Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages

Mārcis Pinnis[1], Nikola Ljubešić[2], Dan Ştefănescu[3], Inguna Skadiņa[1], Marko Tadić[2], Tatiana Gornostay[1]

Tilde[1],
Faculty of Humanities and Social Sciences, University of Zagreb[2],
Research Institute for Artificial Intelligence, Romanian Academy[3]
marcis.pinnis@tilde.lv, nikola.ljubesic@ffzg.hr,
danstef@racai.ro, inguna.skadina@tilde.lv, marko.tadic@ffzg.hr,
tatiana.gornostay@tilde.lv

Abstract. Although term extraction has been researched for more than 20 years, only a few studies focus on under-resourced languages. Moreover, bilingual term mapping from comparable corpora for these languages has attracted researchers only recently. This paper presents methods for term extraction, term tagging in documents, and bilingual term mapping from comparable corpora for four under-resourced languages: Croatian, Latvian, Lithuanian, and Romanian. Methods described in this paper are language independent as long as language specific parameter data is provided by the user and the user has access to a part of speech or a morpho-syntactic tagger.

Keywords: term extraction, term tagging, term mapping, under-resourced languages, comparable corpora

1 Introduction

Term extraction (TE) has been the focus of extensive work in natural language processing for almost 20 years. Approaches may be characterised according to whether they use local grammars, statistical co-occurrence measures, or a combination of the two.

Systems like LEXTER [3], TERMS [13], and Termight [4] use primarily local grammar approaches in the form of hand-authored regular expressions over part-of-speech tags, while systems like Pantel and Lin [19] make no use of linguistic information at all, using solely statistical co-occurence measures between words. Often both approaches are combined in hybrid methodologies [5, 4, 11].

Despite the long history of term extraction, TE tools for Central and East European languages appeared later. Even nowadays there is a significant gap between European analytical languages, on the one side, and synthetic ones, on the other side, due to their under-resourced status with the lack of necessary language resources and tools [15].

For the Croatian language, the first experiments on collocation extraction and TE were presented by Tadić and Šojat [26] using pointwise mutual information as the statistical co-occurence measure for detecting collocations and multi-word term candidates. The *TermeX* system [7] developed later for Croatian and English provides the possibility to use nine different co-occurrence measures for collocations.

For the Lithuanian language, the first experiments on TE were described by Zeller [28]. Grigonyte et al. [12] evaluated the extraction of domain-specific terminology using four approaches: keyword cluster identification, keyword extraction with machine learning, collocation extraction, and grammar-based. The collocation extraction and grammar-based approach appeared to be reliable in terms of recall, but not precision.

For the Latvian language, the first experiment on TE showed that the linguistic method based on morpho-syntactic analysis is more appropriate than the statistical one that proved to be adequate for analytical languages [16]. A semi-automatic TE has been applied to Latvian texts recently [15].

In term tagging, the question "What is a term?" must be addressed not only from the termhood view but also from the unithood, i.e., syntagmatic nature of a term, in case of the so-called nested terms in particular – "those <terms> that appear within other longer terms and may or may not appear by themselves in the corpus" [17] (cf. [8, 14]). In the resulting term candidate list there might be overlaps between term candidates with different length. According to our application-oriented approach, only one of the nested term candidates is considered a valid term (see the example in section 3).

Automatic bilingual term mapping from comparable corpora has received greater attention recently. Methods like contextual analysis [9] and compositional analysis [10, 6] are applied to this task. In view of bilingual lexicon extraction, symbolic, statistical, and hybrid techniques have been implemented [18]. However, term mapping for morphologically rich under-resourced languages received less attention in research [27].

In this paper we present a workflow allowing the extraction of term candidates from text documents (for instance, news articles, technical manuals, knowledge base articles like Wikipedia, etc.), term tagging in the documents (giving evaluation for Croatian, Latvian and Lithuanian), and bilingual term mapping in comparable corpora for four under-resourced languages: Croatian, Latvian, Lithuanian, and Romanian. The paper features also a real world scenario on how to acquire bilingual terms from comparable Web crawled narrow domain corpora.

However, methods described in this paper are language independent and require that the user has access to a part of speech tagger (in order to pre-process text documents for a particular language) and language specific data. That is, the user must have access to: a stopword list, a phrase table for valid term patterns, a list of lemmas with corresponding inverse document frequency (IDF) [22] scores calculated on a large general domain corpora (for instance, all Wikipedia articles of the required language), and an optional bilingual single-word probabilistic dictionary for higher recall bilingual term mapping.

2 Term Candidate Extraction with CollTerm Tool

CollTerm is a tool for collocation and term extraction, and it combines two major approaches: (a) a linguistically motivated approach via morpho-syntactic patterns and (b) a statistically motivated approach via co-occurrence statistics and reference corpus statistics. The diagram of *CollTerm* and its processing flow, as depicted in Fig. 1, defines four processing steps of the system: (a) linguistic (morpho-syntactic) filtering, (b) minimum frequency filter, (c) statistical ranking, and (d) cut-off method.

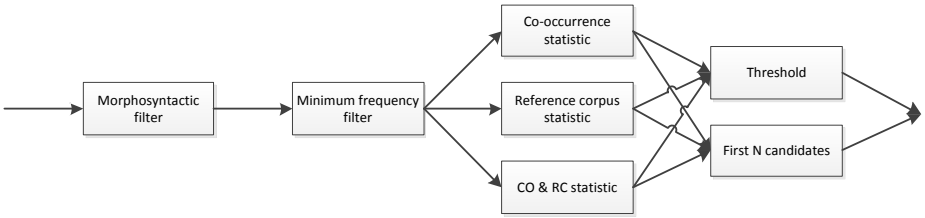

Fig. 1. Diagram of the *CollTerm* processing flow

2.1 Linguistic Filtering

CollTerm starts with linguistic filtering. Linguistic patterns of term candidates are defined in a phrase table containing regular expressions of acceptable phrases (see Fig. 2). In order to find valid term candidates, the regular expressions have to be crafted using the same tagset that is used by the morpho-syntactic (or part of speech) tagger during pre-processing of text documents. Since we deal mostly with morphologically rich languages the phrase table contains a set of syntactic patterns of a language and allows for the definition of agreement between its constituents (as far as the tagset of the morpho-syntactic tagger allows).

```
^[AG].fsn.*     ^N...g.*     ^N.fsn.*
^[AG].fsg.*     ^N...g.*     ^N.fsg.*
^[AG].fsd.*     ^N...g.*     ^N.fsd.*
^A.msg.*        ^N.msg.*     ^N.*
^A.mpg.*        ^N.mpg.*     ^N.*
```

Fig. 2. Fragment of Latvian morpho-syntactic patterns defining agreement between adjective (A) and noun (N) in gender (m-masculine, f-feminine), number (s- singular, p-plural) and case (n –nominative, g-genitive, d-dative)

Additionally, a stopword list can be used to filter out invalid term candidates. Stopword position restrictions can be specified in the phrase table. The example in Fig. 3 states that stopwords are not allowed to be the first and last token of tri-gram and four-gram term candidates.

```
!STOP       *          !STOP
!STOP       *          *          !STOP
```

Fig. 3. Example of morpho-syntactic patterns with stopword restrictions in a phrase table

The pattern lists for Latvian and Lithuanian contain 120 different patterns. Initially these patterns were automatically extracted from morphologically tagged texts [20] in which terms were marked by human annotators. Since this initial list contained patterns for specific cases and not general language rules, the obtained patterns were then manually revised and generalised.

2.2 Minimum Frequency Filter

The second phase consists of the minimum frequency filter where all linguistically accepted phrases occurring less than the set minimum frequency are discarded from further processing, to prevent the necessity of the manual intervention of a domain expert and/or a terminologist who would evaluate the produced list and extract relevant terms (cf. frequency threshold in [8]). As document length and term frequency distribution vary from domain to domain, the minimum frequencies of acceptable term candidates have to be tuned for each type of document (for instance, technical manuals are usually longer than news articles and may have higher minimum frequencies). Evaluation results presented in section 3 show that application of different minimum frequencies influences the recall and precision of term tagging.

2.3 Statistical Ranking

The third phase performs ranking of term candidates concerning co-occurrence or reference corpus statistics. For co-occurrence statistics five different statistic methods can be used:

- the Dice coefficient (DICE);
- modified mutual information (MI);
- the chi-square statistics (CS);
- log-likelihood (LL);
- the t-score statistic (TS).

Evaluation of these methods within the term tagging task for Croatian, Latvian and Lithuanian is discussed in section 3. Table 1 shows the top 10 normalised bigram term candidates extracted from the Wikipedia article “Automobile” using t-score statistics with a minimum frequency of three for English and two for Latvian and Lithuanian (the minimum frequencies differ due to the article length difference).

Since unigrams can’t be ranked via co-occurrence statistics, reference corpus word lemma IDF[22] scores can be provided as additional input information. The reference corpus has to be large enough in order to represent the language (in terms of stopwords in contrast to words that may be important in term extraction). For instance, the

Latvian corpus from which lemma IDF scores have been extracted consists of Wikipedia articles (7.6 million tokens) and Web news articles (8.2 million tokens). Unigram term candidate ranking is calculated as a multiplication of the term's frequency within a document and the corresponding IDF score (TF-IDF [22]). The TF-IDF ranking can also be applied to n-grams of length greater than one. In that case, an average IDF score for an n-gram is calculated.

Table 1. Top 10 normalised English, Latvian, and Lithuanian term candidates consisting of two words and their scores obtained with the t-score statistic

English bigram term candidates		Latvian bigram term candidates		Lithuanian bigram term candidates	
driverless car	1.00	caurejamības automobilis	1.00	antiblokavimas sistema	1.00
Propulsion technology	0.84	iekšdedze dzinējs	0.66	benzininis variklis	0.93
internal combustion	0.83	protektors raksts	0.57	degimas variklis	0.87
combustion engine	0.75	lauksaimniecība traktors	0.52	variklis cilindras	0.85
automotive industry	0.73	tvaiks dzinējs	0.49	sauga diržas	0.84
automotive market	0.64	ciets segums	0.48	dyzelinis variklis	0.82
light truck	0.48	krava pārvadāšana	0.46	Lenktyninis Automobilis	0.78
assembly line	0.40	dzinējs automobilis	0.38	vidus degimas	0.77
automobile use	0.37	sacīkstes automobilis	0.37	vairas mechanizmas	0.75
main article	0.36	ātrums rekords	0.33	įpurškimas sistema	0.72

If the IDF score file is given and a co-occurrence statistic is used for n-gram term candidate ranking, a linear combination of TF-IDF and that co-occurrence statistic is computed. In the case where a non-dummy phrase file with linguistic patterns is given, the term candidates are extracted and ranked using all three information sources – the linguistic, the statistical representing co-occurrence data, and the statistical representing reference corpus data.

2.4 Cut-off Method

In the fourth phase two different "cut-off" methods can be applied on the ranked candidate term list:

- application of a term candidate ranking threshold (every term candidate below the threshold will be filtered out);
- extraction of the first N candidates.

The threshold "cut-off" method is more robust. It is less affected by document length differences and whether the document contains more or less valid term candidates after linguistic filtering.

The resulting list of term candidates can be exported as a sequence of lemmas (suitable for term tagging) or a sequence of most frequent phrases in text (more suitable for human inspection) with or without the lemma rankings.

3 Term Tagging in Documents

The *CollTerm* tool provides a document with a term candidate list of fixed length (up to four tokens) where n-grams (phrases) are ranked according to one of the ranking methods. This requires *CollTerm* to be executed multiple times to cover single and multi-word terms. However, the resulting term candidate list contains lexical overlaps between term candidates with different length, the so-called nested terms (cf. [8]). Consider the following example: "A crash course in physics". As an output *CollTerm* might find two term candidates: a unigram term candidate "crash" and a bigram term candidate "crash course" (both may be correct according to the context). However, in order to capture a more specific representation of terms in the source document, only one of the term candidates is a valid term, e.g., in the example above, an intuitive selection is "crash course" if the document is about education. Our approach is application-oriented: in case of machine translation, e.g., the less specific term may cause wrong translation.

Due to ambiguities, we treat the term candidate lists as intermediate data and tag the terms in the source document with the tool *Tilde's Wrapper System for CollTerm* (*TWSC*). *TWSC* takes as input plaintext or pre-processed tab-separated (broken into sentences, tokenised, and part of speech or morpho-syntactically tagged) documents. *TWSC* then produces either term tagged plaintext where term candidates are marked with tags <TENAME> (see Fig. 4 for an example) or tab-separated documents (see Fig. 5 for an example) where term candidates are marked with tags B-TERM (for the first token) and I-TERM (for the rest of tokens).

<TENAME>Servisa aprīkojumā</TENAME> ietilpst <TENAME>bremžu pārbaudes stends</TENAME>, <TENAME>motora diagnostikas ierīce</TENAME>, <TENAME>riteņu balansēšanas stends</TENAME>, <TENAME>amortizatoru pārbaudes stends</TENAME>, <TENAME>riteņu montēšanas stends</TENAME> u.c.

Fig. 4. Fragment of a term-tagged plaintext document in Latvian

Within one term candidate list, it is possible to select the term candidate that is ranked higher. However, if the overlap is between candidates of different lists, the selection is not straightforward. During our experiments we have applied two methods for combining different n-gram term candidate lists into one. The first approach prioritises the longer n-grams, while the second approach combines all lists within one list using linear interpolation of term candidate confidence scores.

Servisa	N	serviss	N-msg---------n----------f-	28	111	28	117	B-TERM	0.37
aprīkojumā	N	aprīkojums	N-msl---------n----------l-	28	119	28	128	I-TERM	0.37
ietilpst	V	ietilpt	Vp----3--i----------l-	28	130	28	137	O	0
bremžu	N	bremze	N-fpg---------n----------l-	28	139	28	144	B-TERM	0.45
pārbaudes	N	pārbaude	N-fsg---------n----------l-	28	146	28	154	I-TERM	0.45
stends	N	stends	N-msn---------n----------l-	28	156	28	161	I-TERM	0.45
,	T	,	T----------,	28	162	28	162	O	0

Fig. 5. Fragment of a term-tagged tab-separated document in Latvian

3.1 Term Tagging Evaluation for Latvian and Lithuanian

Evaluation of the term tagging tool *TWSC* for Latvian and Lithuanian was performed on manually annotated texts in the IT domain (software manuals, IT news, software reviews, etc.). The human annotated corpora were split in two parts – a development set and a test set. The former was used for tuning of different parameters of *CollTerm* and *TWSC* including: (a) minimum n-gram frequencies, (b) *CollTerm* confidence score thresholds, and (c) linear interpolation coefficients for the second term candidate list combination method. The human annotated corpora statistics of the Latvian and Lithuanian corpora are given in Table 2 below.

Table 2. Latvian and Lithuanian human annotated corpora statistics

	Latvian		Lithuanian	
	Test set	Development set	Test set	Development set
Tokens	15 230	7 795	4 547	2 339
Proportion	66.15%	33.85%	66.03%	33.97%
Terms	2 362	1 127	751	380
Unigram terms	1 540	656	417	198
Multi-word terms	822	471	334	182

During evaluation parameters were tuned on the development set using an iterative approach. At first we tuned the minimum n-gram frequency constraints using the prioritised list combination method and evaluated which ranking methods achieve the highest precision, recall, and F-measure (F1) without application of *CollTerm* confidence score thresholds. Then term candidate confidence score thresholds were tuned in order to achieve better performance. Results using various term candidate ranking methods on the Latvian and Lithuanian test sets are given in Table 3.

The results show that for Latvian the best recall was achieved with the LL ranking method (70.66%), the best precision was achieved with the CS statistic (59.85%), and the best F-measure was achieved with the MI ranking method (54.05). The difference between the different methods is, however, relatively insignificant. For instance, the best achieved F-measure without confidence score threshold tuning with the LL statistic is 54.26 (54.23 on the development set) and with the DICE statistic - 54.05 (54.35

on the development set). As the development set for the Lithuanian language is relatively small, for Lithuanian all term candidate ranking methods produced identical results. Thus, for further tuning of parameters for Lithuanian the MI statistic was selected.

Table 3 also shows that threshold tuning on the Latvian development set improves results (in terms of recall, precision, and F-measure) on the test set as well. Although the evaluation shows an F-measure drop for Lithuanian, we believe that the size of the tuning corpus needs to be increased in order to reliably tune the parameters.

Finally, we tuned the interpolation parameters in order to achieve better F-measure with the interpolation-based term candidate list combination method. The results in Table 3 suggest that the prioritisation method significantly outperforms the interpolation-based method. Moreover, the tuned parameters suggest that longer n-grams are preferred (even in the interpolation-based method).

Table 3. Term tagging evaluation results for Latvian and Lithuanian

Language	Configuration	Term candidate ranking method	Minimum n-gram frequency for n-grams up to length 4				R	P	F1
Latvian	No threshold tuning	LL	1	1	3	3	70.66	42.52	53.09
		MI	2	1	1	2	63.89	46.83	54.05
		CS	11	3	2	3	39.88	59.85	47.87
	Threshold tuning	LL	1	1	3	3	**71.04**	41.70	52.55
		MI	2	1	1	2	57.49	52.74	**55.01**
		CS	11	3	2	3	23.24	**64.14**	34.12
	Prioritized	MI	2	1	1	2	63.89	46.83	54.05
	Linear interpolation	MI	2	1	1	2	63.04	42.58	50.83
Lithuanian	No threshold tuning	MI	1	1	1	1	65.11	46.97	54.57
		MI	4	1	2	2	59.79	53.26	**56.34**
		MI	10	3	2	3	42.08	55.24	47.77
	Threshold tuning	MI	1	1	1	1	**65.78**	47.78	55.35
		MI	4	1	2	2	55.79	52.70	54.20
		MI	10	3	2	2	37.55	**56.97**	45.26
	Prioritized	MI	4	1	2	2	59.79	53.26	56.34
	Linear interpolation	MI	4	1	2	2	60.32	41.79	49.37

The lower performance of the interpolation-based method can partially be explained with the fact that in the term candidate extraction step not only a lot of false term-candidates are filtered out, but also some good term candidates can be filtered due to selection of wrong phrase pattern for overlapping terms. For example, for Latvian and Lithuanian term extraction we use a morpho-syntactic tagger, which makes it possible to define more complex phrase patterns requiring morpho-syntactic property agreements (for instance, agreement in gender, number, and case). Therefore, in many cases, longer n-grams are already valid terms.

The tuning of parameters is very important when it is necessary to tune the system for specific tasks (for instance, document alignment, term mapping and alignment, information retrieval, question answering, etc.), because different tasks may require either higher recall or higher precision.

3.2 Term Tagging Evaluation for Croatian

The evaluation for Croatian was performed on a manually annotated corpus of automotive texts containing 15 603 tokens and 1 430 (849 single word and 581 multiword) tagged terms, of which 652 were unique terms. While working on the Croatian data, we took into account the conclusions drawn from the evaluation on Latvian and Lithuanian by starting the tuning process with MI as the co-occurrence statistic and using the prioritisation method by tagging the terms in a greedy fashion. Besides tuning the parameters for Croatian, we also focused our efforts on the effects of the morpho-syntactic phrase patterns used in linguistic filtering.

We first removed 32 tags longer than 4 tokens from the corpus and split it into a development set (7772 tokens and 645 terms) for tuning and a test set (7831 tokens, 753 terms) for final evaluation.

During the whole tuning process, we were maximising F-measure. The tuning was done in an iterative fashion similar to Latvian and Lithuanian. We started by searching for the optimal n-gram frequency thresholds. In this iteration, we improved the F-measure on the development set from 27.2 to 36.6. The next iteration focused on the optimal co-occurrence statistic and its threshold values. In this step, F-measure was improved from 36.6 to 44.7. It is important to stress that the thresholds had a much higher impact on the performance increase than the statistic itself.

Finally, we evaluated the approach on our test set. We added the tuned parameter values one by one and observed thereby the impact of the tuning process in a more objective fashion. The results are given in Table 4. Obviously both tuning steps improve results significantly.

Table 4. Term tagging evaluation results for Croatian by gradually applying tuned parameters

Minimum n-gram frequency for n-grams up to length 4				Term candidate ranking method	P	R	F1
-				-	17.33	79.55	28.46
5	2	2	1	-	24.20	41.17	30.48
5	2	2	1	LL	39.07	35.59	37.25

An additional insight that we wanted to obtain during our work on Croatian data is the importance of the valid term phrase patterns. For that reason we built three versions of the patterns:

1. 24 detailed morpho-syntactic patterns. The example below specifies a four token term phrase consisting of a noun phrase (adjective + noun) in any case with an additional genitive noun phrase (adjective + noun) attached to it:

   ^A.* ^Nc.* ^Af...g.* ^Nc..g.*

2. 12 more general rules obtained by simplifying the initial ones to just part of speech information (only the first letter of the morpho-syntactic tag). The example below describes the simplified previous example:

   ^A.* ^N.* ^A.* ^N.*

3. 4 rules allowing any morpho-syntactic pattern combination. The example defines a four token phrase without any restrictions to morpho-syntactic properties:

   .* .* .* .*

Results obtained on the test set with these three phrase files are given in Table 5.

Table 5. Term tagging evaluation results for Croatian

Phrase file	P	R	F1
1	39.07	35.59	37.25
2	41.19	35.99	38.41
3	4.55	24.17	7.66

These results show that the simplified phrase file did even slightly outperform the initial one (probably because of some morpho-syntactic annotation errors). The finding that almost identical results can be achieved by using linguistic filtering based only on part-of-speech information is very important since detailed morpho-syntactic taggers are not always available for under-resourced languages. However, the question remains if with more detailed phrase patterns, such as those applied on Latvian and Lithuanian (24 vs. 120 phrase patterns) would still increase the tagging quality in terms of precision. On the other hand, no linguistic filtering at all deteriorates the results drastically which shows the big impact the linguistic filtering step has on the term tagging task.

4 Term Mapping

To find possible translation equivalents of terms tagged in bilingual comparable corpora, the term mapping tool *TerminologyAligner* (*TEA*) was developed. Given term-tagged bilingual document pairs, the term mapping tool is designed to extract two lists of terms and to find pairs of expressions, which are reciprocal translations. The tool analyses candidate pairs, assigning them translation scores based on (a) the translation equivalents and (b) the cognates that can be found in those pairs:

$$translationScore(pair) = max(ete(pair), ecg(pair)) \quad (1)$$

In this case, *ete*(*pair*) is the translation equivalence score and *ecg*(*pair*) is the cognate score for the expressions forming the candidate pair.

The translation equivalence score for two expressions is computed based on the word-level translation equivalents. Each word w_s in the source terminological expression e_s is paired with its corresponding word w_t in e_t such that the translation probability is maximal, according to a Giza++ like probabilistic unigram translation diction-

ary. The score should be normalised with the length of expression e_s. Still, we modify the denominator in order to penalise the pairs according to the length difference between source and target expressions:

$$ete(e_s, e_t) = \frac{\sum_{w_s \in e_s} \max_{w_t \in e_t} wte(w_s, w_t)}{length(e_s) + \frac{|length(e_s) - length(e_t)|}{2}} \quad (2)$$

The cognate score for two expressions is computed as a modified *Levenshtein* distance (LD) between them. The expressions are normalised by removing double letters and replacing some character sequences: "*ph*" by "*f*", "*y*" by "*i*", "*hn*" by "*n*" and "*ha*" by "*a*". This type of normalisation is often employed by spelling and alteration systems [24]. Moreover, the score takes into account the length of the longest common substring of the two expressions, normalised by the maximum value of their lengths:

$$ecg(e_s, e_t) = \frac{1 - \frac{LD\left(normalize(e_s), normalize(e_t)\right)+1}{\min(length(e_s)+1, length(e_t)+1)} + \frac{length(LCS(e_s, e_t))}{\max(length(e_s), length(e_t))}}{2} \quad (3)$$

As probable translation equivalents, term pairs are selected only if the score of *ete*(*pair*) or *ecg*(*pair*) for the bilingual term pair is higher than a specified threshold. The value of the threshold regulates the trade-off between precision and recall of *TEA*.

Table 6. TEA evaluation results for English-Latvian on the Eurovoc thesaurus using in-domain and out-of-domain translation dictionaries

Threshold	In-domain dictionary			Out-of-domain dictionary		
	R	P	F1	R	P	F1
0.0	2.10	2.10	2.10	2.10	2.10	2.10
0.1	3.46	3.48	3.47	3.90	3.96	3.93
0.2	9.39	10.21	9.78	8.84	10.87	9.75
0.3	21.86	29.71	25.19	15.4	28.06	19.89
0.4	29.66	53.76	38.23	**18.11**	**55.00**	**27.25**
0.5	**31.03**	**79.52**	**44.64**	13.74	79.97	23.45
0.6	23.48	89.66	37.22	7.47	85.52	13.75
0.7	15.92	98.54	27.41	4.81	96.46	9.16
0.8	9.92	99.12	18.03	3.59	96.44	6.92
0.9	5.62	98.96	10.64	2.75	97.91	5.35
1.0	3.63	98.41	7.01	2.62	97.80	5.10

In order to evaluate the precision and recall of *TEA*, we used the Eurovoc thesaurus, which is "the thesaurus covering the activities of the EU and the European Parliament in particular" [25]. The Eurovoc thesaurus contains a total of 6 797 unique bilingual terms for every language pair (English-Croatian, English-Latvian, English-Lithuanian and English-Romanian). For the English-Latvian language pair, we used

two different unigram translation dictionaries to show the difference in recall when an in-domain or an out-of-domain dictionary is used.

The results (given in Table 6) show a significantly higher recall if an in-domain dictionary is used (a maximum of 31.03%), in contrast to an out-of-domain dictionary (a maximum of 18.11%). The obvious advantage to using the in-domain translation dictionary is a higher maximum precision of 99.12%, in contrast to 97.91% for the out-of-domain dictionary. However, we believe that in a real life scenario the user won't have an in-domain dictionary at his or her disposal when trying to map terms in an under-resourced domain. Therefore, the recall and precision will be closer to the results obtained with the out-of-domain translation dictionary.

For other language pairs we used only one translation dictionary (see Table 7 below). The results show that the highest F-measure is achieved for English-Romanian (23.48) followed by English-Croatian (21.66), and the lowest results have been achieved for English-Lithuanian (an F-measure of 19.99). For comparison, using a different in-domain dictionary (with higher term coverage) on English-Romanian *TEA* achieves an F-measure of 51.1 [23].

Table 7. TEA evaluation results for English-Lithuanian, English-Croatian, and English-Romanian on the Eurovoc thesaurus

Threshold	English-Lithuanian (in-domain)			English-Croatian (out-of-domain)			English-Romanian (in-domain)		
	R	P	F1	R	P	F1	R	P	F1
0.0	1.79	1.79	1.79	3.94	3.94	3.94	6.08	6.08	6.08
0.1	2.91	3.07	2.99	5.02	5.35	5.18	7.22	7.54	7.38
0.2	5.40	7.40	6.24	7.31	9.93	8.42	9.08	10.31	9.65
0.3	9.96	25.52	14.33	11.71	28.92	16.67	12.06	19.35	14.86
0.4	**12.27**	**53.84**	**19.99**	**13.49**	**54.88**	**21.66**	14.21	38.36	20.74
0.5	10.37	79.21	18.34	12.08	81.94	21.05	**14.24**	**66.8**	**23.48**
0.6	7.00	93.15	13.03	8.50	95.54	15.62	12.81	88.34	22.38
0.7	5.00	96.87	9.51	6.50	98.66	12.20	10.11	95.82	18.29
0.8	3.35	98.28	6.49	4.99	99.41	9.50	8.37	99.13	15.44
0.9	2.15	99.32	4.21	4.08	99.64	7.83	6.19	99.76	11.66
1.0	1.47	80.00	2.89	4.00	99.63	7.69	6.06	99.76	11.43

5 Real World Scenario

In order to show the capabilities of the term extraction, tagging and mapping process chain, we have run a full experiment on a English, Latvian comparable Web crawled corpus in the automotive domain (car service manuals, reviews, marketing materials, etc.). The corpus was collected using the Focused Monolingual Crawler (FMC) [2] and then bilingually aligned at the document level using the DicMetric [1] comparability metric tool. *TWSC* was used to tag terms in both English and Latvian docu-

ments. In order to tag terms in English documents, the documents were pre-processed with TreeTagger [21]. The comparable corpora statistics is given in Table 8.

Table 8. English-Latvian bilingual comparable automotive domain term-tagged corpus statistics

	English	Latvian
Documents	24 124	5 461
Unique sentences	1 114 609	247 846
Tokens in unique sentences	15 660 911	3 939 921
Total number of term phrases	2 851 803	1 792 344
Unique term phrases	432 059	162 312

Table 8 shows that a lot of phrases in both Latvian and English documents have been marked as terms. This is due to the configuration, which in our experiment was set to achieve a better F-measure and not precision.

Once terms were tagged in all documents, we executed *TEA* on the aligned document pairs with a threshold of 0.6. *TEA* produced in total 4 414 term pairs, which were then filtered preserving only the highest scored pair for each Latvian term, thus reducing the final pair count to 972. The results were then manually evaluated in terms of precision, as shown in Table 9.

Table 9. TEA term mapping results with a threshold of 0.6 on the comparable English-Latvian automotive domain corpus

TEA translation equivalence score	Correct mapping	Incorrect mapping	Precision
>=0.60	714	258	73.46
>=0.65	501	115	81.33
>=0.70	331	38	89.70
>=0.75	228	24	90.48
>=0.80	142	14	91.03
>=0.85	93	10	90.29
>=0.90	50	9	84.75
>=0.95	36	7	83.72
>=1.00	30	7	81.08

Error analysis of *TEA* results shows five distinct error types:

1. Term pairs are falsely aligned because too many characters overlap, which results in a high cognate matching score. For instance, "*auto mode*" in Latvian ("*auto fashion*" in English) and "*auto model*" in English get a score of 0.86. This type of error was present in 22.9% of all errors in the experiment.
2. Multi-word terms are misaligned because of different word order. Consider the following example: "*water pressure*" and "*pressure water*". These are two different terms. This type of error was evident in 2.3% of all misalignments.

3. Terms are aligned with longer terms containing additional tokens that change the semantic meaning of the term. For instance, "*modernie dīzeļi*" in Latvian ("*modern diesels*" in English) and "*modern diesel engine*" in English get a translation equivalence score of 0.8. This is the most frequent *TEA* error. 53.1% of all errors in our experiment were of this type.
4. Terms are wrongly aligned with terms in the same language (for instance, English-English instead of the required English-Latvian) because no language identification is performed in the term level. It is frequent (especially in Web crawled documents) that a part of a document or some specific terms are written in another language. In the case of identical terms, this results in a high cognate translation score (for instance, "*combustion process*" both in a Latvian document and English document get a cognate score of 1.0). This type of error was present in 11.6% of all misalignments.
5. Terms are misaligned because of many out-of-domain translations in the probabilistic dictionary. If the dictionary is built from bad quality parallel data or the dictionary features many translations of terms in other domains, false translation equivalents can be produced. For instance, a "*notebook*" may be an electronic device or a book for notes depending on the context. We found that 2.7% of errors in our experiment were of this type.

The remaining 7.4% of errors were caused by either a combination of the above mentioned error types or by other less frequent cases.

Despite the errors *TEA* achieved a precision of 73.46% with the translation equivalence threshold of 0.6, which can be increased up to 91% (as seen in Table 9) using an out-of-domain dictionary.

6 Conclusion

In this paper we presented methods for term extraction and bilingual mapping in comparable corpora, as well as term tagging in comparable documents based on developed term extraction techniques. Term tagging has been applied and evaluated for Latvian and Lithuanian, and bilingual term mapping has been applied and evaluated for Croatian, Latvian, Lithuanian, and Romanian.

The real world scenario, in which bilingual terms were acquired from comparable Web crawled corpus (in a domain unknown to the tools), shows that regardless of the relatively low precision of term tagging, bilingual term mapping in the presented process chain can achieve a precision up to 91%.

The defined process chain combines statistical and knowledge based approaches and can be fine-tuned for specific tasks where different quality measures (recall or precision) apply. The term extraction tool *CollTerm*, the term tagging tool *TWSC*, and the term mapping tool *TEA* presented in the paper are published under the Apache 2.0 license and are freely available as part of the ACCURAT project deliverable D2.6 [1].

7 Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

References

1. ACCURAT D2.6 (2011). Toolkit for multi-level alignment and information extraction from comparable corpora., 31st August, 2011 (http://www.accurat-project.eu/), 123 pages.
2. ACCURAT D3.5 (2011). Tools for building comparable corpus from the Web., 31st October, 2011 (http://www.accurat-project.eu/), 43 pages.
3. 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.
4. 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.
5. 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.
6. 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.
7. Delač, D., Krleža, Z., Dalbelo Bašić, B., Šnajder, J., Šarić, F. (2009) TermeX: A Tool for Collocation Extraction. In Lecture Notes in Computer Science, Computational Linguistics and Intelligent Text Processing: Proceedings of the 10th International Conference, CICLing 2009, March 1-7, 2009, Mexico City, Mexico, Springer, vol. 5449, p. 149-157.
8. 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, Vol. 3, Issue 2, pp. 115-130.
9. Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora**.**In Proceedings of the 5th Annual Workshop on Very Large Corpora, p. 192–202.
10. Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. Translating and the Computer 21, London, UK.
11. Georgantopoulos, B. and Piperidis, S. (2000b). A hybrid technique for automatic term extraction. In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications-ACIDCA'2000.
12. 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, Latvia, p. 82-89.
13. Justeson, J.S. and Katz, S.M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. In Natural Language Engineering 1:9-27.
14. Kageura, K. and Umino, B. (1996). Methods notion of automatic term recognition. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, No. 3, pp. 259-289.

15. Kruglevskis, V. (2010). Semi-Automatic Term Extraction from Latvian Texts and Related Language Technologies. Magyar Terminologia (Journal of Hungarian Terminology).
16. Kruglevskis, V. and Vancane, I. (2005). Term Extraction from Legal texts in Latvian. In Proceedings of the Second Baltic Conference on Human Language Technologies, April 4-5, 2005.
17. Mima, H. and Ananiadou, S. (2000). An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. In Kageura, K. and T. Koyama (eds.), Japanese Term Extraction: Special issue of Terminology 6:2, pp. 1750194.
18. Morin, E. and Prochasson, E. (2011). Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. ACL HLT 2011, page 27.
19. Pantel, P. and D. Lin (2001). A Statistical Corpus-Based Term Extractor. In: Stroulia, E. and Matwin,S. (Eds.) AI 2001, Lecture Notes in Artificial Intelligence, p. 36-46. Springer-Verlag.
20. Pinnis, M. and Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In: Proceedings of the Second Workshop on Systems and Frameworks for Computational Morphology, Communications in Computer and Information Science, Vol. 100, p. 14-22. Springer-Verlag.
21. Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, p. 44-49.
22. Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, volume 28, p. 11-21.
23. Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In Proceedings of BUCC 2012: 5th Workshop on Building and Using Comparable Corpora, May 26, 2012, Istanbul, Turkey.
24. Ştefănescu, D., Ion, R., Boroş, T. (2011). TiradeAI: An Ensemble of Spellcheckers. In Proceedings of the Spelling Alteration for Web Search Workshop, pp. 20-23, Bellevue, USA
25. Steinberger, R., Pouliquen, B., Hagman, J. (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc, Springer-Verlag.
26. Tadić, M. and Šojat, K. (2003) Finding Multiword Term Candidates in Croatian. In Proceedings of IESL2003 Workshop, RANLP2003 Conference, September 8-9, 2003, Borovets, Bulgaria, p. 102-107.
27. Weller, M., Gojun, A., Heid, U., Daille, B., Harastaniv, R. (2011). Simple methods for dealing with term variation and term alignment. In Proceedings of TIA 2011: the 9th International Conference on Terminology and Artificial Intelligence, November 8-10, 2011, Paris, France.
28. Zeller, I. (2005). Automatinis terminu atpazinimas ir apdorojimas. PhD thesis.