# Collecting Language Resources for the Latvian e-Government Machine Translation Platform

**Roberts Rozis, Andrejs Vasiļjevs, Raivis Skadiņš**
Tilde
Vienibas gatve 75A, Riga, Latvia
E-mail: roberts.rozis@tilde.lv, andrejs@tilde.com, raivis.skadins@tilde.lv

**Abstract**

This paper describes corpora collection activity for building large machine translation systems for Latvian e-Government platform. We describe requirements for corpora, selection and assessment of data sources, collection of the public corpora and creation of new corpora from miscellaneous sources. Methodology, tools and assessment methods are also presented along with the results achieved, challenges faced and conclusions made. Several approaches to address the data scarceness are discussed. We summarize the volume of obtained corpora and provide quality metrics of MT systems trained on this data. Resulting MT systems for English-Latvian, Latvian-English and Latvian-Russian are integrated in the Latvian e-service portal and are freely available on website HUGO.LV. This paper can serve as a guidance for similar activities initiated in other countries, particularly in the context of European Language Resource Coordination action.

**Keywords:** corpus, parallel texts, machine translation, web crawling, e-Government, public sector information

## 1. Project Background and Goal

This paper describes the work on language resource collection for the Latvian e-Government Machine Translation (MT) Platform providing multilingual access to the portal of public online services.

To foster broader use of public online services, Latvian Government has created a centralized portal Latvija.lv[1]. The aim of the portal is to ensure quick and convenient access to the services provided by the Latvian State institutions and municipalities. The portal provides guidance on requirements (forms, documents, payments, terms etc.) and administrative procedures in order to receive public and municipal services, as well as direct access to those services that are offered online. Currently this portal hosts 108 e-services.

Considering the nationalities in Latvia, it is important to make this content accessible not only in Latvian (the official state language), but also in Russian (1/3 of population are Russians, Byelorussians, Ukrainians) and English (for most foreigners). Considering dynamically changing content of online services and high costs of human translation, making use of Machine Translation (MT) was adopted as the only viable solution to provide multilingual access.

Latvian Culture Information Systems Centre[2] launched a project to build Machine Translation Platform for e-Government with specific MT systems tuned for state administration domain. The resulting platform is now branded as HUGO.LV. The goal in the long run is to provide MT services not only to portal Latvija.lv but to all Latvian State institutions and keep developing and integrating machine translation in all related public e-services. Language technology company Tilde[3] was chosen as a technology partner in this project, and was commissioned to perform language resource collection and processing, MT systems building and technology delivery.

The goal in the project was to create a large corpora of language resources and build the best possible MT systems for 3 language pairs – English-Latvian, Latvian-English and Latvian-Russian. Two systems had to be provided for every translation pair – generic MT system and system tuned to the specific of the public administration.

## 2. Corpora Composition

### 2.1 Requirements

Latvian e-Government MT Platform (Vasiļjevs et al., 2014) is built by Tilde using LetsMT technologies (Vasiļjevs et al., 2011; 2012) which are based on the Moses toolkit (Koehn et al., 2007). LetsMT includes facilities to process parallel and monolingual corpora and build translation and language models for phrase-based statistical machine translation. From the perspective of language resources, it requires collecting and processing general and domain specific parallel and monolingual data to create MT systems which are customized for particular application area. To ensure optimal quality of resulting MT systems, the project requirements set 5 million sentences as the minimal amount of parallel data in general domain and 2 million sentences as the minimum for the public administration domain. These are significant amounts taking into account that Latvian language is weakly supported by language resources according to the META-NET White Papers (Rehm & Uszkoreit, 2012).

For domain adaptation, in addition to domain specific parallel and monolingual texts Tilde technology allows imposing predefined terminology on the given MT system (Pinnis, 2015). To benefit from this feature, terminology data needed to be specified, collected and attached to the MT system.

[1] http://www.latvija.lv/
[2] http://www.kis.gov.lv/
[3] http://www.tilde.lv

## 2.2 Types and sources of corpora

To collect required language resources, we identified several sources which we present in this section grouped by their type.

**Public corpora**

Several public corpora were identified and used as a source of parallel data:

- Acquis Communautaire corpus JRC-Acquis (Steinberger et al., 2006);
- Corpora published by DGT of European Commission – DGT-TM (Steinberger et al., 2012);
- Digital Corpus of the European Parliament – DCEP (Hajlaoui et al., 2014);
- Corpora of European Parliament proceedings Europarl v6 and v7 (Koehn, 2005);
- OPUS collection (Tiedemann, 2012) – OPUS EMEA, OPUS ECB, EUconst (Tiedemann, 2009);
- Multilingual Corpus from United Nation Documents – MultiUN (Eisele & Chen, 2010);
- WMT News Commentary corpus[4].

Several public corpora were identified as very promising by size, but after more detailed analysis we decided **not to use** them as a source of parallel data for different quality issues:

- OPUS collection of Open Subtitles 2011-2013 – various issues identified: a lot of imprecise translations (unknown source), segment shift, transliteration issues etc.
- OPUS ECB[5] corpus containing data from the European Central Bank – special processing is needed to use this coprora. Diacritics in many European languages are encoded as html entities yet with spaces added before and after. That breaks many words in pieces.

**Crawling Web sources**

Parallel data for the Latvian language is scarce, and effort must be put to find different possible sources, assess them and collect the data. When identifying the sources, attention must be paid how parallel is the data – e.g., seemingly identical web news published on multilingual websites often are not one-to-one translations but are adapted for the target language communities. Although such comparable multilingual data could also be useful to extract data for statistical MT (Skadiņa et al., 2012), it requires significant additional efforts and validation, this is why we discarded sources that are not fully parallel.

We identified and collected the following useful data. We tried and examined different data collection techniques, too.

- Parallel in-domain content from **public institution websites** collected manually and aligned with the Microsoft aligner (see section 2.4 for details). 1200 Latvian, 700 English and 1100 Russian web pages yielded 8493 English-Latvian and 13373 Latvian-Russian segments.
- Parallel in-domain content from translations of **international documents** that are not already included in the existing public corpora. We identified 2500 Latvian language documents on the website of Latvian legislation matching the categories "convention, declaration, international document, international agreement, international law", and we sought the Web for the matching counterpart in English and Russian. We were able to find 150 counterparts in English and 115 in Russian languages resulting in a corpus of 40 000 English-Latvian and 42 000 Latvian-Russian segments respectively. See some document titles as examples:

EN Maritime Labour Convention
LV Konvencija par darbu jūrniecībā
RU Конвенция о труде в морском судоходстве

EN Council Of Europe Convention On The Exercise Of Children's Rights
LV Eiropas Padomes Konvencija par bērnu tiesību piemērošanu
RU Европейская конвенция об осуществлении прав детей

- Parallel in-domain content from **European Commission Press Release Database** RAPID[6]. A custom workflow was created with PERL in multiple steps: generating search requests, getting the document URLs; downloading the html files; converting HTML to TXT, aligning the parallel files with Microsoft aligner. 5000 documents in Latvian were acquired which were aligned with their English counterparts resulting in 285 000 English-Latvian parallel segments[7,8].
- We found and processed some parallel datasets – **classifications from Eurostat's Metadata Server RAMON**[9], such as NACE, PRODCOM, Combined Nomenclature, and added them to the parallel corpus, obtaining 21 100 parallel English-Latvian segments.

We explored numerous other websites which contain multilingual content which may be relevant for the project purpose, but we had to discard them due to complexity of data extraction. Specific structure of these sites would take too much manual work to find, assess and extract pieces of parallel content, which would not be justified by the possible gains. For efficiency reasons we mostly focused on plain text / html sources; limiting processing of other formats like PDF only for most valuable data.

[4] http://www.statmt.org/wmt14/translation-task.html
[5] http://opus.lingfil.uu.se/ECB.php
[6] http://europa.eu/rapid/
[7] https://www.letsmt.eu/CorporaDetails.aspx?id=c-41986cbc-650c-4e27-ab85-b2755453733f
[8] https://www.letsmt.eu/CorporaDetails.aspx?id=c-9259a040-fc82-4d48-aa7c-19a16474259b
[9] http://ec.europa.eu/eurostat/ramon/

**Parallel texts from publishers**

During the project, we identified some publishers who produce and publish parallel content on regular basis. We got a permission from the publisher of magazine which has identical versions in Latvian and Russian. Another publisher performs translation of Latvian legislation into Russian to make it better understandable for the Russian-speaking entrepreneurs. We managed to reach special licensing agreements with these publishers enabling us to process their data and create an in-domain parallel Latvian-Russian corpus of over 450 000 parallel segments.

**Data from public administrations**

Several public institutions provided their taxonomies and collection of documents for customization of MT. We had to admit that it was less than 1% of the total size of corpora that was collected for MT training. The major obstacle is a lack of data management practices in public institutions that would make it easy to select and submit shareable data. A good-will of public institutions to support the project was outweighed by the need to spend additional efforts to prepare the data resulting in a very few resources.

The best data obtained from the public administrations:

- Press releases and multilingual news of home page of Ministry of Foreign Affairs[10]
- Monolingual transcripts of the plenary sittings of Saeima (Latvian Parliament) [11] resulted in 824 000 Latvian monolingual sentences
- State Policy Planning documents database PolSis [12] fcontained 970 000 monolingual sentences
- multilingual website of state e-services[13] provided data for 640 000 parallel segments.

## 2.3 Data processing workflows

The typical workflow that we applied in the project consists of multiple steps. The goal in corpus data processing for MT is to convert the data in Moses (monolingual or parallel plaintext) format or in TMX format files to be imported into LetsMT Resource Repository. Depending on corpora source, it may require doing all or only part of the steps described below:

- **Identifying the source** (URL / publisher/ other source). It may require searching online, 'keeping an eye open', being open-minded to discover unexpected new sources. Another method is brute force scanning of many thousands of domain starting pages and checking for multilingual links – an approach we have not used for production due to resource limitation.
- **Assessing the source for IPR** restrictions and data privacy issues before starting content collection.
- **Collecting the data** for source and target language can mean anything from downloading a zip containing all the data to scripting a crawler to download the content files do be normalised and further processed.
- **Normalizing** of text, files and formats: Converting to UTF-8 encoded plain text format (PDF->HTML->TXT, DOC->TXT etc. conversions). Printed sources have to be scanned to TIFF/JPEG, then OCR-ed to DOC->TXT. The text split for column wrapping must be restored. Hyphenated words should be restored. Headers, Footers, Footnotes, page numbers should be removed. Sentences split by page breaks should be restored. The text in files must be split by sentences.
- **Aligning the documents** – if needed, by analysing document comparability. Selecting the matched pairs and filtering matching documents above a certain threshold.
- **Aligning parallel segments**
- **Evaluating** the results before including the result in MT training.
- Adding the corpus to the Resource Repository and using in the creation of **statistical models for MT**.
- Evaluating the quality of resulting MT system against a baseline to see whether the added data has yielded improvements in MT quality as measured by automated score (e.g. BLEU).

## 2.4 Data processing tools

Each of the corpora type and particular corpus was processed individually in order to be prepared for the use in MT training.

Collecting web data includes crawling techniques, filtering of boilerplate, texts in other languages, noise; tokenisation, alignment, conversion to a single UTF-8 encoding. Project specific methodology was developed which includes application of open source tools as well as custom toolkit developed for this project:

- We use custom PERL scripts, Teleport commercial webspider [14] and *wget* utility to crawl and collect specific content from the Web.
- We use jusText (Pomikálek, 2001) library as the boilerplate removal technique from web pages.
- We use FineReader Pro to do OCR and extract texts from printed sources.
- We use Notepad++ for routine file operations, as well as EditPad Pro to work with huge text files >2GB.
- Custom tokenisation tools were developed as part of the HUGO.LV toolkit.
- We use DICMETRIC of ACCURAT Toolkit[15] (Pinnis et al., 2012) to perform alignment of potentially parallel documents.
- We use Microsoft's Bilingual sentence aligner (Moore, 2002) to align all kinds of parallel texts from aligned files. We tried also Hunalign (Varga et al., 2005), and Vanilla (Gale & Church, 1993). The alignment of Microsoft's Bilingual sentence aligner led to the most accurate results (Skadiņš et al., 2014).
- Custom PERL script tools were developed to convert between encodings, to convert between formats, to merge hyphenated words, to filter data and integrate tools into workflows.

[10] http://www.mfa.gov.lv/
[11] http://saeima.lv/en/transcripts
[12] http://polsis.mk.gov.lv/
[13] https://www.latvija.lv/
[14] http://www.tenmax.com/teleport/home.htm
[15] http://www.accurat-project.eu/index.php?p=accurat-toolkit

HUGO.LV toolkit also includes tools to build data subsets for human evaluation of QA of the new corpora we add to the repository and use in training of new MT systems.

### 2.5 Challenges

**LV-RU parallel texts.** Main challenges in this project are related to insufficiency of language resources for the small Latvian language. Very useful source for parallel Latvian English data are open multilingual corpora of European institutions. Unfortunately, the Russian language is not part of that, and available Latvian-Russian parallel texts are sparse. Collecting significant amount of Latvian-Russian parallel data is among the major achievements of the project.

**Content from State Administration.** We assumed that public institutions produce many text documents, and they should be motivated to contribute them to get better MT systems in return. Although we addressed all the ministries with a request to identify and share the textual data, we got only around 100 text files back. It can be explained by the lack of proper data management procedures that would allow easy selection and provision of the required data.

**Extracting text from PDF format.** Some institutions publish their parallel content online in PDF format only, and no other formats are available. Although technically and theoretically possible, practically extraction of texts from PDFs takes a lot of efforts as they are generated in different ways. There is no single way to extract data from PDFs as there are numerous tools to build them. Each source must be examined individually before building a workflow (Skadiņš et al., 2014).

**Putting PSI Directive into practice**. European Union Directive on the re-use of public sector information (PSI Directive)[16] opens data held by public sector bodies for re-use beyond its initial purpose of collection without restrictions for commercial and non-commercial purposes. We expected that this directive will provide us access to the data translated in public institutions.

Large part of public sector translations is outsourced to translation service providers. As nowadays almost all professional translation bureaus use computer aided translation tools that have translation memories (TM), we expected to receive these TMs which are very valuable parallel data in segment aligned format.

We identified that public procurement results in the area of outsourced translation services are published on the portal of the Procurement Monitoring Bureau[17]. In the past there was no requirement for service providers to deliver Translation Memories (TM) as part of the service. None of the recipients of such services, mostly state institutions, had any TM files to contribute. It is still a challenge for State Administration to put PSI directive in relation of translations done in public sector to be made available for public, free and open reuse.

**IPR restrictions.** A lot of useful data is protected by the intellectual property rights which do not allow data sharing and reuse without explicit permission from the data owners. Publishers and data owners are protecting and securing their intellectual assets, however after close consideration they are sometimes positive to contribute for a very specific use in MT.

To use such texts with a permission from the data owners while still protecting their intellectual property, other type of data can be generated and shared instead of the original text such as n-grams, shuffled extracts, phrase tables or binary models.

To foster development of machine translation and other data-driven language technologies, it would be necessary to modernize European Union copyright legislation to open copyrighted data for use in research and development that does not infringe the normal exploitation of the copyrighted work, such as creation of statistical models and machine learning.

## 3. Addressing resource scarceness challenge

### 3.1 Terminology data

**Terminologies and taxonomies.**

If MT system is trained on a very large parallel corpus, it can "learn" how to translate terms from the term occurrences in the data. Since the parallel corpora is scarce for Latvian language, we must apply other approach to ensure proper translation of in-domain terms. MT systems in this project were enhanced by using dynamic terminology and named entity integration in statistical machine translation (Pinnis, 2015).

We collected in-domain terminology and taxonomies and added to the MT system. The following types of named entities were prepared:

- Names of state institutions and their translations;
- Names of professions and their translations;
- Street and place names and their transcriptions / translations;
- Popular person names and surnames and their transcriptions;
- Geographical names – cities, states, villages etc. and their matching counterparts.

These lists underwent special filtering to ensure that these named entities do not conflict with common names. We ended with 9300 entries for English-Latvian and 8200 entries for Latvian-Russian.

**Resources from IATE[18] (Inter-Active Terminology for Europe) database.**

IATE data was made public during the project and we considered using its terminology data as either additional MT training data or as a source for terminology to be used in dynamic integration. Closer analysis showed that:

- A term entry in one language may be matched with abbreviation instead of full term in another language;
- Definitions of terms in different languages are not always direct translations;

[16] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

[17] http://www.iub.gov.lv/

[18] http://iate.europa.eu/

- Abbreviations and full names are used inconsistently.

We concluded that this data is not suitable as parallel content "as is". IATE can serve as a helpful reference to translators, but we could not include this data as parallel data source in this project. Elaborated content cleaning / filtering techniques must be applied to IATE data to select the parallel terms or definitions to be used as MT training data, but this was beyond the scope of this project.

### 3.2 Using MT to produce additional parallel data

In order to deal with data scarceness for Latvian-Russian language pair, we experimented with building a Latvian-Russian parallel corpus using a pivot language. We added such automatically generated corpus to the MT training data and evaluated quality improvements of resulting MT system to determine whether such approach could be used in production.

We performed an experiment using a large English-Russian parallel corpus – MultiUN (Eisele & Chen, 2010) with 9.4M segments of translated UN documents. We translated this corpus from English to Latvian, and aligned Russian part of the segment with the machine-translated Latvian part.

We translated this corpus from English to Latvian using Tilde MT engine which outperforms in quality other English-Latvian MT systems (Skadiņš et al., 2014).

One of the challenges to perform the test was to get the this large corpus translated in a reasonable time. Assuming 1 sec/segment translation speed, it requires 4 months. MT in multiple queues was used, and the total time was reduced to less than 1 calendar month.

To decide whether the obtained data are good and shall be used in a production system, we built two MT systems for comparison. One was a baseline system with 3.8M parallel segments of quality training data. Another system was the experimental system made from the baseline clone with the Latvian-Russian data added. We used the same training settings and tuning and evaluation data sets. The results obtained are presented in Table 1.

| MT System | Parallel Data Volume | BLEU |
|---|---|---|
| Baseline | 3.8M | 59.41 |
| System with test data added | 12.3M | 58.15 |

Table 1. Evaluation of MT-generated parallel corpus

We were anticipating quality improvement to consider including the MT-generated parallel corpus in production systems. With 1-point BLEU drop it means this approach is not feasible yet, so we decided not to use this method in building production MT system in this project.

## 4. Results Achieved

### 4.1 Corpora size obtained

At the end of the project we collected a significant amount of new MT training data (See Table 2) for all project language pairs and for both general domain and state administration domain, exceeding the requirements set for the project.

| Corpus language (pair) | Corpus Type | Domain | Corpus size (millions of sentences) |
|---|---|---|---|
| English-Latvian | Parallel | General | 5.8 |
| Latvian-Russian | Parallel | General | 5.1 |
| English-Latvian | Parallel | State Adm. | 3.3 |
| Latvian-Russian | Parallel | State Adm. | 2.0 |
| English | Monol. | General | 50 |
| Latvian | Monol. | General | 75 |
| Russian | Monol. | General | 75 |
| English | Monol. | State Adm. | 15 |
| Latvian | Monol. | State Adm. | 25 |
| Russian | Monol. | State Adm. | 24 |

Table 2. The amount of collected MT training data

### 4.2 Human evaluation of corpora

Before putting to use, we evaluated each newly created parallel corpora. Some of the corpora may be small – contain 5 to 10 thousand segments, other bigger corpora may contain 100 to 500 thousand segments. We applied a human evaluation method of evaluating a subset of the corpus of 50..200 randomly selected segments to represent the entire corpus. The annotation was very basic – Good/Still Acceptable/Bad.

- **Good** means that source and target language sentences are parallel.
- **Still acceptable** (still good) means one or two typos or minor errors beyond the alignment process; style aspects attributable to the translator preferences.
- **Bad** means content in the supposedly parallel sentences has major differences; or if two or more words are split (possibly due to hyphenation or extraction from PDF), or have incorrect characters in them (due to OCR or PDF, or encoding issues).

After annotation we check the percentage score – a simple formula of dividing the number of good segments by the number of total segments.

We have set a quality threshold of over 90% good segments to consider a corpus to be good for use in MT system training. If the quality was below this threshold, we checked the process of building the corpus – its source files, the segment alignment process. If problems cannot be fixed, the lower quality data is rejected. Although our previous research (Skadiņš et al., 2014) shows that even data with much lower quality level can lead to improvement in BLEU score, in this case we set high corpus quality criteria to avoid random erroneous words in the MT output.

### 4.3 Evaluation of MT systems

The collected corpus was used to build both general domain and state administration domain MT systems. We evaluated MT systems using BLEU score metric (Papineni et al., 2002) and compared results to Google Translate (See Table 3).

General domain MT systems were trained using all collected data, and state administration domain MT systems were trained using two language models – in-domain language model, and general domain language model. We used domain adaptation methods suggested in earlier research by Koehn and Schroeder (2007) and Lewis et al. (2010). Both language models have different weights determined with system tuning by MERT (Och, 2003) using in-domain tuning corpus.

| **System** | | **BLEU** | |
|---|---|---|---|
| **Language pair** | **Domain** | **HUGO.LV** | **Google Translate** |
| English-Latvian | General | 34.85 | 31.05 |
| English-Latvian | State administration | 55.58 | 26.05 |
| Latvian-English | General | 44.11 | 42.92 |
| Latvian-English | State administration | 60.93 | 28.00 |
| Latvian-Russian | General | 40.66 | 14.41 |
| Latvian-Russian | State administration | 65.88 | 19.72 |

Table 3. Results of automatic evaluation

We compared general domain systems to Google Translate - for all three systems our results were significantly better. Careful collection and procession of training data have made a major contribution to these results.

## 5. Conclusions

The project successfully achieved its goal to collect the maximum of data useful for training MT systems adopted to the needs of e-Government. Data was collected from various sources that are described in this paper. Up to the knowledge of authors this is the largest collection of data readily available for the generation of Latvian SMT systems.

Collected data was used to build six MT systems for the needs of Latvia public sector, including general domain systems, all of which outperform Google Translate in both BLUE score and human evaluation. Another indicator of success is that the resulting MT systems are integrated in the portal Latvija.lv and public MT service HUGO.LV. The platform is nominated for the World Summit Award[19] and World Summit on the Information Society Prize[20].

Project partners Centre of Cultural Information Systems and Tilde take active part in the newly started European Language Resource Coordination action that will open the use of project results for the CEF Automated Translation digital service infrastructure[21].

Another positive effect of the project is a growing awareness in the public sector institutions about the importance of their textual data for language technology development. State institutions are encouraged to take care of their translation data – to require translation memories to be returned as part of each translation contract, to collect, anonymise, reuse and publish the translation data. This process will gradually lead to more parallel data available for research and practical developments.

## 6. Bibliographical References


Eisele, A., Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, B. Maegaard, … N. C. (Conference Chair) (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation* (pp. 2868–2872). European Language Resources Association (ELRA).

Gale, W.A., Church, K.W. (1993). A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1), pp. 75-102.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., Varga, D. (2014). DCEP–Digital Corpus of the European Parliament. In N. C. (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit 11*, 79–86. Retrieved from http://mt-archive.info/MTS-2005-Koehn.pdf

Koehn, Philipp, Federico M., Cowan B., Zens R., Duer C., Bojar O., Constantin A., Herbst E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, 177-180.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague

Lewis, W., Wendt, C., and Bullock, D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*

Moore, R.C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation:*


[19] http://www.wsis-award.org/news/world-summit-award-nominees-2015-136420150820

[20] http://groups.itu.int/stocktaking/WSISPrizes/WSISPrizes2016.aspx#nominated-projects

[21] http://lr-coordination.eu

*From Research to Real Users*. (pp. 135-144). London, UK: Springer-Verlag.

Rehm, G., Uszkoreit, H. (eds). (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age.* Springer, Heidelberg.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *ACL 2003: Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 160–167).

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*.: ACL

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., Babych, B. (2012). ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 91–96). Association for Computational Linguistics. Jeju, South Korea.

Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation. In *Proceedings of the Eighteenth Annual Conference of the European Association for Machine Translation (EAMT 2015)* (pp. 89–96). Antalya: European Association for Machine Translation.

Pomikálek. J., (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis. Masaryk University. Brno.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlič, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M.L., Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 438-445) Istanbul, Turkey.

Skadiņš, R., Tiedemann, J., Rozis, R., Deksne, D. (2014). Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In N. C. (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1850–1855). Reykjavik, Iceland: European Language Resources Association (ELRA)

Skadiņš, R., Šics, V., & Rozis, R. (2014). Building the World's Best General Domain MT for Baltic Languages. In *Human Language Technologies–The Baltic Perspective-Proceedings of the Sixth International Conference Baltic HLT*. Kauunas, Lithuania.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, (pp. 24-26). Genoa, Italy

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P. (2012). DGT-TM: A freely available Translation Memory in 22 languages. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA)

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, (pp. 590-596)

Vasiļjevs, A., Skadiņš, R., Tiedemann, J. (2011). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In *Proceedings of MT Summit XIII: the Thirteenth Machine Translation Summit* (pp. 507–511). Retrieved from http://www.mt-archive.info/MTS-2011-Vasiljevs.pdf

Vasiļjevs, A., Skadiņš, R., Tiedemann, J. (2012). LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In M. Zhang (Ed.), *Proceedings of the ACL 2012 System Demonstrations* (pp. 43–48). Jeju Island, Korea: Association for Computational Linguistics

Vasiļjevs, A., Kalniņš, R., Pinnis, M., Skadiņš, R. (2014). Machine translation for e-Government – the Baltic case. In O. Beregovaya, M. Dillinger, J. Doyon, R. Flournoy, P. O'Neill-Brown, & C. Simmons (Eds.), *Proceedings of AMTA 2014, vol. 2: MT Users* (pp. 181–193). Vancouver, BC. doi:10.13140/2.1.4498.4323

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing* (Vol. 5, pp. 237-248).

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).