even nowadays there is a significant gap between analytical languages, on the one side, and synthetic ones, on the other side, due to their under-resourced status with the lack of language resources and tools [Kruglevskis, 2010]. First experiments on TE for Lithuanian were described by Zeller [2005] in his PhD thesis. A recent paper by Grigonyte et al. [2011] explores the problem of extracting domain-specific terminology in the field of science and education from Lithuanian texts from the perspective of existing term extraction tools. Four different TE approaches have been applied and evaluated - three statistical and one symbolic. One of the three statistical approaches (keyword cluster identification, keyword extraction with machine learning and collocation extraction) - collocation extraction and the linguistic approach (that extracts term candidates on basis of morphosyntactic patterns and is language-dependent) appeared to be quite reliable with the second one to be more promising, according to the measure of recall. However, in terms of precision, the symbolic, or linguistic, approach produced much noise.

For the Latvian language, the first experiment in TE showed that a linguistic method based

on morphosyntactic analysis is more appropriate than a statistical one which is more adequate for analytical languages [Kruglevskis and Vancane, 2005]. The applied method was based on morphosyntactic analysis of the sentence and aimed at identifying term candidates in a text on basis of morphosyntactic patterns (e.g. noun phrase patterns). One of the advantages of the linguistic method is that term candidates with low frequency are not discarded (Ibid.). A semi-automatic TE has been applied to Latvian texts recently [Kruglevskis, 2010].

Obviously, one of the solutions that has not been researched so far, for example, for Lithuanian, could be the implementation of a hybrid approach with symbolic (knowledge-rich or linguistically-based) methods, e.g. morphological analysis, Part-of-Speech (POS) tagging, syntactic parsing, including a set of morphosyntactic patterns of multiword terminological constructions relevant to this or that language (see Table 3.3 for Latvian and Lithuanian[9]). As we can see, terminological constructions, or morphosyntactic patterns of terms, are mostly nominal groups with nouns, adjectives and participles [Kruglevskis and Vancane, 2005; Kruglevskis, 2010; Skujina, 2010; Grigonyte et al., 2011]. However, verbs can be also used as terms, for example, according to Grigonyte et al. [2011], a verb has the tenth position in the top-ten frequency list. Also, the dominant case is genitive for both languages. In Latvian, for example, a half of all nominal groups and genitive ones as it was observed on the material of technical multiword terms [Skujina, 2010]. Single-word terms are also very frequent, however, the majority of terms in any domain are multiword terms [Skujina, 2010]. We can conclude that the structure of multiword terms in both languages is similar.

Moreover, it was observed, that not all grammatical categories are important for TE, and demonstrated that grammatical categories of gender and number of Lithuanian nouns, for example, cannot help a lot in TE while the category of case can serve as a marker of a terminological construction, e.g. genitive case in a two-element nominal group $N_{genitive}$ Grigonyte et al. [2011], and this refers to Latvian as well. The situation becomes even more complicated in case of a three-(and more)element nominal group, and syntactic parsing might be required to identify relations between elements of a term or existing terminological data can be used for verification and/or splitting an n-element nominal group into at least two term candidates.

A list of top-five and three more Latvian and Lithuanian morphosyntactic patterns of terminological constructions is represented in Table 3.3.[10] This data is used during TE for these two languages in ACCURAT.

For terminology extraction *CollEx* tool, developed by FFZG for automatic collocation extraction, was adapted for Baltic languages - Latvian and Lithuanian. The language independent *CollEx* version developed in ACCURAT project requests only syntactic pattern list and stopword list for particular language as it is described in Section 3.5. Since originally *CollEx* was designed for collation extraction, it extracts multiword term candidates consisting of 2, 3 or 4 words only.

The evaluation of *CollEx* tool was performed on manually annotated IT domain texts for Latvian and Lithuanian. The Latvian test corpus consists of 15 documents (part of IT subdomain corpus collected by ILSP for Task 5.3). The total size of test corpus is 19456 words. It contains 3434 terms in total, 1288 terms are multiword units consisting of 2, 3 or 4 words. The Lithuanian test corpus contains 6 documents, the total size of test corpus is 5726 words. It contains 1131 terms in total, 513 terms are multiword units.

[9] A list of Latvian patterns was compiled during the research in the TTC project that received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no: 248505; a list of Lithuanian patterns was taken from Grigonyte et al. [2011]

[10] N stands for noun, A - for adjective, P - for participle. Note, that all patterns in the table are given in their canonical form.

Table 3.3: **Latvian, Lithuanian morphosyntactic terminological patterns.**

Latvian	Lithuanian
$N_{genetiv}$ $N_{nominative}$	$N_{genetiv}$ $N_{nominative}$
$A_{nominative}$ $N_{nominative}$	$A_{nominative}$ $N_{nominative}$
$G_{nominative}$ $N_{nominative}$	$A_{genetiv}$ $N_{genetiv}$ $N_{nominative}$
$N_{genetiv}$ $N_{genetiv}$ $N_{nominative}$	$N_{genetiv}$ $N_{genetiv}$ $N_{nominative}$
$A_{genetiv}$ $N_{genetiv}$ $N_{nominative}$	$A_{nominative}$ $N_{genetiv}$ $N_{nominative}$
$A_{nominative}$ $N_{genetiv}$ $N_{nominative}$	$P_{nominative}$ $N_{nominative}$
$N_{genetiv}$ $A_{nominative}$ $N_{nominative}$	$A_{nominative}$ $A_{nominative}$ $N_{nominative}$
$G_{genetiv}$ $N_{genetiv}$ $N_{nominative}$	ABBR $N_{nominative}$
ABBR $N_{nominative}$	$N_{nominative}$ ABBR
	ABBR $N_{genetiv}$ $N_{nominative}$

The *CollEx* tool used for evaluation performs scoring of the n-grams by five different association measures: Dice coefficient, modified pointwise mutual information, chi-square statistic, log-likelihood ratio and t-score statistic. We applied these measures to (1) corpus as a single document and (2) to each individual document in the corpus. Precision, recall and F-measure were calculated for multiword units (terms) only. Results obtained for both languages are rather similar. However, in terms of F-measure better results are achieved for Latvian with chi-square statistics, while for Lithuanian with modified pointwise mutual information.

In Table 3.4 we provide evaluation results for the Latvian language. In terms of F-measure better results achieved with chi-square statistic measure. However, results are rather similar between measures, except for Dice coefficient.

Table 3.4: **Evaluation results on Latvian language.**

	Corpus			Separate Documents		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Dice coefficient	21.42	46.25	29.28	30.71	**50.53**	38.20
Modified pointwise mutual information	**44.96**	46.65	45.79	40.07	50.18	44.56
chi-square statistic	47.52	**47.18**	47.35	41.70	49.45	**45.25**
log-likelihood ratio	46.52	45.94	46.23	**41.35**	48.75	44.75
t-score statistic	47.09	46.37	46.37	**41.77**	49.62	45.36

Table 3.5 provides evaluation results for the Lithuanian language. Although in general results are similar to Latvian, the best results in terms of F-measure are achieved with modified pointwise mutual information measure. Also usually recall for particular method is higher in Latvian texts, while precision is higher in Lithuanian texts.

Table 3.5: **Evaluation results on Lithuanian language.**

	Corpus			Separate Documents		
	Recall	Precision	F-measure	Recall	Precision	F-measure
Dice coefficient	31.29	**55.06**	39.90	32.73	55.32	41.13
Modified pointwise mutual information	42.27	53.41	**47.19**	37.41	54.59	**44.40**
chi-square statistic	**43.17**	51.17	46.83	**37.90**	52.38	43.77
log-likelihood ratio	**43.17**	51.06	46.78	37.59	52.25	43.72
t-score statistic	**43.17**	51.17	46.83	37.59	52.38	43.77

Obtained evaluation results of *CollEx* tool demonstrate potential of this tool for term extraction task. However, these results could be insufficient for multilingual term mapping task. Thus we plan to perform more detailed evaluation of these methods to improve terminology extraction for Baltic languages.

3.5 Croatian TE

For usage within the ACCURAT project FFZG developed a tool *CollEx* that perform automatic collocation extraction.[11] Although in the ACCURAT project DoW (page 35) another tool - *TermeX* - was planned for terminology extraction, we were forced to depart from this description because the original *TermeX* creators' conditions could not be met by the ACCURAT project. The main issue was the request to make *TermeX* an open source application so that it can be included into ACCURAT Toolkit freely. Since agreement that would satisfy both sides could not be achieved, we decided to develop the replacement application - *CollEx* - that would cover the command line functionality of *TermeX*, but also add some more features that were necessary for smoother pipelining into ACCURAT Toolkit.

3.5.1 CollEx

Croatian language features all the inflectional properties that most of Slavic languages adhere to. For nouns, adjectives, pronouns and numerals morphosyntactic categories of case, gender and number have 7, 3 and 2 possible values respectively. Adjectives mark definiteness with endings as well and add the 3 grades of comparison to it. In the system of simple verbal word-forms there are categories of person, number, tense, mood which can have 3, 2, 3, 2 possible values respectively, while in the compound tenses and moods a category of gender (3 values) comes into play. This illustrates the complexity of the Croatian inflectional system that can lead to, e.g. 227 regular different word-forms for every adjective.

On the syntactical level Croatian has a flexible word order with SVO being marked as stylistically neutral and considered a basic one. While the order of words within the constituents (e.g.

[11] See the technical description of the CollEx tool in D2.6

PPs, NPs etc.) is generally fixed, the very constituents can be freely shuffled around the sentence. Besides, the rules of clitics positioning, usually after or within the first phonetic word, can break the constituents apart leading to syntactic phenomena such as branch-crossing and long-distance dependencies.

All these characteristics of Croatian makes the use of knowledge-poor or statistical methods to language processing to underperform compared to e.g. English, mainly because of the data sparseness with multiple word-forms. For inflectionally rich languages this problem is usually tackled by lemmatization and we also apply this procedure here. Thus the approach that uses morphosyntactic analysis combined with statistics of collocation association measures over lemmas instead of word-forms only is more favorable than the pure statistical approach. Besides, very large corpora are important for statistical modeling of specific domains and their terminologies and for under-resourced languages they are not always around.

3.5.2 Application of terminology extraction in Croatian

For usage within the ACCURAT project FFZG developed a tool CollEx that perform automatic collocation extraction. Since this tool is language independent, the only language specific data it needs are list of terminological and/or collocational MSD-patterns (it could be written also as RegEx), and a list of stop-words. It extracts the n-grams of length 2, 3 and 4, applies the MSD-patterns as filters and calculates the requested association measure selected from a set of five measures.

It can be noted that only the morphosyntactic category of case is used in defining the Croatian terminological MSD-patterns since the categories of gender and number are irrelevant for this detecting procedure. The scoring of the n-grams that pass the POS/MSD filters and stop-word filters is performed by five different association measures. Association measures, loosely speaking, measure how much words in a sequence of words co-occur more than by chance.

The five association measures implemented in this tool are the following:

1. **Dice coefficient:**

$$DICE(w_1...w_n) = \frac{nf(w_1...w_n)}{\sum_{i=1}^{n} f(w_i)} \tag{3.14}$$

   where $f(.)$ is the frequency of a specific n-gram.

2. **modified pointwise mutual information:**

$$I^{'}(w_1...w_n) = log_2 \frac{f(w_1...w_n)P(w_1...w_n))}{\prod_{i=1}^{n} P(w_i)} \tag{3.15}$$

   where $f(.)$ is the frequency of a specific n-gram and $P(.)$ is the probability of a n-gram calculated as a maximum likelihood estimate.

3. **is the frequency of a specific n-gram:**

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.16}$$

   where O_{ij} and E_{ij} are observed and expected frequencies in a contingency table of two dimensions for bigrams (contingency tables for n-grams have n dimensions).

4. **log-likelihood ratio:**

$$G^2 = 2 * \sum_{i,j} O_{ij} log \frac{O_{ij}}{E_{ij}} \tag{3.17}$$

where observed and expected frequencies are calculated as in the chi-square statistic

5. **t-score statistic:**

$$tscore = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \tag{3.18}$$

where observed and expected frequencies are calculated as in the chi-square statistic and the log-likelihood ratio.

These association measures have been selected from an exhaustive list of existing association measures since previous research for bigrams [Evert, 2005; Pavel, 2009] and n-grams [Petrovic et al., 2010] has shown that these measures show the most consistent results on different data sets and languages. Additionally, only these association measures are implemented since other measures do not show consistent and statistically significant improvements over each other.

3.5.3 Evaluation

The evaluation of the CollEx tool for term extraction was performed on the test corpus of 18,000 tokens collected from the web as comparable corpora. In this corpus terms were manually annotated. Two association measures were used for evaluation: the Dice coefficient and the modified pointwise mutual information. The frequency minimum for a phrase to be included in potential terms was three. Since we use a collocation extraction tool for extracting terminology only the terms of length 2-4 words were taken into account. From a list of extracted bigrams, trigrams and tetragrams, precision, recall and the F-measure were recorded in case the extracted term was an exact overlap of the annotated term. The results are given in Table 3.6.

Table 3.6: **Evaluation results of CollEx applied to Croatian test corpus.**

	precision	recall	F1
Dice	39.37	23.63	29.53
modified pointwise mutual information	39.15	24.19	29.90

The results show a F1 measure slightly below 0.3 which is a result very similar to the result obtained on Greek language with a different tool. Regarding different association measures, both produce very similar results. We expect to get a greater difference between results by implementing additional association measures, namely the chi-square statistics, the log-likelihood ratio and the t-test. In this evaluation the F1 measure was maximized. By changing the beta parameter, higher numbers for precision, or recall, can be obtained.

4 Mapping

Mapping of named entities (NE) or terms is the process of mapping multilingual variants of the same named entity (NE) or term with each other. Recall that named entities can be person, location or organization names but also dates, day names, currencies, etc. Terms are word or multiword units used only in specific domains, for instance, only in medicine or in the automotive domain. Mapping named entities such as dates, day names or currencies can be achieved via manual rules. However, the same approach is not valid for person, location and organization names as well as domain specific terms which is why most related work has put effort into investigating mapping methods for these linguistic units. Once multilingual named entity or term variants are mapped with each other they can be used for machine translation (MT) and cross-language information retrieval (CLIR) Feng et al. [2004].

4.1 Related Work

Mapping NEs multi lingually has been investigated in various earlier studies. One of the early studies is that of Al-Onaizan and Knight [2002b]. The authors use a probabilistic approach to map Arabic NEs to English. In the process of mapping a combination of probabilistic phoneme [Stalls et al., 1998] and spelling based methods [Al-Onaizan and Knight, 2002a] are used. In the phoneme method the probability of transforming an Arabic NE to an English one is computed using phonemes. Similarly, the spelling task computes the probability of mapping spellings or characters of an Arabic NE to an English NE. In the end for an Arabic NE a list of English NEs with probability scores is returned. These scores are updated with search engine hit results where the English NEs are used to query the web. This final step ensures that the most likely English NE is taken as an English variant to the Arabic NE. Moore [2003] uses parallel corpora to learn NE mappings between English-French and English-Spanish languages. Their NEs are in the domain of computer software manuals. Moore computes the log-likelihood of a word in the source language being a translation mapping of another in the target language. Parallel corpora are also used by Huang et al. [2003], Feng et al. [2004], Klementiev and Roth [2006] and Ehrmann and Turchi [2010]. Huang et al. [2003] and Feng et al. [2004] work with English-Chinese NEs, Klementiev and Roth [2006] with English-Russian and Ehrmann and Turchi [2010] with English, French, Spanish, German and Czech texts. In general these approaches make use of the fact that the two NEs come from parallel sentences and learn how to map one NE in e.g. English into the other language e.g. Chinese.

In summary all these approaches require some initial training data such as phoneme mappings or parallel corpora and thus are only limited to language pairs where such data is available. To avoid the problem with training data others have investigate ideas/methods which do not require such data. For instance, Aswani and Gaizauskas [2010] use cognate-based approaches such as the Levenshtein Distance [Levenshtein, 1966], Longest Common Subsequence (LCS), etc. to map NEs in English-Hindi and English-Gujarati languages. Cognates are words that have a similar spelling between several languages because they have similar meaning. The cognate-based methods rely on the languages having similar alphabets and writing styles, but Church [1993] suggests that even very different languages can share a large number of proper nouns, numbers, and punctuation. We adopt some of the cognate-based methods to perform NE mapping between the ACCURAT languages.

4.2 Language Independent NE Mapper

For language independent NE mapping we have implemented two scenarios.

In the first scenario the mapper takes as input two comparable documents in text format and outputs pair of NEs with scores indicating their level of mapping. On both sides we use OpenNLP[1] to identify sentence boundaries.[2] Next, on the English text the mapper applies the OpenNLP NER to extract English NEs. On the foreign text it uses case information to identify candidates as foreign NEs. It treats all capitalized words as NEs and compares them with the English NEs. Consecutive capitalized words are treated as a single NE. For each word in the beginning of each sentence we compare its lowercase variant with a list of lowercase words. If the lowercase variant is found in the list then it is not treated as NE. After having collected NEs in English and so called NEs in the foreign language, we compare each English NE with all the foreign NEs. The comparison is computed using cognate-based methods described later in this section.

In the second scenario the mapper uses proper NE identification on both sides. On the English side it uses the OpenNLP NER as before. On the foreign text side it assumes that the NEs are identified using the NER systems described in Chapter 2. Having both lists of NEs with their types (PERSON, LOCATION, ORGANIZATION) it uses the following cognate-based methods to map them with each other. However, instead of comparing every English NE with every foreign NE it compares every English NE with type X with every foreign NE of the same type. For the comparison we use the following string similarity measures which we view as cognate-based approaches on the assumption that cognates share a higher proportion of character grams.

- **cosine similarity:** The *cosine similarity* method [Salton and Lesk, 1968] is a string similarity metric between different text units. We adopt it for computing the similarity between two NEs. To compute the cosine similarity vector space models [Salton et al., 1975] also called term or word vectors are used. The metric requires two word vectors whose values are numbers representing the strings. Our strings are NEs. From each NE we extract character n-grams and use their counts (tf-term frequency) to create its vector representation. We experiment with bi-gram and tri-gram character grams. We compute the similarity between two NEs X and Y using the following equation:

$$cos(X,Y) = \frac{\sum_i tf(X_i) * tf(Y_i)}{\sqrt{\sum_{i \in X} (tf(X_i))^2} * \sqrt{\sum_{i \in Y} (tf(Y_i))^2}} \tag{4.1}$$

  where $tf(Z_i)$ is the tf value of character gram i in NE Z.

  The cosine similarity gives the cosine angle between two vectors representing NEs. If the angle is *1* it means that the two NEs are identical and *0* indicates that the two NEs have zero character grams in common.

- **Longest Common Subsequence Ratio (LCSR):** The *longest common subsequence* (LCS) measure measures the longest common non-consecutive sequence of characters between two strings. For instance, the words "*dollars*" and "*dolari*" share a sequence of 5 non-consecutive characters in the same ordering. We make use of dynamic programming

[1] http://incubator.apache.org/opennlp/

[2] We apply the OpenNLP English sentence boundary detector on all ACCURAT languages.

[Cormen et al., 2001] to implement LCS, so that its computation is efficient and can be applied to a large number of possible word pairs quickly. The LCS method is used to identify the most likely mapping for every source word by choosing the word in target word that has the longest (normalized) subsequence in common. We normalize by the length of the longest word (see Equation 4.2). If more than one word gives the same score then ties are broken by choosing the word which occurs more frequently in the target data. For more details about LCS see D2.1.

$$LCSR(X,Y) = \frac{\text{length}[LCS(X,Y)]}{\max[\text{length}(X),\text{length}(Y)]} \tag{4.2}$$

where *LCS* is the longest common subsequence between two strings and characters in this subsequence need not be contiguous.

- **Longest Common Substring (LCST):** The *longest common substring* (LCST) measure is similar to the LCS measure, but measures the longest common *consecutive* string of characters that two strings have in common. This measure can be thought of as finding a word which contains the longest *n*-gram of characters in common with a given word. The formula we use for the LCST measure is a ratio as in the previous measure:

$$LCSTR(X,Y) = \frac{\text{length}[LCST(X,Y)]}{\max[\text{length}(X),\text{length}(Y)]} \tag{4.3}$$

- **Dice Similarity:**

$$dice = \frac{2 * LCST}{\text{length}(X) + \text{length}(Y)} \tag{4.4}$$

- **Needleman-Wunch Distance:**

$$needleman - wunch = \frac{LCST}{\min[\text{length}(X),\text{length}(Y)]} \tag{4.5}$$

- **Levenshtein Distance:** This method computes the minimum number of operations necessary to transform one string into the other. The allowable operations are insertion, deletion, and substitution. Compared to the previous methods this one returns a score x which is between 0 and n. The number n represents the maximum number of operation to convert an arbitrarily dissimilar NE to another one. To have a uniform score between all cognate methods we normalize x so that it lies between 0 and 1 using the following formula:

$$normalizeScore_{Levenshtein} = \frac{1}{1 + LevenshteinDistance} \tag{4.6}$$

Each of these cognate methods return a score between 0 and 1. We use a weighted linear combination to compute a final score for the pair of NEs:

$$finalScore_{EngNE-foreingnNE} = \sum_{i=1}^{n} cognate_i * weight_i \tag{4.7}$$

However, if the English NE and target language NE are spelled in the same way then these two are treated as a perfect mapping and none of the previous methods are applied on these two NEs.

4.2.1 Adaptation for the Greek language

The cognate methods assume that the characters in both English and foreign languages are the same. However, this is not the case for the Greek language. To also be able to apply our cognate based approach to the Greek language we first map the Greek NEs into English characters and apply the cognate metrics on the mapped characters.

We use a list of Greek-English place name variants[3] to learn character mappings. For this purpose we used the Giza++ tool. The input to Giza++ is a list of aligned NEs (Greek and English) where each NE is split into single characters. The output of the tool is a dictionary with character mappings. We use these mappings to transliterate a Greek NE into English characters and use the transliterated version for the cognate mapping. Usage of the UN/ELOT standard (http://en.wikipedia.org/wiki/Romanization_of_Greek) for transliterating Greek letters into English characters is also feasible.

4.2.2 Evaluation

To evaluate the performance of the NE mapper we use comparable corpora collected from the web. The corpora contain news article pairs for all ACCURAT languages. For evaluation purposes we used 100 randomly chosen document pairs from each language pair corpora. The documents in the foreign languages are annotated for NEs (PERSON, LOCATION and ORGANIZATION). The English comparable documents are NE tagged using the OpenNLP tools. As mentioned earlier the mapper outputs pairs of NEs with scores varying from 0 to 1. For the evaluation we only considered the pairs which had equal or higher score than a threshold which was set experimentally to 0.5. We showed the output of the mapper to native speakers who judged each pair as "correct" (meaning that the mapping is a correct mapping), as "partial correct" (meaning that the NEs are partially matching, e.g. only the first or last name of a person is matching but the other part not) and as "incorrect" (meaning that the mapping is not correct).

Please also note that the "partial correct" category is also used to judge pairs where the case information is different (e.g. *Europe* (EN) and *Europi* (HR)). Actually, this pair would be judged as correct if the context where these NEs are used were the same (*to the Europe* (EN) and *Europi* (HR)). However, in the evaluation we only showed the extracted pairs without any context. Thus, any case problematic pair was judge as "partial ok".

For each language pair we showed the results only to a person. The results of the evaluation are shown in Table 4.1.[4]

Table 4.1: **NE mapping evaluation.**

Language pair	Correct	Partial Correct	Incorrect	Total	Correct in %
en-lv	49	1	4	54	90%
en-lt	80	20	5	105	76%
en-ro	113	4	4	121	93%
en-de	141	11	7	159	88%
en-el	60	6	0	66	90%
en-hr	59	51	4	114	51%

[3] http://en.wikipedia.org/wiki/List_of_Greek_place_names

[4] The table does not show results for the Slovenian and Estonian languages because for these languages narrow domain systems are not planned.

4.3 Language Dependent Mapper - English-Romanian

NERA maps the named entities of the input bi-text based on two criteria: (i) a GIZA-like translation equivalents table and (ii) the Levenshtein distance between candidates. While the latter accounts for most of the PERSONs and ORGANIZATIONs, the former is indispensable when dealing with certain LOCATIONs (e.g.: Black Sea (en) vs. Marea Neagră (ro)). Currently, the application is tested using different parameters and thresholds.

4.4 Terminology mapping

Bilingual terminologies are important for various applications of human language technologies e.g. cross-language informational search and retrieval, adjusting machine translation to narrow domain, etc. During recent years automatic bilingual terminology mapping (and then extraction) in comparable corpora has received greater attention in view of scarceness of parallel data for under-resourced languages, and several methods were applied to this task e.g. contextual analysis [Rapp, 1995; Fung and McKeown, 1997], compositional analysis [Grefenstette, 1999; Daille and Morin, 2008]. In view of bilingual lexicon extraction symbolic, statistical and hybrid techniques have been implemented [Morin and Prochasson, 2011].

In this work we apply the same cognate-based approach as in language independent NE mapper to map terminologies. On English side an English terminology extractor is used. On the target one the ACCURAT specific tools are used. Extracted terminologies from both sides are aligned using cognate-based methods.

For English term extraction we use the *KEA TE* extractor[5]. On the foreign sites we use the TE tools described in Chapter 3. For the Greek TE we used the same adaptation approached as described in Section 4.2.1.

To evaluate the performance of the terminology mapper we used the same data set as in the evaluation scenario of NE mapping. The English data was tagged for terms using the KEA tool and the foreign documents were annotated using the tools described in Chapter 3. We run the mapper on these data sets. Unfortunately, we haven't obtained any mapping pairs. We think that one of the reasons for this is the training data used to train all the different systems. KEA mainly tags single words as terms whereas e.g. the CollEx tags terms containing at least 2 words. E.g. the Lithuanian TE tool tags *gyvos bakteriyos* or *baltas kivis* as terms in the Lithuanian documents and KEA does not do so in the English text (*live bacteria* and *white kiwi* although these terms occur in the English documents). These examples also show that the cognate methods alone might fail on these examples as they contain translations rather than transliterations. Thus for successful term mapping one also requires translation resources such as lexical dictionaries (see Chapter 5) to translate e.g. *gyvos* to *live*.

For NE we believe that there is a common understanding what e.g. persons or locations are so that the training data is uniformed across all the systems. However, for terms the training data is biased by the annotators' understanding about what constitutes a technical term. We believe that a clear term annotation framework has to be implemented in order to find intersections of terms between different systems applied on different languages. Finally, in NEs mostly transliteration is used instead of translation so that the cognate based methods work well on NE mapping.

[5] http://www.nzdl.org/Kea/

5 Extracting lexical dictionaries from comparable corpora to improve alignment and information extraction

The task of extracting translation equivalents from bilingual corpora has been approached in different manners, according to the degree of parallelism between the source and target parts of the corpora involved. For a well sentence aligned parallel corpora one can benefit from reducing the search space for a candidate translation to the sentence dimension and external dictionaries are not required. In the case of comparable corpora, the lack of aligned segments can be compensated by external dictionaries [Rapp, 1999] or by finding meaningful bilingual anchors within the corpus based on lexico-syntactic information previously extracted from small parallel texts [Gamallo, 2007].

The word alignment of parallel corpora has received significant scientific interest and effort (see for a review D2.1). There are already various free software aligners used in the industry and research, from which we mention only the well-known GIZA++ [Och and Ney, 2003]. Moreover, the error rate goes down to 9% in experiments made with some of these approaches [Och and Ney, 2003]. By comparison, the efforts and results in extracting bilingual dictionaries from comparable corpora are much poorer. Most of the experiments are usually done on small test sets, containing words with high frequency in the corpora (>99) and the accuracy percentages do not rise above 65%.

The most popular method to extract word translations from comparable corpora, on which we based the construction of our tool, is described and used in Fung and McKeown [1997]; Rapp [1999]; Chiao and Zweigenbaum [2002]. It relies on external dictionaries and is based on the following hypothesis:

*word **target1** is a candidate translation of **source1** if the words with which **target1** co-occurs within a particular window in the target corpus are translations of the words with which **source1** co-occurs with in the same window in the source corpus.*

The translation correspondences between the words in the window are extracted from the external dictionaries and considered seed word pairs.

Gamallo and Pichel [2005] used as seed expressions pairs of bilingual lexico-syntactic templates previously extracted from small samples of parallel corpus. This strategy led to a context-based approach, reducing the searching space from all the target lemmas in the corpus to all the target lemmas that appear in the same seed templates. In the improved version of the approach [Gamallo, 2008], the precision-1 (the number of times a correct translation candidate of the test word is ranked first, divided by the number of test words) and precision-10 (the number of correct candidates appearing in the top 10, divided by the number of test words) scores go up to 0.73 and 0.87 respectively.

In the following we will describe the algorithm implemented by our tool as introduced by Rapp [1999] and we will highlight the modifications and the adaptations we made, based on the experimental work we conducted.

5.1 The original approach

In a previous study, Rapp [1999] had already proposed a new criterion (the co-occurrence clue) for word alignment appropriate for non-parallel corpora. The assumption was that “there is a correlation between co-occurrence patterns in different languages” and he demonstrated in a study that this assumption is valid even for unrelated texts in the case of English-German language pair.

Starting from a more or less small seed dictionary and with the purpose of extending it based on a comparable corpus, a co-occurrence matrix is computed both for the source corpus and for the target corpus. Every row in the matrix corresponds to a type word in the corpus and every column corresponds to a type word in the base lexicon. The intersection of a row i and a column j in the co-occurrence matrix of the source corpus contains a frequency value of common occurrence of word i and word j in a window of pre-defined size.

The target and source corpora are lemmatized and POS-tagged. Function words are not taken into consideration for translation (they are identified by their POS closed class tags: pronouns, prepositions, conjunctions, auxiliary verbs, etc.).

For any row in the source matrix, all the words with which the co-occurrence frequency is bigger than 0 are sent for translation to the seed lexicon. An entry in the seed lexicon is identified by a unique identifier *id*. The unknown words (absent from the lexicon) are discarded and a vector of co-occurrence for the word correspondent to the row is computed versus the list of ids resulted after translation. The same procedure is applied to all the rows in the target matrix.

Experiments suggested the need to replace the co-occurrence frequency by measures able to eliminate word-frequency effects and favour significant word pairs. Measures with this purpose were previously based on mutual information [Church and Hanks, 1990], conditional probabilities [Rapp, 1996], or on some standard statistical tests, such as the chi-square test or the log-likelihood ratio [Dunning, 1993b]. In the approach we based our tool on, the measure chosen was the log-likelihood ratio computed as below:

$$\begin{aligned} LL(w_1,w_2) &= \sum_{i,j\in\{1,2\}} k_{ij} log \frac{k_{ij}N}{C_i R_j} \\ &= k_{11} log \frac{k_{11}N}{C_1 R_1} + k_{12} log \frac{k_{12}N}{C_1 R_2} + k_{21} log \frac{k_{21}N}{C_2 R_1} + k_{22} log \frac{k_{22}N}{C_2 R_2} \end{aligned} \tag{5.1}$$

where

$C_1 = k_{11} + k_{12}$
$R_1 = k_{11} + k_{21}$
$C_2 = k_{21} + k_{22}$
$R_2 = k_{12} + k_{22}$
$N = k_{11} + k_{12} + k_{21} + k_{22}$

and

k_{11} = frequency of common occurrence of w_1 and w_2 in a specific window in the corpus
k_{12} = corpus frequency of word w_1 - k_{11}
k_{21} = corpus frequency of word w_2 - k_{11}
k_{22} = size of corpus - corpus frequency of word w_1 - corpus frequency of word w_2

Finally, similarity scores are computed between all the source vectors and all the target vectors computed in the previous step, thus setting translation correspondences between the most similar source and target vectors. Different similarity scores were used in the variants of this approach; see Gamallo [2008] for a discussion about the efficiency of several similarity metrics combined with two weighting schemes: simple occurrences and log likelihood.

5.2 Adaptations, Experiments and Results

With the aim of obtaining a dictionary similar to a translation table of the type a decoder like Moses would need to produce its translation, we decided that the lines and columns of the matrixes will be populated in our approach by word forms and not by lemmas, as in the standard approach. The option for lemma entries in the matrix was assumed also by works like Gamallo and Pichel [2005] and Gamallo [2008].

As the purpose of this tool (and of all the other tools described so far) is to extract from comparable corpora data that would enrich the information already available from parallel corpora, it seems reasonable to focus on the open class (versus closed class) words. Obviously, this approach reduces the space and time necessities. Moreover, the closed class words we decided to ignore (pronouns, prepositions, conjunctions, articles, auxiliary verbs) are too general, they are not specific to any semantic field; therefore, they are not useful in an approach that is based on the tendency of some words to occur in the same semantic context as other words. Because in many languages, the auxiliary verbs can also be main verbs, frequently basic concepts in the language (see "be" or "have" in English), and most often the POS-taggers do not discriminate correctly between the two roles, we decided to eliminate their main verb occurrences as well. For this purpose, the user is asked to provide a list of all these types with all their forms in the language of interest, other than English.

We gave the user the possibility to specify the length of the text window in which co-occurrences are counted by modifying a parameter in the configuration file. We use a text window of length five as default.

Being based on word counting, the method is sensitive to the frequency of the words: the bigger the frequency, the better the performance. In previous works, the evaluation protocol was conducted on frequent words, usually on those with the frequency bigger than 100. Even in works presented by Gamallo [2008], where the evaluation was made on a list of nouns whose recall was 90% (those nouns that together come to the 90% of noun tokens in the training corpus), this corresponded to a bilingual lexicon constituted by $1,641$ noun lemmas, each lemma having a token $frequency >= 103$, for a bilingual comparable corpus of around 15 million tokens for each part. It doesn't seem too efficient to extract only a small amount of tokens from a big size corpus. Therefore, even if it brings loss of precision, the frequency threshold must be lowered when we are interested in extracting more data. In our tool, this parameter can be set by the user, according to his/her needs, but it should be bigger than 3 (our minimal threshold) and it should take into account the corpus dimension.

Our seed lexicon is based on a general domain translation table automatically extracted (with GIZA++) and this is consistent with the idea that we want to improve translation data obtained from parallel corpora. But as a consequence, we deal with high ambiguity and erroneous data in the seed lexicon. In the following you can see an excerpt from the base lexicon displaying all the possible translation for the word form "creates" with their translation probabilities in the last column (see Table 5.1).

Only the first three entries are exact translations of the word form "creates" while 3 of them ("instituie", "stabileşte" and, in a lesser extent, "ridicã" are acceptable translations in certain contexts). The two bold entries, "naştere" (birth) and "duce" (carries), may seem wrong translations learned from the training data, having a translation probability score similar to some correct translation (like "creând" or "crea"). We think we need to have access to all these possible translations as the semantic content of a linguistic construction is rarely expressed in another language through an identical syntactic or lexical structure. This is true especially in the case of a comparable corpus.

Table 5.1: **Translation for the word form "creates".**

72083 creates|crea|0.0196078
72084 creates|creează|0.686275
72085 creates|creând|0.0196078
72076 creates|duce|0.0196078
72087 creates|instituie|0.117647
72088 creates|naştere|0.0196078
72089 creates|ridică|0.0392157
72090 creates|stabileşte|0.0196078

Our solution was to distribute the log-likelihood of a word pair (w_1, w_2) in the source language to all the possible translations of w_2 in the target language as follows:

$$LL(w_1, w_2) = \sum_i LL(w_1, w_2) * p(w_2, t) \tag{5.2}$$

where $p(w_2, t)$ is the probability of a word w_2 to be translated by t_i and $sumlimits_i p(w_2, t) = 1$.

Every translation pair (w_2, t_i) is identified in the base lexicon by an unique *id*, making it possible to compute a similarity score across the languages.

For example, $LL(man, creates) = 12$ will be transformed in a list of *LLs* as following:

```
LL(man, 72083) = 12*0.0196078 = 0.2352936
LL(man, 72084) = 12*0.686275 = 8.2353
LL(man, 72085) = 12*0.0196078 = 0.2352936
LL(man, 72086) = 12*0.0196078 = 0.2352936
LL(man, 72087) = 12*0.117647 = 1.411764
LL(man, 72088) = 12*0.0196078 = 0.2352936
LL(man, 72089) = 12*0.0392157 = 0.4705884
LL(man, 72090) = 12*0.0196078 = 0.2352936
```

Previous to the LLs distribution, there is a step of LL filtering, in which all the words that occur with an LL smaller than a threshold are eliminated. This was motivated by the need to reduce the space and time computational costs and is also justified by the intuition that not all the words that occur at a specific moment together with another word are significant in the general context of our approach and the LL score is a good measure of this significance.

Following the conclusions of Gamallo [2008] experiments, we used as a vector similarity measure the *DiceMin* function:

$$diceMin(w_1, w_2) = \frac{2 * \sum_{k \in S_{ids} \cap T_{ids}} \min[LL(w_1, k), LL(w_2, k)]}{\sum_{i \in S_{ids}} LL(w_1, i) \sum_{j \in T_{ids}} LL(w_2, j)} \tag{5.3}$$

where S_{ids} and T_{ids} are the sets of dictionary entries identifiers with which w_1 and w_2 co-occur.

In computing the similarity scores, we did not allowed the cross-POS translation (a noun can be translated only by a noun, etc.); the user can decide if he/she allows the application to cross the boundaries between the parts of speech, through a parameter modifiable in the configuration

file. Each choice has its rationales, as we know that a word is not always expressed through the same part of speech when translated in another language. On the other hand, putting all the words in the same bag increases the number of computations and the risk of error.

Tests have been conducted on different sizes and different types/registers of comparable corpora:

1. A comparable corpora of small size representing the civil code of Romania (184081 words) vs. the civil code of Canada (199401 words).
2. A corpus of articles extracted from Wikipedia: 743194 words for Romanian, 809137 words for English.
3. Comparable corpora derived from the web (1396009747 words for English, 3764654484 words for Romanian).

We manually compiled a gold standard lexicon of around 1500 words (common nouns, proper nouns, verbs and adjectives) from the Wikipedia corpus. For these words the precision-1 and precision-10 scores introduced earlier were computed:

```
common nouns::
precision-1: 0.573948439620081 precision-10: 0.738127544097693
proper nouns::
precision1: 0.695652173913043 precision-10: 0.733695652173913
adjectives::
precision-1: 0.49438202247191 precision-10: 0.629213483146067
verbs::
precision-1: 0.662068965517241 precision-10: 0.827586206896552
```

6 Conclusion

In this report we first motivated the usefulness of named entities and technical terms for Machine Translation. Our hypothesis is that if two texts written in different languages share such linguistic terms then they are likely to share translation units. Those translation units can be used to extend the training data for Statistical Machine Translation. However, we argued that finding two corresponding linguistic terms in different texts of different languages is a challenging task. We focused on this challenge.

We first introduced different named entity extraction tools which are necessary for identifying corresponding named entities. We also described tools for identifying technical terms. Then we showed how to map such named entities and technical terms after they are pre-processed by these tools. We showed that identifying corresponding named entities in different languages works reasonably well. However, this is not the case for mapping technical terms. We argued that one of the reasons for this is the training data used in different term extraction systems. Each of these datasets contains terms which are result of distinct annotators' understanding of what "term" means rather than the outcome of a well-defined guideline. We believe if terms are tagged according to specific guidelines then mapping of these terms will lead to better results. We also observed that successful term mapping requires translation resources. Based on this observation, we also presented work about how to learn lexical dictionaries from comparable corpora.

In the future we plan to focus further on extracting translation units from comparable corpora using the linguistic term information as a guide. We plan to integrate the lexical dictionaries extracted from comparable corpora into the mapping workflow. We also plan to improve our tools and deliver a second version of the toolkit later in the project. The current toolkit that combines all the described systems/tools is explained in D2.6.

Bibliography

Al-Onaizan, Y. and Knight, K. (2002a). Machine transliteration of names in Arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–13. Association for Computational Linguistics.

Al-Onaizan, Y. and Knight, K. (2002b). Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408. Association for Computational Linguistics.

Aswani, N. and Gaizauskas, R. (2010). English-Hindi Transliteration using Multiple Similarity Metrics. In *7th Language Resources and Evaluation Conference (LREC), La Valletta, Malta*.

Basili, R., Moschitti, A., Pazienza, M., and Zanzotto, F. (2001). A contrastive approach to term extraction. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA), France*.

Bekavac, B. and Tadić, M. (2007). Implementation of croatian nerc system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 11–18. Association for Computational Linguistics.

Bender, O., Och, F., and Ney, H. (2003). Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 148–151. Association for Computational Linguistics.

Bonin, F., DellOrletta, F., Venturi, G., and Montemagni, S. (2010). A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Malta*, pages 19–21.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 977–981. Association for Computational Linguistics.

Chen, S. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Chiao, Y. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pages 1–5. Association for Computational Linguistics.

Chieu, H. and Ng, H. (2003). Named entity recognition with a maximum entropy approach. In *In CoNLL 2003, shared task.*

Choueka, Y. (1988). Looking for needles in a haystack. In *Proceedings of RIAO88*, pages 609–623.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1989). Parsing, word associations and typical predicate-argument relations. In *Proceedings of the workshop on Speech and Natural Language*, pages 75–81. Association for Computational Linguistics.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, USA. Association for Computational Linguistics.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. The MIT Press, 2nd revised edition.

Curran, J. and Clark, S. (2003). Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 164–167. Association for Computational Linguistics.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40. Association for Computational Linguistics.

Daille, B. (1994). Combined approach for terminology extraction: lexical statistics and linguistic filtering.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1:49–66.

Daille, B. and Morin, E. (2008). Effective Compositional Model for Lexical Alignment. *In Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India*.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447.

Dunning, T. (1993a). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Dunning, T. (1993b). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Ehrmann, M. and Turchi, M. (2010). Building Multilingual Named Entity Annotated Corpora exploiting Parallel Corpora. *Workshop on Annotation and Exploitation of Parallel Corpora (AEPC), University of Tartu, Estonia*.

Evert, S. (2005). The statistics of word cooccurrences. *Word Pairs and Collocations. Phil. Diss. Institut f*
"ur maschinelle Sprachverarbeitung. Stuttgart.

Feng, D., Lv, Y., and Zhou, M. (2004). A new approach for english-chinese named entity alignment. In *Proc. of EMNLP*, pages 372–379.

Finkel, J., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Firth, J. and Palmer, F. (1968). *Selected papers of JR Firth, 1952-59*. Indiana University Press.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Frantzi, K. and Ananiadou, S. (1997). Automatic term recognition using contextual clues. *In Proceedings of Mulsaic 97, IJCAI, Japan*.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Gamallo, P. (2007). Learning bilingual lexicons from comparable English and Spanish corpora. *Proceedings of MT Summit XI*, pages 191–198.

Gamallo, P. (2008). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora*, pages 19–26. Citeseer.

Gamallo, P. and Pichel, J., R. (2005). An Approach to Acquire Word Translations from Non-Parallel Texts. *Lecture Notes in Computer Science, vol. 3808.Springer Verlag*.

Georgantopoulos, B. and Piperidis, S. (2000a). A Hybrid Technique for Automatic Term Extraction. In *In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications-ACIDCA'2000*, pages 124–128.

Georgantopoulos, B. and Piperidis, S. (2000b). A hybrid technique for automatic term extraction. In *In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications-ACIDCA'2000*.

Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. *Translating and the Computer 21, London, UK*.

Grigonyte, G., Rimkute, E., Utka, A., and Boizou, L. (2011). Experiments on Lithuanian Term Extraction. *In Proceedings of NODALIDA 2011 Conference, May 11-13, 2011, Riga, University of Latvia, Latvia, p. 82-89*.

Huang, F., Vogel, S., and Waibel, A. (2003). Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*, pages 9–16. Association for Computational Linguistics.

Ion, R. (2007). Word Sense Disambiguation Methods Applied to English and Romanian. *Romanian Academy, Bucharest*.

Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Justeson, J., S. and Katz, S., M. (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. In *Natural Language Engineering (1), pp. 9-27, Cambridge University Press*.

Klein, D., Smarr, J., Nguyen, H., and Manning, C. (2003). Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 180–183. Association for Computational Linguistics.

Klementiev, A. and Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics.

Kochanski, I. (1995). Lecture 4 - Good-Turing probability estimation, Oxford. http://kochanski.org/gpk/teaching/0401Oxford/GoodTuring.pdf.

Kruglevskis, V. (2010). Semi-Automatic Term Extraction from Latvian Texts and Related Language Technologies. *Magyar Terminologia (Journal of Hungarian Terminology)*.

Kruglevskis, V. and Vancane, I. (2005). Term Extraction from Legal texts in Latvian. *In Proceedings of the Second Baltic Conference on Human Language Technologis, April 4-5, 2005*.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Manning, C., Schutze, H., and MITCogNet (1999). *Foundations of statistical natural language processing*, volume 59. MIT Press.

Moore, R. (2003). Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 259–266. Association for Computational Linguistics.

Morin, E. and Prochasson, E. (2011). Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. *ACL HLT 2011*, page 27.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nenadié, G., Ananiadou, S., and McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 604–es. Association for Computational Linguistics.

Ng, H., Goh, W., and Low, K. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–73. ACM.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Paukkeri, M., Nieminen, I., Polla, M., and Honkela, T. (2008). A language-independent approach to keyphrase extraction and evaluation. In *Proceedings of COLING*, pages 83–86. Citeseer.

Pavel, P. (2009). Lexical association measures: Collocation Extraction. *Studies in Computational and Theoretical Linguistics. Institute of Formal and Applied Linguistics, Prague, Czech Republic*.

Petrovic, S., Snajder, J., and Basic, B. (2010). Extending lexical association measures for collocation extraction. *Computer Speech & Language*, 24(2):383–394.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Rapp, R. (1996). *Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz*, volume 16. Georg Olms Verlag.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Salton, G. and Lesk, E., M. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schutze, H. (1998). The hypertext concordance: a better back-of-the-book index. In *Proceedings of First Workshop on Computational Terminology*. Citeseer.

Skadina, I., Veisbergs, A., Vasiljevs, A., Gornostay, T., Keisa, I., and Rudzite, A. (2011). Languages in the European Information Society - Latvian. *META-NET White Paper Series. Early Release Edition, META-FORUM 2011, June 27-28, 2011, Budapest, Hungary*.

Skujina, V. (2010). Latviesu terminologijas izstrades principi: Morfologiskais aspekts. *Raksti un gramatas, 2010*.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

Stalls, B., Knight, K., et al. (1998). Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41.

Stefanescu, D. (2010). Intelligent Information Mining from Multilingual Corpora, PhD thesis (Romanian). *Romanian Academy, Bucharest*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Arxiv preprint cs/0609058*.

Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2008). Racais linguistic web services. In *Proceedings of the 6th Language Resources and Evaluation Conference–LREC*. Citeseer.

Tufiş, D. and Irimia, E. (2006). RoCo-News: A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy*, pages 22–28. Citeseer.

Wermter, J. and Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 843–850. Association for Computational Linguistics.

Wong, W., Liu, W., and Bennamoun, M. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 47–54. Australian Computer Society, Inc.

Zabarskaite, J. and Vaisniene, D. (2011). Languages in the European Information Society - Lithuanian. *META-NET White Paper Series. Early Release Edition, META-FORUM 2011, June 27-28, 2011, Budapest, Hungary*.

Zeller, I. (2005). Automatinis terminu atpazinimas ir apdorojimas. *PhD thesis*.