# Service model for semi-automatic generation of multilingual terminology resources

Andrejs Vasiljevs, Marcis Pinnis, Tatiana Gornostay

## To cite this version:

Andrejs Vasiljevs, Marcis Pinnis, Tatiana Gornostay. Service model for semi-automatic generation of multilingual terminology resources. Terminology and Knowledge Engineering 2014, Jun 2014, Berlin, Germany. 10 p, 2014. <hal-01005874>

HAL Id: hal-01005874

https://hal.archives-ouvertes.fr/hal-01005874

Submitted on 13 Jun 2014

# Service model for semi-automatic generation of multilingual terminology resources

Andrejs Vasiļjevs, Mārcis Pinnis, Tatiana Gornostay

Tilde, Vienibas gatve 75a, Riga, Latvia

{andrejs|marcis.pinnis|tatiana.gornostay}@tilde.lv

**Abstract.** The authors present a service-based model for semi-automatic generation of multilingual terminology resources which, if performed manually, is very time consuming. In this model, the automation of individual terminology work tasks is rendered as a set of interoperable cloud-based services integrated into workflows. These services automate the identification of term candidates in user documents, and the lookup of translation equivalents in online terminology resources and on the Web by automatically extracting multilingual terminology from comparable and parallel online resources. Collaborative involvement of users contributes to the refinement and enrichment of the raw terminological data. Finally, we present the TaaS platform, which implements this service-based model, particularly focusing on the processing of Web content.

**Keywords:** terminology, service, language resource, multilingual, cloud computing, automation

## 1 Introduction

Terminology resources are among the most used language data providing lexical designations assigned to concepts, term equivalents in different languages, their definitions, usage contexts, and other data. Digital collections of terminology data are in everyday use by language workers – terminologists, translators, technical writers, and are also increasingly used in automated tools for machine translation, information extraction, semantic search, and other applications. In this paper we present a service-based approach to automate the creation and use of multilingual terminology resources, which is a very time-consuming process if performed manually.

The creation and maintenance of terminology resources is usually organised in one of two settings. The objective of descriptive terminology work is to document terms used to designate concepts of a given discipline. It usually involves manual or semi-automated analysis of documents to identify candidate terms, which are then checked in existing terminology resources for corresponding entries to create a terminology glossary. This is a typical practise in preparing terminology for translation and writing (Wright and Budin, 2001). Prescriptive terminology work is practiced by standardisers, government regulators, and nomenclature specialists to ensure consistent and unambiguous use of terminology in regulated areas as well as to facilitate precise communication in general usage. Standardisers may also perform descriptive work as they collect data on usage prior to agreeing on standardised terms.

The traditional model for the creation of a bi-(multi-)lingual terminology resource in the descriptive scenario involves (1) collection of domain specific documents, (2) term identification and preparation of a glossary, (3) lookup for matching terminology data in prescriptive sources, (4) creation of new terminology data for glossary entries that have not been found in other sources. This work is usually carried out by an individual expert or a group, involves a great deal of manual work, and is rarely shared to the community.

In the last decade, significant progress has been achieved in the automation of these terminology work steps. Web crawling tools have been successfully adapted and applied to collect corpora for terminology needs. Baroni and Bernardini (2004) use a list of seed terms and bootstrapping approach in Web crawling, Blancafort et al. (2010) and Pinnis et al. (2012a) collect comparable corpora for term extraction. Methods for term extraction range from language independent n-gram extraction using relative frequency based termhood estimation (Delač et al., 2009; Pantel and Lin, 2001) to linguistically motivated methods based on syntactic analysis and the application of term phrase patterns (Bourigault, 1992). The combination of statistical and linguistically motivated techniques (Justeson and Katz, 1995; Dagan and Church, 1994) is the most used approach in the practical tools. The automation of term lookup is hindered by fragmentation of terminology resources in numerous databases with differing data structure and coverage. In recent years, several efforts have been invested to consolidate and harmonise terminology resources in national and international online term banks (e.g., *Rikstermbanken*, *IATE* and related lookup tool *Quest*, *EuroTermBank* and its term lookup API). But these term banks mostly incorporate terminology resulting from prescriptive work and are still limited in coverage.

Due to laborious manual work and the incompleteness of terminology data in prescriptive sources, it is still very time consuming to find and prepare the terminology data needed in practical translation work. Several surveys show that technical translators spend more than 30% of translation time on terminology work (Massion, 2007; Gornostay, 2010). Creators of terminology resources and operators of term banks struggle to cope with the need to incorporate an increasing number of new terms resulting from the rapid developments in technological, scientific and social spheres.

Although several tools are available to automate individual steps of terminology work, there is no solution that covers all major tasks for terminology creation. As a result, the major deficiencies of terminology resources are high cost and time needed to create them, insufficient coverage of terminology, particularly for the most recent concepts, poor language coverage, insufficient sharing of terminology resources, and a lack of collaborative mechanisms for involving terminology practitioners.

## 2 Overview of the Service Model

To facilitate the creation and usage of terminology resources and to benefit from the recent advances in computational linguistics, we propose a cloud-based service model that automates the major steps in terminology work. The automation of individual tasks in terminology work is rendered as a set of interoperable cloud-based services integrated into workflows. These services automate identification of term candidates in user-provided monolingual documents and the lookup of translation equivalents for extracted monolingual term candidates. Translation equivalents are retrieved from online terminology term banks, automatically extracted multilingual terminology from comparable and parallel resources on the Web (in online and cached scenarios), as well as from terminology collections created by the platform's users.

An essential element of this model is the collaborative involvement of users in the refinement and enrichment of raw terminological data, automated sharing and synchronisation of the terminology in various use cases by language workers and language processing applications (e.g., computer-assisted translation tools, machine translation systems, terminology management and terminology lookup platforms etc.). The model is based on a reciprocity principle. Users process their documents and refine and enrich resulting terminological data, which can be shared and provided to other users and contributed to term banks.

This model is being successfully piloted in the TaaS platform (Pinnis et al., 2013) serving all 24 official EU languages and providing the following services:

- Automatic extraction of monolingual term candidates from user-uploaded source documents using terminology extraction techniques;
- Automatic retrieval of target translation equivalents for the extracted monolingual term candidates from various existing public and industry terminology resources;
- Acquisition of translation candidates from parallel or comparable Web data for terms not found in existing terminology resources using terminology extraction and bilingual terminology alignment methods;
- Facilities for refinement and enrichment of the resulting automatically extracted terminological data by the platform's users;
- Data sharing and integration via API and export tools for sharing terminological data with major existing terminology resources and reuse it in various applications;
- Instant access to term translation equivalents and translation candidates for professional translators via CAT tools;
- Domain adaptation of statistical machine translation systems by integration with provided terminological data.

## 3 Service for Term Candidate Identification

Terms in user-provided monolingual documents are identified using the term tagging system TWSC (Pinnis et al., 2012b), which identifies term candidates in three steps:

1. At first, documents are pre-processed using part-of-speech or morpho-syntactic taggers (and optionally also lemmatisers if such exist for a language).
2. Then term candidates are extracted using linguistic filtering and statistical ranking methods. The filtering is performed with morpho-syntactic term phrase regular expressions and the ranking is performed with co-occurrence measures (e.g., log likelihood, modified mutual information etc.) for terms of two or more tokens and the TF*IDF (Spärck Jones, 1972) measure for unigram terms.

3. Finally, identified and extracted term candidates are marked in the user-provided documents using n-gram prioritisation and the term rankings.

The system has been extended to the languages supported by the platform by integrating existing part-of-speech taggers (e.g., the OpenNLP [1] models for Dutch, English, French, German, Italian, and Spanish, the system by Pinnis and Goba (2011) for Estonian, Latvian, and Lithuanian, and HunPOS[2] for Hungarian and Portuguese), building projected part-of-speech taggers for under-resourced languages using parallel corpora (Aker et al., 2014), and generating term phrase patterns from parallel corpora following a similar approach to the part-of-speech tagger projection.

We have evaluated the quality of the system for four languages in two subject fields (information technology and mechanical engineering). Two annotators (language specialists with a focus on terminology) identified terms in two documents. The documents across all languages were on similar topics and of similar difficulty levels. Each of the annotators has a subjective view on what comprises a term in a given context and what does not. This is because termhood and unithood of terms can be very ambiguous as well as subjective to the opinions of specialists who work with the terminology. Therefore, in our evaluation we use a union of their annotations and performed a precision analysis of the documents tagged by the system (see Table 1).

**Table 1.** Evaluation results

| **Language** | **Information Technology** | | | **Mechanical Engineering** | | |
|---|---|---|---|---|---|---|
| | **Correct** | **Total** | **Precision** | **Correct** | **Total** | **Precision** |
| English | 213 | 365 | 58.36% | 254 | 503 | 50.50% |
| German | 198 | 338 | 58.58% | 132 | 380 | 34.74% |
| Hungarian | 147 | 605 | 24.30% | 199 | 603 | 33.00% |
| Latvian | 316 | 540 | 58.52% | 331 | 662 | 50.00% |

The results show that on average around 50% of the identified terms are true positives. Although seemingly average, the results are acceptable considering that termhood and unithood simultaneous identification is very challenging. This difficulty is supported also by comparing the annotator outputs. The average agreement rate of the two Latvian annotators was only at 63.3%. Also the remaining term candidates are not necessarily wrong. Because of the linguistically motivated term phrase filtering, the system produces syntactically justified term candidates, which can still be useful in some application scenarios, e.g., machine translation (Pinnis et al., 2012c).

For users who work on morphologically rich languages, term identification may produce very redundant term candidate lists. This can be due to the inflective nature of many languages. For example, in Czech, Latvian, Estonian, etc., nouns, verbs, adjectives (and other parts of speech) may have numerous different inflected surface forms. Terms are also affected by this inflective nature and, therefore, the platform addresses this issue with term normalisation. Term normalisation is a process of trans-

[1] http://opennlp.sourceforge.net/models-1.5/

[2] http://code.google.com/p/hunpos/

forming terms from their surface forms into their corresponding canonical forms as they are found in dictionaries and term banks. We use rule-based methods for term normalisation that for each of the term phrase regular expressions define a rule for term normalisation. For single-word terms the normalised forms often correspond to the term lemmas, however, for multi-word terms the normalised forms in many cases differ from the corresponding token lemma sequences. For example, the Latvian term "datoru tīklu" (transl. "computer network") is normalised as "datoru tīkls", however, the lemma sequence is different – "dators" "tīkls". Using a rule-based approach we can remove redundancy in the monolingual term lists.

# 4 Service for Translation Equivalent Retrieval from the Web

One of the main sources for translations is the Web. It contains a vast amount of multilingual information that can be used to acquire up-to-date knowledge. Using novel workflows for the collection of Web corpora in automatic and on-demand scenarios, extraction of multilingual terminology from the corpora, and integration of the acquired terminology into the platform, we can provide users with up-to-date term candidate translation equivalents. All the multilingual terminology that is acquired from the Web is stored in a Statistical Data Base (SDB), which is accessible for translation candidate lookup when users create their bilingual terminology collections. There are four distinct workflows for bilingual terminology extraction from the Web:

1. On-demand bilingual terminology extraction from parallel data. For terms, for which existing resources do not yield any translation equivalents, users can manually trigger a bilingual terminology collection task that collects parallel corpora from the Web by identifying Web sites simultaneously containing the unknown terms as well as parallel content, and extracts bilingual terminology from the parallel corpora using bilingual phrase alignment techniques.
2. On-demand bilingual terminology extraction from focused comparable corpora. As parallel corpora may be scarce and not in all subject fields identifiable on the Web, an alternative path is to search for comparable corpora, which is much wider available (e.g., news, encyclopaedias, press releases etc.).
3. Automated bilingual terminology extraction from comparable RSS news corpora.
4. Automated bilingual terminology extraction from Wikipedia.

The three comparable corpora processing workflows are depicted in Fig. 1 and we further describe the comparable corpora processing workflows in more details.

## 4.1 Corpora Collection

Comparable corpora are collected using the following three different methods:

- On-demand comparable corpora in a specific subject field are collected using the Focussed Monolingual Crawler – FMC (Mastropavlos and Papavassiliou, 2011).

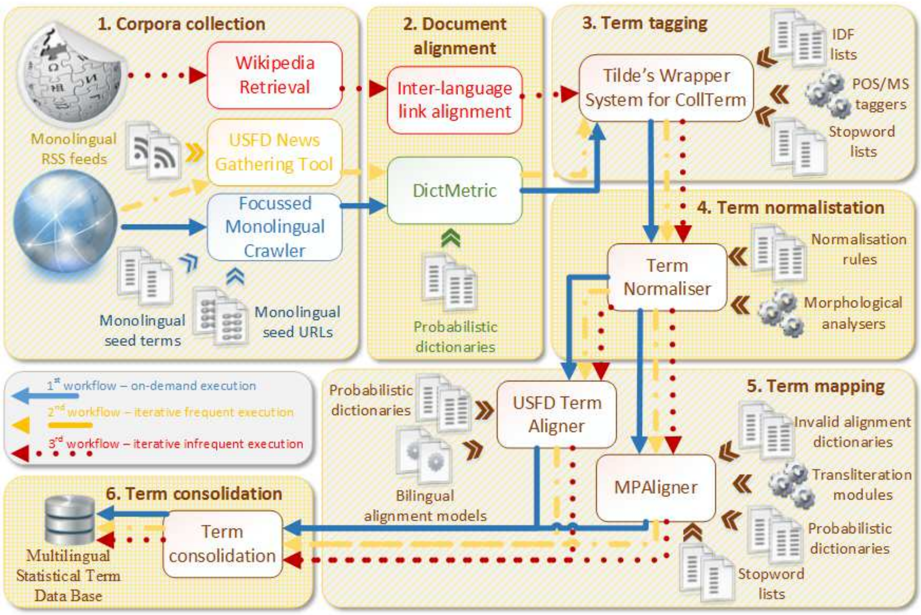

**Fig. 1.** Workflows for multilingual terminology extraction from comparable Web corpora

The corpus collection can be triggered by the user after the completion of a bilingual terminology extraction task and it can be performed using two scenarios:

- If the user requires translation equivalents for unknown terms only, a monolingual target language corpus is collected with FMC using seed URLs provided by the user. The bilingual terminology alignment tools then identify translation candidates of the unknown source terms in the collected monolingual corpus.
- If the user requires bilingual terminology of a specific subject field, two monolingual corpora are collected using seed URLs and optional seed terms for both languages. Then the corpora are cross-lingually aligned at the document level to acquire document-aligned comparable corpora. We use the DictMetric tool (Su and Babych, 2012) to estimate the comparability between two lists of monolingual documents. All documents are translated into English and a vector-based similarity metric is applied to calculate the document pair comparability.

- Iterative (automatic) comparable corpora are collected from RSS feeds using the tool proposed by Aker et al. (2012). The motivation behind this approach is that focused subject field oriented data is produced on a daily basis in the news articles from different sources (e.g., medical news, IT news etc.). Such news are often published in a full or adapted form in multiple languages, which makes them a valuable source of comparable corpora. The platform utilises such news sources in an automatic scenario in order to acquire up-to-date terminology once per week. In this scenario, DictMetric is also used to acquire bilingual comparable corpora.
- Comparable corpora are acquired from Wikipedia using the Wikipedia Retrieval tool (Paramita et al., 2012). Due to Wikipedia's multilingual nature where articles

are linked between different languages using inter-language links, Wikipedia offers access to the largest comparable corpus found on the Web. The platform iteratively downloads Wikipedia dump files and extracts comparable corpora with the Wikipedia Retrieval tool for all the languages supported by the platform.

### 4.2 Bilingual Term Extraction Workflows

After corpora collection, we use the term tagging system in order to identify term candidates. Then, for languages with term normalisation support, terms are normalised. Finally, the bilingual terminology is extracted using cross-lingual term mapping:

- Context independent term mapping proposed by (Pinnis, 2013) using maximised character alignment maps. The mapper maps terms in two steps: 1) at first, possible translation and transliteration equivalents of monolingual terms (in the respective other language) are identified using probabilistic dictionaries, and rule-based and statistical transliteration techniques; 2) then the translation and transliteration equivalents are analysed in order to identify the maximum character level content overlap between source and target terms. The method has been evaluated automatically for 22 language pairs and has shown to achieve a precision from 72.3% up to 91.1% (with an average of 84.8%) with recalls ranging from 33.7% to 71.5% depending on the source language when paired with English. Manual evaluation of the mapper on an English-Latvian term-tagged comparable corpus collected with FMC in the field of medicine has shown to achieve a precision of up to 91.3%.
- Supervised term mapping using the method proposed by Aker et al. (2013). Similarly to the first method, this method operates in two steps, however both methods differ significantly. The supervised method requires language-specific models, which are trained on term pairs and is, therefore, limited to language pairs for which such models exist. At first, the mapper analyses whole term pairs and tries to identify dictionary based and cognate-based features. Then for all term pairs of two documents, a binary classifier is used to estimate whether two term pairs can be considered valid translation equivalents or not. Although the authors report automatic evaluation results of up to 100% (the automatic evaluation scenarios between the mappers are not comparable), manual evaluation shows that the precision for true translation equivalents for English-German ranges from 63% up to 82.

Both mappers produce output data where each term is described by its surface form, sequence of lemmas, sequence of part-of-speech (or morpho-syntactic) tags, normalised form (if normalisation is available), sequence of the normalised form's part-of-speech (or morpho-syntactic) tags, and a concordance (up to 5 words around the term) that is extracted from the input document. An example is given in Fig. 2.

```
Lv → acs audus → en → eye tissue → 2200 → http://taas.eurotermbank.com/
FirstMapper/FMC-Medicine-Corpus/v12-05-2013 → 0.836885 → N-fsg---------
n-----------l- N-mpa---------n-----------l- → acs audi → acs audi → N-
fsg---------n-----------l- N-mpn---------n-----------l- → var ietekmēt
jebkurus acs audus, eksistē ļoti → NN NN → eye tissue → eye tissue → NN
NN→ to the underlying eye tissue. Symptoms of→ lv_2739.txt→ en_18915.txt
```

**Fig. 2.** Example of a mapped term pair ("→" denotes a tabulation character)

### 4.3 Delivery of Raw Terminological Data to Users

After cross-lingual term mapping, bilingual term pairs are integrated into the SDB by simultaneously performing term pair morphological consolidation. Depending on the linguistic tool support for languages, term consolidation is performed in three levels:

- For languages, for which lemmatisation of words is not integrated in the term tagging system, terms are grouped together only by their surface forms and part of speech sequences. This level ensures that SDB can support term translation candidate lookup even if linguistic support for certain languages is scarce.
- For languages with lemmatisation support, terms are grouped by their lemmatised forms and part of speech sequences. This consolidation level ensures that for morphologically rich languages redundancy, which is caused by having numerous surface forms of a single word, can be eliminated. However, this method can also group together surface forms belonging to different terms. For example, the term candidates “personālais dators” and “personāls dators” from Fig. 2 both have identical lemma sequences. This issue is solved by the third level.
- For languages with term normalisation support, different surface form terms are grouped by their normalised forms and the normalised form part of speech sequences. This level provides the highest level of morphological consolidation.

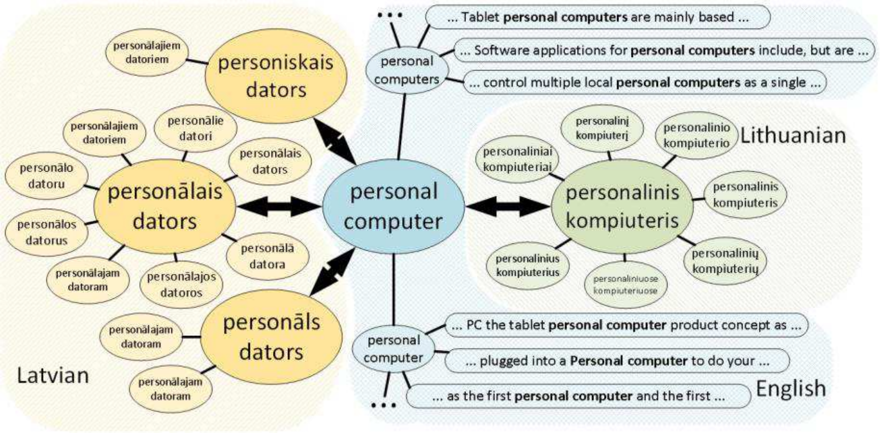

**Fig. 3.** Visualised example of terminological data from the Statistical Data Base

The consolidation levels are used in order to provide the most appropriate term translation equivalents for a term lookup query. If no translation equivalents are identified in the higher consolidation levels, the data from the lower levels is used.

## 5 Service for Collaborative Refinement of Raw Terminology

The platform provides facilities for collaborative refinement of raw term pairs that are noisy and need validation. Term validation can be regarded as a three-step procedure: 1) monolingual validation (deletion of “unwanted” or unreliable term candidates, definition of termhood, term variant identification, deduplication, etc.), bilingual validation (checking whether translation candidates are reciprocal translations, defining the right translations, etc.), and 3) validation in context.

The platform also provides a service for sharing terminology among users. Sharing that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology is highly rated by users. A recent survey (Gornostay et al., 2013) showed that up to 60% of users would share their data with the community.

## 6 Service for Terminology Sharing and Application

Approved and enriched terms can be exported and then used in other working environments with the help of the TBX[3], and comma or tab-separated value formats. Approved terms can also be used in other terminology projects by the user(s) who owns the data as well as other users, provided that the term collection is shared.

Terminology resources are important not only for language workers but also for various language processing applications ("machine users") such as computer-assisted translation tools and machine translation (MT) systems. The platform provides an API for external systems to access the terminology services and terminology data. This API-level integration is currently implemented by the memoQ CAT tool and the LetsMT statistical MT system (Vasiļjevs et al., 2013). The objective of this work is to create project specific terminology resources for dynamic adaptation of MT systems.

## 7 Conclusion

We have presented a service-based model and its implementation in a novel platform for translators, terminologists, and language workers that streamlines the work on multilingual terminology generation, management, sharing, and use in various working environments. We described the services and workflows provided by the platform and presented evaluation results for the separate platform components. Although term identification is very challenging even to human annotators, we can achieve comparable precision with automatic methods using the extended term tagging system. For example, for Latvian an average precision of 53.8% was achieved in comparison to an average annotator agreement rate of 63.3%. The work within the TaaS project has received funding from the European Union under grant agreement n° 296312.

## 8 References

1. Aker, A., Kanoulas, E., and Gaizauskas, R. (2012). A light way to collect comparable corpora from the Web. In Proceedings of LREC 2012 (pp. 15–20). Istanbul, Turkey.
2. Aker, A., Paramita, M., and Gaizauskas, R. (2014). Bootstrapping Term Extractors for Multiple Languages. In Proceedings of LREC 2014.
3. Aker, A., Paramita, M., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In Proceedings of ACL 2013. Sofia, Bulgaria.

[3] Term Base eXchange standard (ISO 30042:2008) with extensions defined in https://demo.taas-project.eu/tbx/taas.xcs.

4. Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of LREC 2004.
5. Blancafort, H., Daille, B., Gornostay, T., Heid, U., Méchoulam, C., and Sharoff, S. (2010). TTC: Terminology extraction, translation tools and comparable corpora. In Proceedings of the 14th EURALEX International Congress (pp. 263-268).
6. Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In Proceedings of COLING 1992, Volume 3 (pp. 977-981).
7. Dagan, I., and Church, K. (1994). Termight: Identifying and translating technical terminology. In Proceedings of ANLP 1994 (pp. 34-40).
8. Delač, D., Krleža, Z., Šnajder, J., Bašić, B. D., and Šarić, F. (2009). TermeX: A Tool for Col-location Extraction. In Proceedings of CICLing 2009 (pp. 149-157).
9. Gornostay, T. (2010). Terminology management in real use. In Proceedings of the 5th Inter-national Conference Applied Linguistics in Science and Education (pp. 25-26).
10. Gornostay, T., Vopodiyanova, O., Vasiļjevs, A., and Schmitz, K.-D. (2013). Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. In Proceedings of TRALOGY II.
11. L'Homme M.-C. (2004) La terminologie: principes et techniques. Montréal: Les Presses de l'Université de Montréal.
12. Justeson, J. S., and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering, 1(1), 9-27.
13. Mastropavlos, N., and Papavassiliou, V. (2011). Automatic acquisition of bilingual language resources. In Proceedings of the 10th International Conference of Greek Linguistics.
14. Muegge U. (2012) The Silent Revolution: Cloud-Based Translation Management System. In TCWorld Journal, July, 2012.
15. Pantel, P., and Lin, D. (2001). A statistical corpus-based term extractor. In Advances in Artificial Intelligence (pp. 36-46). Springer Berlin Heidelberg.
16. Paramita, M. L., Clough, P., Aker, A., and Gaizauskas, R. J. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In Proceedings of LREC 2012 (pp. 790–797), Istanbul, Turkey.
17. Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of RANLP 2013 (pp. 562–570). Hissar, Bulgaria.
18. Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., and Babych, B. (2012a). ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In Proceedings of ACL 2012 System Demonstrations (pp. 91–96).
19. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012b). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In Proceedings of TKE 2012 (pp. 193–208). Madrid.
20. Pinnis, M., and Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In Proceedings of Baltic HLT 2012 (Vol. 247, pp. 176–184).
21. Pinnis, M., Gornostay, T., Skadiņš, R., and Vasiļjevs, A. (2013). Online Platform for Extracting, Managing, and Reusing Multilingual Terminology. In Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013. Tallinn, Estonia.
22. Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, Volume 28, pp. 11-21.
23. Vasiljevs, A., Gornostay, T., and Skadins, R. (2010). LetsMT!--Online Platform for Sharing Training Data and Building User Tailored Machine Translation. In Proceedings of Baltic HLT 2010 (pp. 133–140).
24. Wright, S. E., and Budin, G. (Eds.). (2001). Handbook of Terminology Management: Basic Aspects of Terminology Management. Volume 1. John Benjamins Publishing.