Bridging the Language Barriers -the Baltic Case. Session 1 - Problems and Solutions on Baltic Shores

Andrejs Vasiļjevs

TRALOGY

Bridging the Language Barriers – the Baltic Case

Andrejs Vasiļjevs
Tilde
andrejs@tilde.com

TRALOGY II - session 1
Date d'intervention : 17/01/2013

This paper is an overview of the work that the software company Tilde is doing to develop technologies and applications for the languages of the Baltic countries. We present the challenges faced by these languages due to their complex morphological structure and limited availability of language resources. We talk about different approaches explored at Tilde to address these challenges in the development of machine translation solutions and automatic methods for acquisition of parallel data from the Web. In addition the combination of data driven techniques with knowledge based methods are discussed. We present Tilde work on the consolidation of terminology resources, creation of advanced terminology services and the application of domain specific terminology for customized machine translation. We describe our activities in the creation of an open European language resource infrastructure and populating it with numerous resources for our languages.

lien video : http://webcast.in2p3.fr/videos-bridging_the_language_barriers_the_baltic_case

Introduction

The rapid speed of technological development poses a critical challenge to numerous languages. We are increasingly dependent on technologies in our information, communication and entertainment needs. Language is and will remain our key instrument for conveying information and socialization. But language usage is more and more mediated and facilitated by technological means. Word processors correct mistakes in our writings, we depend on search engines to find information, we use machine translation to access information in foreign language, we give voice commands to cars and voice queries to our smartphones – language and technologies are intertwined everywhere.

Language technology is a key enabler of the knowledge society (Rehm and Uszkoreit, 2012a). Comprehensive support and usage of language in popular technological platforms nowadays is as important as was the introduction of language into printing a few hundred years ago. The so called secondary Guttenberg effect puts a language under the threat of extinction if it is not sufficiently armed for the digital age. This resembles the process of the gradual disappearance of many languages that were not used in printed publications.

For this reason the leading European experts have raised the alarm about the differences in technology support between the various European languages. More than 200 experts from academia and industry participated in the assessment of the European language landscape and the preparation of the White Paper Series "Europe's Languages in the Digital Age". The 30 volumes of this series describe language technology support for the 23 official European Union languages and several other major European languages (Rehm and Uszkoreit, 2012b). According to the expert findings only the largest European languages have comprehensive technological support. Major gaps and deficiencies in key technological tools and resources put the long term survival of at least 21 European languages under serious risk.

These risks are particularly worrying for the Baltic languages Latvian and Lithuanian that are among the smallest official EU languages. The Lithuanian economy is in the 83rd position and Latvian in 101st position on the global scale[1] (IMF, 2012). Consequently these markets are not sufficiently large to motivate global IT companies to make significant investments in technological development for the Baltic languages. National research budgets in the Baltic countries are very limited. Government budget appropriation for research and development in these countries are among the very lowest in EU (Eurostat, 2011). Public investments are by far insufficient to lay the necessary research foundation for scientifically grounded language technology development. The technological gap has further widened due to the lack of dedicated language technology funding in the EU 6th Framework Programme for Research and Development after significant investments for other EU languages in the 80-ies and 90-ties when the Baltic countries were not yet members of the EU.

In this paper we describe the approach taken by the language technology company Tilde[2] to address these challenges in its mission to provide Baltic users with the same technological opportunities that are enjoyed by the larger language communities. Tilde was established in 1991 in Latvia right after this country regained its independence. In these years Tilde has grown to become the leading language technology and localization service company in the Baltic countries with offices in Riga, Vilnius and Tallinn. Tilde develops widely used software tools for Baltic languages like spelling and grammar checkers, electronic dictionaries, spoken language technologies, terminology databases and machine translation systems. Tilde's tools are very popular in the Baltics and have more than 350 000 users.

The following sections provide a general characterization of the Baltic languages, overall assessment of their language technology support, and gives and overview of Tilde's work

(1) Gross domestic product based on purchasing-power-parity (PPP) valuation of country GDP
(2) http://www.tilde.com/

advancing statistical machine translation for the needs of the Baltic and other smaller languages, developing terminology systems and services and participating in the creation of an open European language resource infrastructure.

1. Baltic languages in the digital age

1.1 General overview of Baltic languages

Latvian and Lithuanian are the only two surviving languages of the Baltic branch of the Indo-European language family. These are among the oldest European languages. It should be noted that Estonian – the language of the third Baltic country – is not in the same linguistic group and is very different from Latvian and Lithuanian. It does not belong to the family of Indo-European languages but to the Finnic branch of Uralic languages.

Though apparently small, Lithuanian and Latvian rank 144th and 150th respectively among the most spoken languages out of about 6,900 languages on our planet. Latvian is spoken by about 1.5 million native speakers. Lithuanian is spoken by twice as many native speakers totaling more than 3 million speakers in Lithuania and other countries.

Both Latvian and Lithuanian are highly inflected languages that have rich morphology with various derivational means. In Baltic languages word order is relatively free and syntactic relations are determined by morphological forms.

Both languages are based on the Latin script alphabet supplemented with diacritical marks that differ for these languages. The Lithuanian alphabet has 32 letters and the Latvian 33 letters. The Lithuanian language is considered the most conservative of the living Indo-European languages preserving many features that have since disappeared in other languages. Although more evolved over time Latvian also shares many similar features like rich inflections, free word order and complex grammatical structure.

Latvian and Lithuanian are among the 23 official languages of European Union and are the sole official languages in the Republic of Latvia and Republic of Lithuania respectively. Although Latvian and Lithuanian belong to the same language group and are spoken in neighboring countries, speakers of both languages cannot communicate with each other freely (Rehm and Uszkoreit, 2012b).

Both languages are well represented on the Web. The number of internet domains with the extension .lv and .lt and content mostly in Latvian and Lithuanian exceeds 100 000 and 130 000 respectively.

2. Language technologies for Latvian and Lithuanian

A comparative assessment of language technology development for Baltic languages was carried out within the framework of the META-NORD project (Skadiņa et al., 2012). META-NORD is a cooperation project coordinated by Tilde involving leading language technology research institutions in the Baltic and Nordic countries – the Institute of Lithuanian Language, Tartu University, University of Gothenburg, University of Bergen, University of Helsinki, University of Copenhagen, and University of Iceland. META-NORD covered the Baltic and Nordic parts of the META-NET collaboration network.

Assessment included the evaluation of language technology support and the core application areas of language and speech technology (e.g., language checking, web search, speech interaction,

machine translation, etc.) based on the following criteria: quantity, availability, quality, coverage, maturity, sustainability, and adaptability.

The comparison presents the situation for four key areas: machine translation, speech processing, text analysis, and resources. This study puts the smaller languages of the Baltic region – Latvian, and Lithuanian – in the last cluster, defined as having major gaps for all of the four key areas (see Table 1). The relative ranking of the remaining five languages is slightly higher, although none of them comes close to the "larger" languages (English, French, Spanish, and German). "Moderate" support is provided only for Finnish in speech technologies and for Swedish with respect to language resources.

The results indicate that only with respect to the most basic tools and resources, such as tokenisers, PoS taggers, morphological analysers/generators, reference corpora, and lexicons/terminologies, the status is reasonably positive for all languages. However, tools for information extraction, machine translation, and speech recognition, as well as resources such as parallel corpora, speech corpora, and grammar, are rather simple and have limited functionality for some of the languages. For the most advanced tools and resources, such as discourse processing, dialogue management, semantics and discourse corpora, and ontological resources, most of the languages either have nothing of the kind, or their tools and resources have a quite limited scope.

Besides objective limitations dealing with complex languages there are also other obstacles like the lack of continuity in research and development funding. Due to limited funding, Latvian language technology support has not reached the quality and coverage not only of that for English, but also for many under-resourced languages of the Baltic and Nordic region with a smaller number of speakers. Although different resources have been developed for Latvian by research and industry institutions (Vasiļjevs and Skadiņa, 2012), targeted national research and development activities are urgently needed to fill the gaps in language resources and tools.

3. Developing machine translation to bridge language barriers

3.1 Approaches in creating statistical MT systems

With internet infrastructure becoming omnipresent, language diversity remains one of the last barriers in accessing information and cross-national communication. The exponentially growing volume of multilingual information by far exceeds the capacity of human translators to meet demand for translation. Machine translation is the only viable solution for instant and cheap access to information in foreign languages. This is why machine translation (MT) is among the most critical language technologies also for the Baltic languages.

Machine translation has been a particularly difficult problem in the area of natural language processing since it was first proposed as a concept in the early 1940-ies. Till recently the dominant approach in MT was the so-called rule-based strategy. It is based on linguistic rules and rich translation lexicons to analyze source language text and generate translation in the target language. This approach was widely used in developing MT solutions for larger languages and resulted in numerous commercial MT systems, e.g. Systran, PROMT and others. Tilde's work on rule-based MT took almost a decade and resulted in English-Latvian and Latvian-Russian MT systems released to the market in 2008. The major drawback of the rule-based MT strategy is the immense time and human resources that are needed for every language pair and for enhancing the quality of translation.

Table 1: Availability of LRT for languages of the Baltic and Nordic countries

Speech processing				
Excellent	**Good**	**Moderate**	**Fragmentary**	**Weak/None**
	English	Czech, Dutch, **Finnish,** French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, **Danish, Estonian,** Galician, Greek, Hungarian, Irish, **Norwegian,** Polish, Serbian, Slovak, Slovene, **Swedish**	Croatian, **Icelandic, Latvian, Lithuanian,** Maltese, Romanian
Machine Translation				
Excellent	**Good**	**Moderate**	**Fragmentary**	**Weak/None**
	English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian., Czech, **Danish, Estonian, Finnish,** Galician, **Icelandic,** Irish, **Latvian, Lithuanian,** Maltese, **Norwegian,** Portuguese, Serbian, Slovak, Slovene, **Swedish**
Text Analysis				
Excellent	**Good**	**Moderate**	**Fragmentary**	**Weak/None**
	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, **Danish, Finnish,** Galician, Greek, Hungarian, **Norwegian,** Polish, Portuguese, Romanian, Slovak, Slovene, **Swedish**	Croatian, **Estonian, Icelandic,** Irish, **Latvian, Lithuanian,** Maltese, Serbian
Resources				
Excellent	**Good**	**Moderate**	**Fragmentary**	**Weak/None**
	English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, **Swedish**	Basque, Bulgarian, Catalan, Croatian, **Danish, Estonian, Finnish,** Galician, Greek, **Norwegian,** Portuguese, Romanian, Serbian, Slovak, Slovene	**Icelandic,** Irish, **Latvian, Lithuanian,** Maltese

In recent years statistical machine translation (SMT) has provided a major breakthrough in MT development. SMT provides a cost effective and fast way to create MT systems. In this approach statistical models are built by analyzing huge volumes of parallel and monolingual text to guide the MT system in translating (translation model) and establishing the target language sentences (language model). In the translation process, from all the possible translation candidates the SMT system selects the one with the highest statistical probability to be the translation of a given source sentence.

Rapid adaption of SMT was particularly facilitated by the open-source corpus alignment tool GIZA++, and the MT training and decoding tool Moses. Another factor which facilitated the development of MT for many languages was the availability of the EU translation corpus and other

parallel data on the internet. The EuroMatrix project[3] demonstrated how open source tools and publicly available data can be used to generate SMT systems for all language pairs of the official EU languages.

However, the quality of an SMT system largely depends on the size of training data. Obviously the majority of parallel data is in major languages. As a result SMT systems for larger languages are of much better quality compared to systems for under-resourced languages.

This quality gap is further deepened due to the complex linguistic structure of Baltic languages. To learn the complexity of rich morphological structure and free word order from corpus data by statistical methods, much larger volumes of training data are needed than for languages with simpler linguistic structure. For example, although the popular SMT system Google Translator currently covers more than 70 languages, translation quality for Baltic languages is significantly worse than for English, French, Spanish and other larger languages.

Tilde is putting a great deal of effort into collecting data to train SMT systems for Baltic languages. The parallel training corpus includes DGT-TM, OPUS and localization corpora. The DGT-TM corpus is a publicly provided collection of legislative texts available in 22 European Union languages (Steinberger et al., 2012). The OPUS translated text collection (Tiedemann, 2009) contains publicly available texts from the web in different domains. Among the proprietary corpora collected by Tilde is localization parallel corpus obtained from translation memories created by the Tilde Localization department during translation of software content, appliance user manuals and software help content. To increase word coverage we supplement parallel texts with word and phrase translations from bilingual dictionaries.

Assessing baseline SMT systems trained on this data we spotted that they were weak at picking the correct inflectional forms for translated lexical units. This very negatively affected the fluency of translation, particularly for adjective-noun and subject-object agreement. This showed that for highly inflectional languages a language model over surface forms might not be sufficient to estimate the probability of target sentence reliably. To address that, we introduced an additional language model over morphologic tags in the English-Latvian system (Skadiņš et al., 2010). The tags contain relevant morphologic properties (case, number, gender, etc.) that are generated by a morphologic tagger. We applied these tags as additional factors to the so called factored SMT to improve local word agreement and inter-phrase consistency.

Although in automated evaluation BLEU metric scores showed only a slight improvement (21.7% vs. 23.8%), human evaluation demonstrated a clear preference (58.67%) for factored SMT over the baseline SMT, which operates only with the surface forms.

To make it easier and faster to use potential of existing open SMT technologies for Latvian, Lithuanian and other smaller languages, Tilde initiated the development of the online cloud-based SMT platform called LetsMT! (Vasiljevs et al., 2010). This platform was created in the framework of the EU CIP-PSP programme project where Tilde cooperated with the Universities of Edinburgh, Zagreb, Copenhagen and Uppsala, the localization company Moravia and the semantic technology company SemLab.

LetsMT! is an online platform[4] that enables users to share translation data for MT training and to build tailored MT systems for different languages and domains on the basis of this data. An easy to use online interface allows the user to select parallel and monolingual data from the cloud repository, specify a few parameters and launch cloud-based training of custom SMT system (Figure 1). The user can also train a system on his own proprietary data uploaded to the system. Depending on the amount of data selected, training of SMT system may take from an hour to a couple of days.

(3) http://euromatrix.net/
(4) http://letsmt.eu

User trained systems are automatically evaluated using BLEU and other popular metrics. This is particularly handy for running several experiments to find the combination of training data that results in the best quality for a particular translation domain.

LetsMT! translation services can be used in several ways: through the web portal, through a widget provided for free inclusion in web-pages, through browser plug-ins, and through integration in computer-assisted translation (CAT) tools and various online and offline applications. Localization and translation businesses as well as other professional translators can use the LetsMT! platform to build custom SMT solutions from their translation memory data, and access these solutions in their productivity environments (typically, various CAT tools).

LetsMT! platform has dramatically reduced the time and resources needed for experimenting and developing SMT systems for Baltic and other languages. It helped Tilde to create numerous domain and task specific Latvian and Lithuanian SMT systems that provide significantly better translation quality than Google Translate.

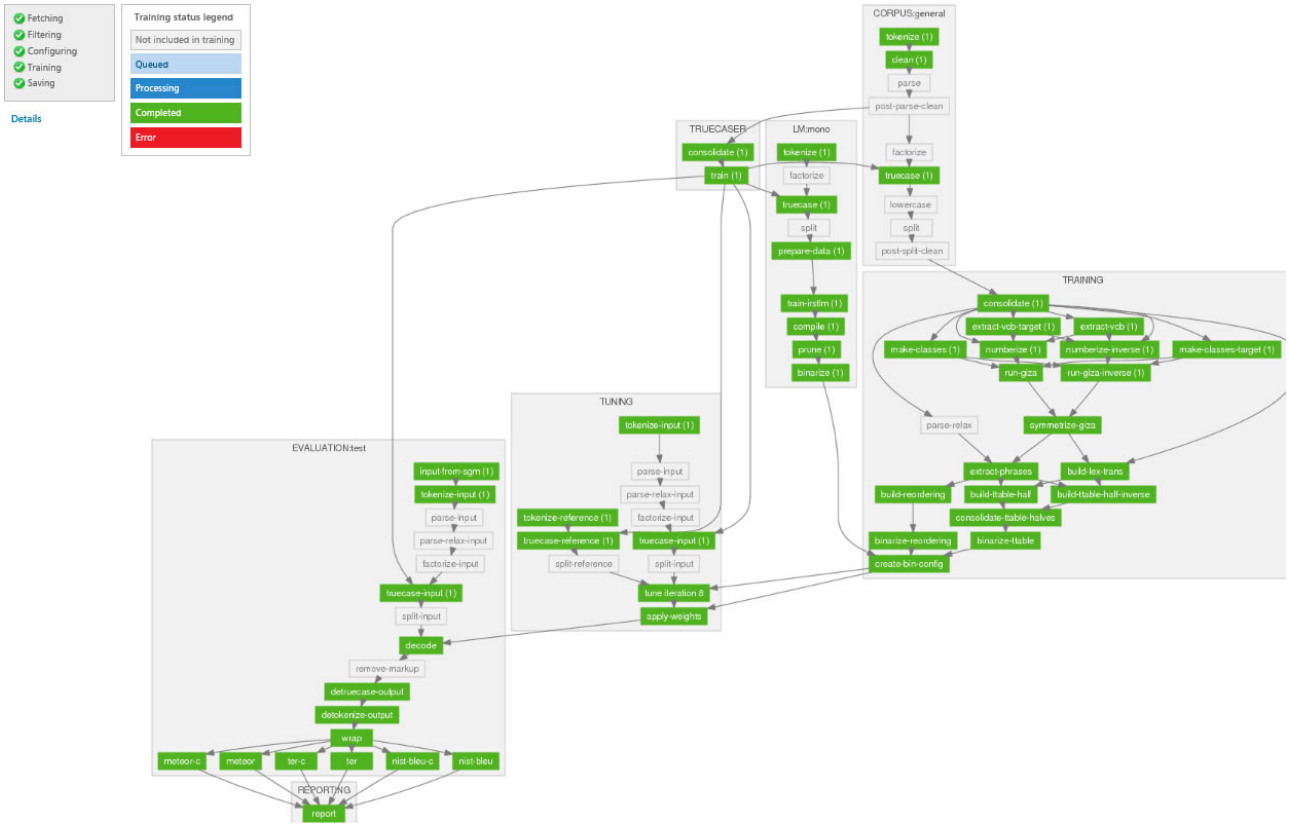

Figure 1: LetsMT! workflow diagramm allows tracking the progress of fully automated generation of custom SMT systems.

4. Acquiring more data for machine translation

Currently available parallel corpora for Baltic languages is not sufficient to further advance the quality of SMT. The solution for this problem proposed by Tilde in the ACCURAT project (Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation) was to exploit the fact that comparable corpora, i.e., non-parallel bi- or multilingual text resources are more widely available than parallel translation data.

A comparable corpus is a relatively recent concept in MT, corpus linguistics and NLP in general. In contrast to the notion of a parallel corpus, a comparable corpus can be defined as collection

of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period (McEnery and Xiao, 2007) in more than one language or variety of languages (EAGLES, 1996) that contain overlapping information (Munteanu and Marcu, 2005).

Comparable corpora have several obvious advantages over parallel corpora – they can draw on much richer, more available and more diverse sources which are produced every day (e.g. multilingual news feeds) and are available on the Web in large quantities for many languages and domains. Although the majority of these texts are not direct translations, they share a lot of common paragraphs, sentences, phrases, terms and named entities in different languages. Expansion of Web content with daily multilingual news feeds and large knowledge bases like Wikipedia make comparable corpora more widely available than parallel corpora.

The ACCURAT project was industry-research collaboration within the EU FP7 Framework Programme coordinated by Tilde with research partners University of Sheffield, University of Leeds, University of Zagreb, RACAI, ILSP and DFKI, and industry partners Linguatec and Zemanta. The key output of the project is an open source ACCURAT Toolkit**.** It contains tools for collecting comparable corpora, measuring comparability, data alignment at different levels and extraction of data useful for training statistical machine translation (SMT) systems. Besides the task specific tools, the toolkit also contains two general-purpose workflow chaining tools for particular usage scenarios: (1) the parallel data mining workflow which, given a comparable corpus, will output parallel or quasi-parallel phrases useful for SMT training, and (2) the named entity and term mapping workflow which, given a comparable corpus, will output translation lexicons of either named entities or terminology.

The ACCURAT toolkit produces:

- comparable document pairs with comparability scores, allowing to estimate the overall comparability of corpora;
- parallel sentences which can be used as additional parallel data sources for statistical translation model learning;
- terminology dictionaries — this type of data is expected to improve domain specific translation;
- named entity dictionaries.

Latvian and Lithuanian languages were the particular focus of the ACCURAT project. Tilde experimented with SMT domain adaptation for Baltic languages utilizing bilingual terms and bilingual comparable corpora collected from the Web. The results of these experiments showed that integration of terminology within SMT systems even with simple techniques (adding translated term pairs to the parallel data corpus or adding an in-domain language model) can achieve an SMT system quality improvement of up to 23.1% over the baseline system. Transformation of translation model phrase tables into term-aware phrase tables can boost the quality up to 24.1% over the baseline system mostly because of wrong translation candidate filtering in the translation process.

Data collected for Baltic languages supplements parallel and monolingual data stored in the repository of LetsMT! Platform. Currently this repository includes 34.2M parallel sentences for English-Latvian, 22.9M for English-Lithuanian, 3.4M sentences for Russian-Latvian and 4.5M for Russian-Lithuanian (Table 2).

Table 2 Paralell and monolingual data for training Baltic language SMT systems collected in LetsMT! repository.

	Latvian	*Lithuanian*
Bulgarian	3.5 M	3 M
Czech	5.9 M	5.9 M
Danish	6.1 M	6.3 M
German	6.6 M	7.6 M
Greek	5.3 M	5.3 M
English	34.2 M	22.9 M
Spanish	5.9 M	6.1 M
Estonian	8.2 M	8.3 M
Finnish	5.7 M	5.7 M
French	6.3 M	6.5 M
Hungarian	6.4 M	6.4 M
Italian	5.9 M	6 M
Lithuanian	7.6 M	169 M
Latvian	181.2 M	7.6 M
Dutch	5.7 M	5.9 M
Polish	6.6 M	9.1 M
Portuguese	6 M	6.3 M
Romanian	3.4 M	3.5 M
Russian	3.4 M	4.5 M
Slovak	6.2 M	6.3 M
Slovenian	6.3 M	6.4 M
Swedish	5.5 M	5.7 M
Turkish	0.1 M	0.2 M

5. Application of Baltic language SMT in localization

One of the main areas where we target development of statistical MT for Baltic languages is its application in translation and localization. This industry is experiencing growing pressure on time, efficiency and performance. Global vendors want to adapt their products for the small Baltic markets as inexpensively as possible. Volumes of texts to be translated are growing at a higher rate than the capacity of human translation, and translation results are expected in real-time.

Translation memories (TM) have been in use in localization for more than 10 years to increase productivity. Translation memories can significantly improve the efficiency of localization if the new text is similar to the previously translated material. However, if the text is in a different domain than the TM or in the same domain from a different customer using different terminology, support from the TM is minimal.

The objective of Tilde's work in this area is to increase the efficiency of the translation process without a degradation of quality. This can be achieved by combining traditional TMs with machine translation solutions adapted for the particular domain or customer requirements. Customization of SMT for a particular translation domain can be achieved by using previously translated data in the training of adapted SMT system.

We elaborated and evaluated this approach for translation in the IT domain. To create an English-Latvian SMT for this domain, Tilde used a corpus of 5.37M parallel sentence pairs from various fields, including 1.29M pairs in the IT domain. Additional tweaking was made by manually adding a factored model over the disambiguated morphological tags.

To integrate the SMT system in SDL Trados we developed a plug-in using the standard MT integration approach described in the SDL Trados SDK. If the source language segment is not found in the translation memory Trados translates it using the designated MT. Automatically translated segments are provided for the translator's consideration in the same way as translation memory suggestions. We clearly mark MT suggestions to distinguish them from TM suggestions, because MT output may be inaccurate, ungrammatical, it may use the wrong terminology etc.

We evaluated such MT assisted process against typical translation work where just translation memories are used. 5 translators with different skill levels participated in the evaluation. The results showed clear benefits from MT integration. Assistance from the machine translation increased the translation productivity by an average 32.9% – from 550 to 731 words per hour (Skadiņš et al., 2011). We have to note that there were significant performance differences in the various translation tasks and by individual translators.

In addition a quality assessment for texts was performed according to the standard internal quality assessment procedure. Although the error score increased for all translators (from 20.2 to 28.6 points in Tilde's proprietary metrics), it still remained at the quality evaluation grade "Good" (<30 points). This degradation is not critical and the result is acceptable for production purposes.

6. Developing terminology resources

In both human and machine translation a critical requirement for translation quality is the appropriateness and consistency of domain and project specific terminology. To facilitate development and accessibility of Latvian and Lithuanian terminology Tilde actively participates in terminology creation and standardization work, and the development of online terminology databases and services.

7. Consolidation of terminology in online databases

In partnership with the Terminology Commission of Academy of Science of Latvia Tilde has developed the Latvian online terminology database termnet.lv[5]. Numerous terminology glossaries have been integrated, many of which had to be digitized from the paper form. Currently termnet.lv includes more than 140 000 terms in 30 fields.

Experience in consolidating Latvian terminology served as the background for expanding terminology consolidation work on a pan-European level. In partnership with several national and international terminology centers Tilde created single online access point to multilingual European terminology – the EuroTermBank portal[6] (Rirdance and Vasiļjevs, 2006). It enables searching almost 2 million terms in over 25 languages. Under the term bank federation principle, it provides a single access point to the central database along with interlinked national and international term banks, consolidating terms from such major collections as IATE, WebTerm, Microsoft Terminology Collection, Terminology database of the Latvian Terminology Commission, and others.

(5) http://www.termnet.lv/
(6) http://www.eurotermbank.com

8. Application of terminology resources for customized MT

Most of the online terminology databases offer not much more than the typical database features of storing and querying terminology entries. The evolution of the Internet and cloud-

computing opens the opportunity to advance the automation of terminology and translation work by creating cloud-based terminology services for key terminology tasks. Such work is being carried out in the FP7 project Terminology as a Service (TaaS).

TaaS platform will provide a variety of online terminology services, to serve the needs for automated acquisition, processing, and application of terminological data by human users (i.e., language workers), for example:

- Automatic extraction of monolingual term candidates, using state-of-the-art terminology extraction techniques, from documents uploaded by users;
- Automatic lookup of translation equivalent term candidates in user-defined target language(s) from different terminology data- bases (for automatically extracted monolingual term candidates);
- Automatic extraction of translation equivalent term candidates from parallel and/or comparable Web data, using state-of-the-art terminology extraction and bilingual terminology alignment techniques (for automatically extracted monolingual term candidates);
- Facilities for cleaning up automatically acquired raw terminological data;
- Facilities for exporting terminological data in different formats, e.g., TSV, CSV, TBX, and others.

Terminology services can also be used by machine users (i.e., language processing applications), such as CAT tools, machine translation systems, search engines, and others. Thus, terminology services have the potential to significantly enhance the quality of language tools, and machine translation in particular.

The easiest method for terminology integration in SMT training is by adding the bilingual term collection to the parallel corpus that is used for generation of SMT system. Although the size of the term collection usually is relatively small in comparison to the whole parallel corpus, namely the presence of a term collection in training data helps the SMT training engine to build better word and phrase alignments, and it also fills gaps in the vocabulary by allowing translation of previously unknown terms.

In addition to this simple approach, we also propose to use online terminology services to tag terms in both parallel and monolingual corpora used in SMT training (Figure 2). We introduce an additional feature indicating phrases containing in-domain term translations in an SMT system's translation model. For this we have developed the phrasal level term tagging method. Using online terminology services we acquire a bilingual term collection from the corpus/corpora specified by a user and identify bilingual terms in SMT phrase tables (Pinnis et al., 2013).

Another process were terminology identification is helpful is in the translation phase. Preprocessing of the source text and marking terms and their possible translations assist th SMT system in making lexical choices of translation candidates.

Our experiments building an English-Latvian SMT system in the mechanical engineering domain show that such an approach achieves a relative SMT quality improvement of up to 6% according to BLEU (Pinnis and Skadiņš, 2012).

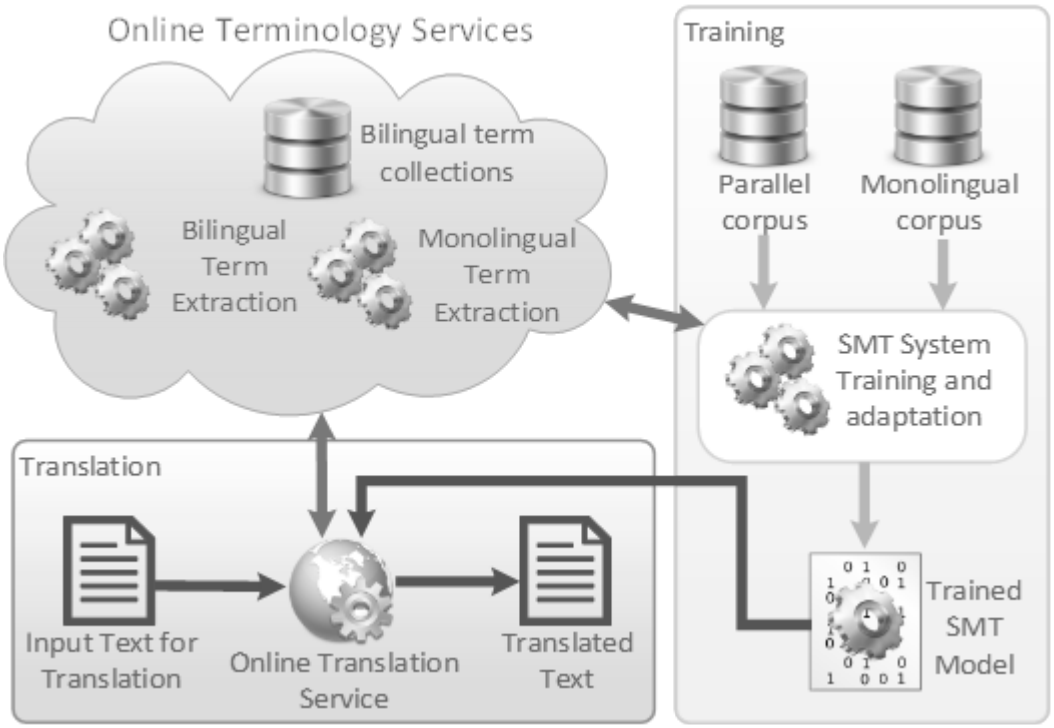

Figure 2 The conceptual design of the terminology service integration into statistical MT.

9. Baltic languages in language resource infrastructure

To facilitate the availability and usage of numerous Baltic language resources and tools developed by Tilde and other institutions it is important to ensure they are easy to find, that they follow commonly accepted standards and are interoperable, that they are free to use or there are clear licensing conditions, that sufficient description and documentation is provided.

For this purpose we take an active part in the development of the European linguistic infrastructure and the establishment of the META-NET cooperation network of research and industry institutions. META-NET is a network of excellence dedicated to fostering the technological foundations of a multilingual European information society through facilitating cooperation across different research fields, development of a common vision and strategic research agenda and establishment of an open language resource distribution platform.

Several projects were launched in the framework of META-NET supported by the FP7 and CIP ICT-PSP programmes of the European Commission. The META-NORD project coordinated by Tilde (Vasiļjevs et al., 2011; Skadiņa et al., 2011) worked on the assessment, description, enhancement and cataloguing of language resources in Baltic and Nordic countries.

For distribution and sharing of language resources, we and other META-NET projects use the distributed online platform META-SHARE (Piperidis, 2012). It consists of independent META-SHARE nodes set up in different countries and interlinked into a federated repository. This freely accessible distributed online infrastructure provides facilities for describing, storing, preserving language resources, and making them publicly available. Among various language resources that can be considered useful for different purposes, META-SHARE places a strong focus on language data that are important in development of language technology applications that are useful to EU citizens in their everyday communication and information search needs. META-SHARE

is intended for providers and users of language resources and technologies such as language technology developers, researchers, students, translators, technical writers and others. In the scope of the META-NORD project we augmented META-SHARE with 69 Latvian and 38 Lithuanian language resources and tools.

In Latvia the META-SHARE node is set up and maintained by Tilde[7]. In Lithuania the Institute of Lithuanian Language is responsible for maintaining the META-SHARE node[8]. According to the architecture of META-SHARE, all nodes are networked, and the content of the individual LR repository is harvested into the managing META-SHARE node, which for the META-NORD consortium resides at Tilde. In the managing node, information about the catalogued language resources is collected and synchronised with other managing nodes across Europe, thus providing access to the full catalogue of the pan-European infrastructure.

As a part of the activities related to populating META-SHARE with language resources, we wanted to extend the open linguistic infrastructure with multilingual terminology resources. Our intention was not to duplicate the resources stored on EuroTermBank, but to interconnect this content-specific language resource repository with META-SHARE. We implemented the integration of a language resource-specific node with META-SHARE via proxy: it connects to a META-SHARE managing node just like any other META-SHARE node – the LR metadata provider is proxied to the rest of the META-SHARE network.

Using this approach we enabled the harvesting of EuroTermBank metadata and integrated the EuroTermBank within the META-SHARE infrastructure. This interlinking yielded 99 additional terminology resources now listed in META-SHARE, including Latvian and Lithuanian terminology dictionaries. META-SHARE can use all the metadata of the terminology resources to catalogue, retrieve and link them. When users want to access actual terminology data they are provided with a direct link to the corresponding resource on the EuroTermBank (Skadiņa et al., 2013).

10. Conclusions

We described some major activities that foster the development of language technologies and resources for the Baltic languages Latvian and Lithuanian. Although the assessment of the META-NET experts includes Baltic languages in the cluster of the less supported languages of Europe in all key language resource categories, significant progress has been achieved in several areas.

Development of the cloud-based platform LetsMT! fully automates the generation of statistical machine translation systems. It not only serves as a cost and time efficient way to build SMT systems for Baltic languages but can greatly serve MT development for other smaller languages and specialized domains.

We showed how the quality of SMT, particularly the fluency of translation, can be improved by supplementing basic statistical models with morphology-based factorized models. We described the application of the ACCURAT Toolkit for extracting training data for SMT from comparable corpora which further improves the application of SMT for under-resourced languages and domains.

This helped to achieve a sufficient quality of SMT to be used for production purpose in localization and translation work. We described this SMT application scenario and its evaluation which showed a 32.9% increase in translation productivity for English-Latvian localization in the IT domain.

We outlined recent achievements in terminology – consolidation of dispersed terminology resources in the online EuroTermBank database providing a single access point to multilingual

(7) http://metashare.tilde.com/
(8) http://meta-share.lki.lt/

and multi-domain terminology. We introduced the latest work in developing the TaaS terminology services to automate the term identification, extraction and acquisition process. This will lead to further productivity increases in human translation and will improve the quality of domain specific machine translation.

These developments in Baltic language resources and technologies were enabled through participation in Pan-European research and industry cooperation projects where technologies for Baltic languages were developed in line with other European languages. Although methods elaborated in these projects were tested on Baltic and other selected languages, they can be easily adapted for other languages as well. These resources and tools are accessible on the META-SHARE language resource infrastructure.

Bibliography

EAGLES. (1996). «Preliminary recommendations on corpus typology». Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Eurostat (2011), Science, technology and innovation in Europe, 2011 edition, Publications Office of the European Union

Hutchins, John (2007) «Machine translation: a concise history», in *Computer aided translation: Theory and practice*, ed. Chan Sin Wai. Chinese University of Hong Kong, 2007.

IMF (2012), International Monetary Fund, World Economic Outlook Database, October 2012, http://www.imf.org/external/pubs/ft/weo/2012/02/weodata/index.aspx

McEnery, A., Xiao, Z. (2007). «Parallel and comparable corpora?» in *Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters*, Clevedon, UK.

Munteanu, D. and Marcu, D. (2005). «Improving Machine Translation Performance by Exploiting Non-Parallel Corpora» in *Computational Linguistics*, 31(4), pp. 477--504.

Pinnis, Mārcis, Inguna Skadiņa, and Andrejs Vasiļjevs (2013). «Domain adaptation in statistical machine translation using comparable corpora: case study for english latvian IT localisation» in *Computational Linguistics and Intelligent Text Processing*, pp. 224-235. Springer Berlin Heidelberg, 2013.

Piperidis, S. (2012). «The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions» in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pages 36-42.

Rehm, Georg, Uszkoreit, Hans (editors) (2012a). Strategic Research Agenda for Multilingual Europe 2020, Springer

Rehm, Georg, Uszkoreit, Hans (editors) (2012b). META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 2012. This series comprises 31 volumes on the following 30 European languages: Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian (available in Bokmål and Nynorsk), Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, Swedish. http://www.meta-net.eu/whitepapers.

Rirdance, Signe and Andrejs Vasiļjevs ed. (2006) Towards Consolidation of European Terminology Resources: Experience and Recommendations from EuroTermBank Project. Tilde.

Skadiņa, Inguna, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş, and Tatiana Gornostay (2010). «Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation» in *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora LREC 2010*, ELRA, pp. 6–14.

Skadina, Inguna, Andrejs Vasiljevs, Lars Borin, Koenraad De Smedt, Krister Lindén, and Eiríkur Rögnvaldsson (2011). «META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries» in *Workshop on Language Resources, Technology and Services in the Sharing Paradigm* (pp. 107–114). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Skadiņa, Inguna, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte (2012). «Latvian in the European Information Society» in *The Latvian Language in the Digital Age*, pp. 49-59. Springer Berlin Heidelberg.

Skadiņa, Inguna, Andrejs Vasiļjevs, Lars Borin, Krister Lindén, Gyri Losnegaard, Sussi Olsen, Bolette S. Pedersen, Roberts Rozis, and Koenraad De Smedt (2013). «Baltic and Nordic Parts of the European Linguistic Infrastructure» in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (pp. 195–211). Oslo: Linköping University Electronic Press.

Skadiņš, Raivis, Kārlis Goba, and Valters Šics (2010). «Improving SMT for Baltic Languages with Factored Models» in *Proceedings of the Fourth International Conference Human Language Technologies -The Baltic Perspective* (*Baltic HLT 2010)*. IOS Press,

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter (2012). «DGT-TM: A freely available Translation Memory in 22 languages» in *Proceedings of Language Resource and Evaluation Conference LREC 2012*, pp. 454-459.

Tiedemann, Jörg (2009) «News from OPUS (2009) - A Collection of Multilingual Parallel Corpora with Tools and Interfaces», in N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) in *Recent Advances in Natural Language Processing (vol V)*, 237-248, John Benjamins, Amsterdam/ Philadelphia.

Vasiljevs, Andrejs, Tatiana Gornostay, and Raivis Skadins (2010). «LetsMT!--Online Platform for Sharing Training Data and Building User Tailored Machine Translation» in *Proceedings of the Fourth International Conference Human Language Technologies -The Baltic Perspective (BalticHLT 2010)*. IOS Press, 2010.

Vasiļjevs, Andrejs, Bolette Sandford Pedersen, Koenraad De Smedt, Lars Borin, and Inguna Skadiņa (2011). «META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure» in *Proceedings of the NODALIDA 2011 workshop on Visibility and Availability of LT Resources*, Riga, Latvia, pp. 18-22.

Vasiljevs, Andrejs, Raivis Skadins, and Indra Samite (2012). «Enabling users to create their own web-based machine translation engine» in *Proceedings of the 21st international conference companion on World Wide Web (WWW 2012) Companion* (pp. 295–298). New York, New York, USA: ACM Press.

Vasiļjevs, Andrejs, and Inguna Skadiņa (2012), «Latvian Language Resources and Tools: Assessment, Description and Sharing» in *Proceedings of the Fifth International Conference Human Language Technologies -The Baltic Perspective (Baltic HLT 2012)*, pp. 265-272. 2012.