Terminology Extraction from Comparable Corpora for Latvian

Tatiana Gornostay, Anita Ramm, Ulrich Heid, Emmanuel Morin, Rima Harastani, Emmanuel Planas

Terminology Extraction from Comparable Corpora for Latvian

Tatiana GORNOSTAY[a,1], Anita RAMM[b], Ulrich HEID[b],
Emmanuel MORIN[c], Rima HARASTANI[c], Emmanuel PLANAS[c]
[a] *Tilde, Latvia*
[b] *IMS, University of Stuttgart, Germany*
[c] *LINA, University of Nantes, France*

Abstract. This paper presents the work on terminology extraction from comparable corpora for Latvian. In the first section we introduce our work; the second section briefly describes the concept of the project and the implemented general terminology processing chain; the following two sections focus on terminology extraction workflow for Latvian and evaluation of results, respectively.

Keywords. terminology extraction, comparable corpora, under-resourced language, Latvian language

1. Introduction

Terminology is multidisciplinary and comprises primarily such tasks as the analysis of concepts and conceptual systems; creation of new terms; identification, recognition, extraction of existing terms from texts; compilation of terminology resources, i.e., terminography, e.g., dictionaries, term banks and databases; application of terminology resources, e.g., in translation, including computer-assisted and machine translation; management of terms and, as a new trend in terminology – the consolidation of different, usually dispersed, multilingual terminology resources.

After a long history of monolingual terminology extraction which has been a research object for more than 20 years starting with pioneer experiments for the so-called big languages, such as English, German or French [1], this task still remains a less-researched and thus crucial for small languages, such as Latvian. Small languages are usually under-resourced and existing terminology extraction tools underperform for such languages. This can be explained by the lack of necessary language resources, on the one hand, and by poor performance of language independent methods for some of these languages, on the other hand, often depending on the typological nature of a given language. For the Latvian language, e.g., the first experiment on terminology extraction showed that a linguistic method based on morphosyntactic analysis is more appropriate than a statistical one that proved to be adequate for analytical languages [2, 3, 4].

[1] Corresponding Author: Tatiana GORNOSTAY, Terminology service manager, Tilde, Vienibas gatve 75a, Riga LV-1004, Latvia; E-mail: tatiana.gornostay@tilde.lv.

Classical bilingual terminology extraction methods have so far relied on the assumption that there is a collection of parallel texts available for processing. As a rule, parallel corpora are available for certain language pairs, usually including English, and are scarce for small languages. Recent work has focused on automatic terminology extraction in such languages from comparable corpora [2]. The project TTC (Terminology Extraction, Translation Tools and Comparable Corpora)[2] presented in this paper contributes to the leveraging of computer-assisted translation tools, machine translation systems, and multilingual content management tools by generating bilingual terminologies automatically from comparable corpora in seven languages belonging to five language families. In this paper we describe the terminology extraction workflow being developed within the project with special attention to the Latvian language.

2. Project overview

The project is developing tools for automatic bilingual terminology extraction from comparable corpora [5, 6] since parallel corpora are very scarce, especially when one of the languages under consideration belongs to the group of small languages, such as Latvian.[3]

The tool for bilingual terminology extraction being developed in the project is domain-independent and can be used for 7 languages: Chinese (ZH), English (EN), French (FR), German (DE), Latvian (LV), Spanish (ES), and Russian (RU), and their pairs, respectively. In order to handle all of the languages listed above in a homogeneous way, we have implemented knowledge-poor and language-independent procedures for monolingual and bilingual terminology extraction.

2.1. Terminology extraction – general processing chain

The multilingual terminology extraction workflow based on the processing of comparable corpora and implemented in TTC is shown in Figure 1.

The first step in the terminology extraction chain is the collection of domain-specific corpora in two languages. In TTC, we have developed a focussed web crawler *Babouk* [7] which collects documents from the web for all of the project languages. The texts are subsequently pre-processed by enriching them with word categories (part-of-speech (POS) tags) and lemmas. For this purpose, we use TreeTagger [8] for all of the languages, with the exception for Latvian. For Latvian we use the web service which provides the procedures for tagging and lemmatization of Latvian texts based on the proprietary POS tagger for Latvian developed by Tilde [9]. Its tagset is richer than the usual 50-70 tags applicable for languages with less rich morphology than Latvian.

Monolingual terminology extraction is based on language-specific POS patterns which have been manually collected by language experts. Term candidates identified in the corpus are filtered using the *domain specificity* defined in [10]. The general corpus data, mostly newspaper articles, needed for the term filtering has been collected within the project. Monolingual terminology extraction also identifies term variants. For each

[2] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248005 (www.ttc-project.eu).

[3] Most previous work on automatic terminology extraction relies on parallel domain-specific corpora (cf. e.g., [20, 21, 22]). In TTC, however, we use comparable corpora for this task.

language, we have collected a list of POS patterns identifying base terms along with POS patterns which denote their synonymous or morphologically-related variants. The monolingual term lists provide a lot of information about the extracted term candidates: POS pattern, term complexity, frequency, domain specificity, inflected forms, and variants. The results are provided in the TBX format.

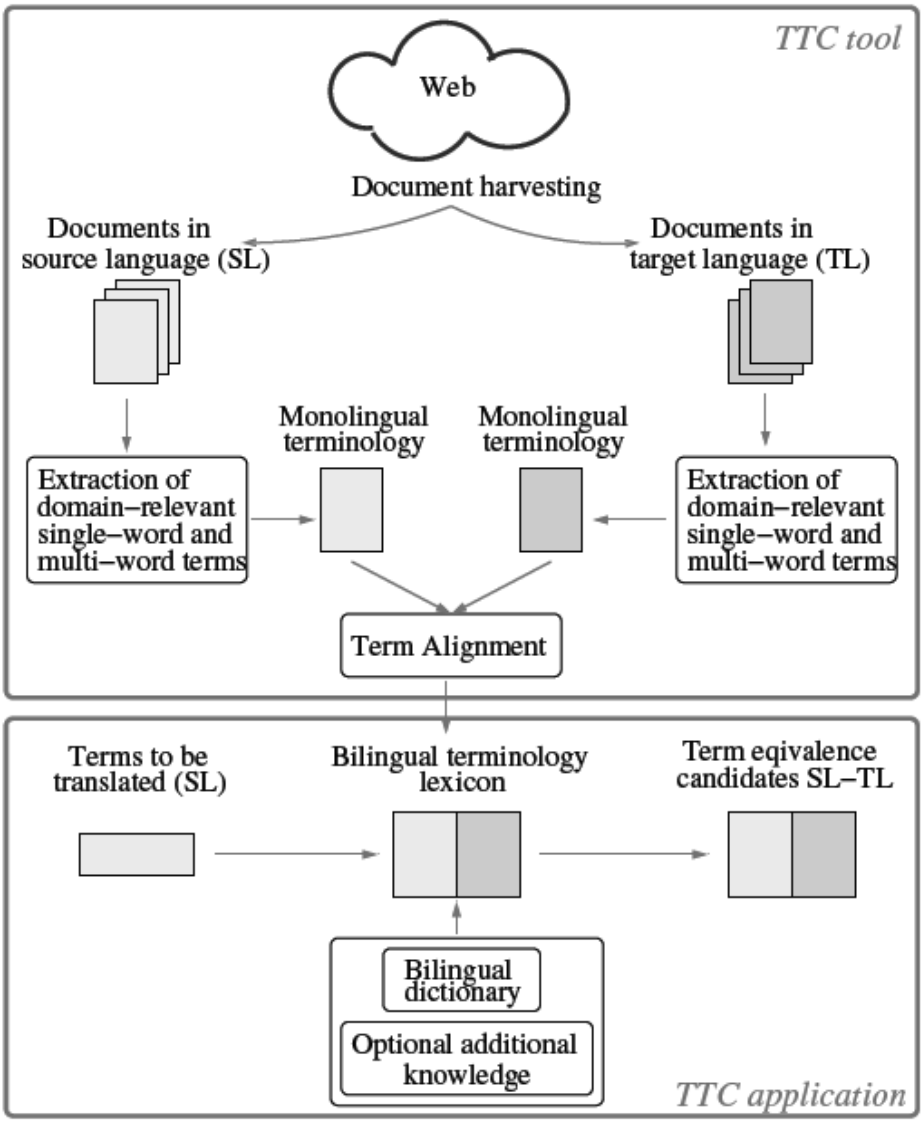

Figure 1. Multilingual terminology extraction workflow from comparable corpora.

Using the list of term candidates produced by the tool for a given (source) language (SL), the user may select a set of SL terms for which terminology alignment is to be computed. Terminology alignment is a combination of three approaches: a version of the standard vector-based approach for single-word terms [11], a compositional method for multi-word terms [12], and a special method for handling neoclassical compounds [13]. The output of terminology alignment is a list of translation candidates for each SL input term.

2.2. Use of the extracted terminology data

The output of bilingual terminology extraction can be fed into computer-assisted tools, such as SDL Trados (used within the evaluation usage scenario at Tilde), as well as into machine translation systems, such as Systran (a rule-based system) and Moses (a statistical system) (used within the evaluation usage scenarios at Sogitec and Tilde). The user can also import the extracted terminology data into My EuroTermBank[4], a web platform for storing, editing and sharing terminology resources based on EuroTermBank[5] and developed within the project.

3. Terminology extraction for Latvian

3.1. Crawled domain-specific corpora for Latvian

We have collected domain-specific corpora in the two domains (wind energy and mobile technology) for Latvian with a focused crawler *Babouk* [7]. The Latvian corpus collected within the project has the smallest size out of the 7 project languages: 220 823 running words from specialized texts in the domain of wind energy. For other languages, we managed to compile bigger corpora (e.g., EN: 313 954 words, DE: 358 602 words, and FR: 314 954). The task of obtaining more corpora is currently under consideration within the project work towards the domain adaptation of the existing English-Latvian SMT system.

3.2. Latvian terminology

We have analysed the Latvian terminology in both domains treated in the project (wind energy and mobile technology) and identified morphosyntactic term patterns of nominal groups and variation term patterns (graphical, morphological, paradigmatic, syntactic, and transpositional term variants) for single word and multi-word terms [14] (Examples 1-3).

lādēšana – uzlādēšana (charging): morphological addition (prefixation) *(1)*

ģeotermāls – ģeotermisks (geothermal): morphological substitution (suffixation) *(2)*

saules enerģija – saules un vēja enerģija (wind energy – wind and solar energy): syntactical addition (coordination of dependent element) *(3)*

[4] https://my.eurotermbank.com/

[5] http://www.eurotermbank.com

The nominal group is a dominant in the Latvian domain-specific texts. Non-prepositional genitive nominal groups are highly frequent in the Latvian terminology. In the domain of wind energy, e.g., 5 268 Latvian terms have been analysed and 2 703 of them represent a *Noun2:genitive Noun1* pattern. Much more rarely, nouns are used with prepositions and other cases (e.g., dative, accusative or locative). 1 413 multi-word terms in the analysed list are noun phrases made up of an adjective and a noun. Thus, a basic multi-word term pattern in the Latvian domain-specific texts is a two-element nominal group with the head noun modified by another noun in genitive case or an adjective (Examples 4-5).

gondola/N:fsg dzin'ejs/N:msn (nacelle engine) *(4)*

geostrofisks/Adj:msn vējš/N:msn (geostrophic wind) *(5)*

A three-element nominal group is observed in 667 cases with the following distribution: *Noun3:genitive Noun2:genitive Noun1* in 335 cases, *Adj Noun2:genitive Noun1* in 247 cases, *Noun2:genitive Adj Noun1* in 56 cases, *Adj2 Adj1 Noun* in 29 cases.

3.3. Reference term lists for Latvian

For evaluation purposes (to evaluate the output of the terminology extraction tools developed in the project) we manually compiled monolingual and bilingual reference term lists (RTLs) in both domains [15]. The process of the reference term list compilation for Latvian was fourfold:

(1) initially a linguist extracted term candidates manually;

(2) then a terminologist validated the list;

(3) then the list was checked against another list of automatically extracted term candidates to ensure the frequency of the manually extracted term candidates in the corpus;

(4) finally, a domain specialist was consulted on the termhood and/or unithood of term candidates [16, 17].

The quality of the extracted term candidate lists is being evaluated against the reference term lists.

4. Evaluation of bilingual terminology extraction for Latvian

4.1. Alignment of Latvian single word terms using vector-based approach

We have evaluated bilingual terminology extraction for EN→LV and LV→RU by checking the alignment of Latvian single word terms (SWTs). We ran two sets of alignment experiments. The first one takes Latvian SWTs from the corresponding RTLs as input, while in the second one we align larger sets of Latvian words (which need not to be domain-specific) derived from a big bilingual dictionary. The alignments were computed using the comparable wind energy corpora for the two language pairs crawled with the project's tools, cf. section 2.

The evaluation results (accuracy for top *n* alignment candidates) are given in Table 1. The results indicate that the vector-based approach performs better for LV→RU than for EN→LV.

Table 1. Vector-based alignment of LV SWTs with EN and RU SWTs.
The number of LV terms to be aligned and evaluated is given in the second row.

	EN - LV		**LV - RU**	
	SWTs from RTL (22)	SWTs from dictionary (254)	SWTs from RTL (32)	SWTs from dictionary (179)
top 5	4,54	1,18	18,75	2,79
top 10	18,18	5,90	37,50	10,61
top 20	18,18	9,05	43,75	16,76
top 50	18,18	17,32	53,12	24,02
top 100	22,72	25,98	62,50	32,96

4.2. Neoclassical alignment of Latvian multi-word terms

Neoclassical compounds are terms that contain at least one neoclassical element (Latin or Greek). Implementing the neoclassical approach in TTC [18], we translate each neoclassical element within a word (a neoclassical compound) separately (e.g., aero into aero and dynamic into dinamisks when translating EN: aerodynamic into LV: aerodinamisks). For this, we researched neoclassical elements in the both languages and translated them from English into Latvian. The evaluation results are as follows:

- for EN-LV translation direction: precision top 5 are 85% and 100% for the two domains of wind energy and mobile technology;
- for LV-RU translation direction: precision top 5 are 83% and 80% for the two domains of wind energy and mobile technology.

5. Conclusion

The project is at the beginning of its third year now and so far it has made significant progress towards the main scientific and technological objectives for the first two years [19].

We have compared the evaluation results for Latvian with the results obtained for other language pairs treated in the project, such as EN→ES and EN→FR. The comparisons showed that there is no big difference in the alignment accuracy. This leads us to the conclusion that the alignment performance of our tools is almost equal (with some small deviations) for both small and big language pairs and across language pairs from different typological language families.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248005.

References

[1] Cabré Castellví, M. T., R. Estopa Baot, J. Vivaldi Palatresi. Automatic Term Detection: A Review of Current Systems. In *D. Bourigault, C. Jacquemin, M-C. L'Homme. Recent Advances in Computational Terminology*, pp. 53-88.

[2] Pinnis, M. Ljubešić N., Ştefănescu D., Skadiņa I., Tadić M., Gornostay T., Term Extraction, Tagging and Mapping Tools for Under-Resourced Languages. In *Proceedings of the Terminology and Knowledge Engineering Conference*, 2012.

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

[4] Kruglevskis, V., I. Vancane, Term Extraction from Legal texts in Latvian. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, 2005.

[5] Fung, P. A., Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the Association for Computational Linguistics*, 1995, pp. 236-243.

[6] Rapp, R., Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 320-322.

[7] Groc, C. de, Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011.

[8] Schmid, H., Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland, 1995.

[9] Pinnis, M. and Goba K., Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In *Proceedings of the 2nd Workshop on Systems and Frameworks for Computational Morphology (SFCM2011), Zurich, 26 August 2011, Springer, Heidelberg, Communications in Computer and Information Science*, 2011(1), Vol. 100, pp. 14-22.

[10] Ahmad, K., What is a term? The semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Poechhacker, and K. Kaindl, editors, *Translation studies: an interdiscipline*, pp. 267-278, 1st edition, 1992.

[11] Rapp, R., Automatic Identification of Word Translation from Unrelated English and German Corpora. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL '99). College Park, Maryland, USA*, 1999, pp. 519-526.

[12] Morin, E. and Daille, B., Compositionality and lexical alignment of multi-word terms. In *Language Resources and Evaluation*, 2009, Vol. 44, pp. 79-95.

[13] Harastani, R., Daille, B., Morin, E., Neoclassical Compound Alignments from Comparable Corpora. In *Proceedings of CICLing*, 2012(2), pp. 72-82.

[14] Daille, B., Variants and application-oriented terminology engineering, *Terminology, Vol. 1*, 2005, pp. 181-197.

[15] Loginova, Reference Lists for the Evaluation of Term Extraction Tools. In *Proceedings of the Terminology and Knowledge Engineering Conference*, 2012.

[16] Frantzi, K., S. Ananiadou, and H. Mima, Automatic Recognition of Multi-Word Terms: the *C-value/NC-value* Method. In *International Journal on Digital Libraries, Vol. 3, Issue 2*, pp. 115-130.

[17] Kageura, K. and B. Umino, Methods notion of automatic term recognition. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3, 1996, pp. 259-289.

[18] Harastani, R., B. Daille, E. Morin, Neoclassical Compound Alignments from Comparable Corpora. In Proceedings of CICLing 2012(2), pp. 72-82.

[19] Gornostay, T., Gojun A., Weller M., Heid U., Morin E., Daille B., Blancafort H., Sharoff S., Mechoulam C., Terminology Extraction, Translation Tools and Comparable Corpora: TTC Concept, midterm progress and achieved results. In *Proceedings of CREDISLAS 2012: Workshop on Creating*

Cross-language Resources for Disconnected Languages and Styles co-located with LREC 2012, Istanbul, Turkey.

[20] Koehn, P., Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand, 2005.

[21] Singh, A. K. Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 2008, pp. 7–12.

[22] Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, pp. 2142-2147.