UNIVERSITY OF LATVIA
Faculty of Computing

MĀRCIS PINNIS

TERMINOLOGY INTEGRATION IN STATISTICAL MACHINE TRANSLATION

DOCTORAL THESIS
Submitted for the degree of Doctor of Philosophy in Computer Science (Dr.sc.comp.)

Field: Computer Science
Subfield: Software and Systems Engineering

Scientific supervisor:
Dr.sc.comp. **Inguna Skadiņa**

Riga 2015

The doctoral thesis was carried out at the Faculty of Computing from 2011 to 2015.

EIROPAS SOCIĀLAIS FONDS

EIROPAS SAVIENĪBA

IEGULDĪJUMS TAVĀ NĀKOTNĒ

This work has been supported by the European Social Fund within the project «Support for Doctoral Studies at University of Latvia».

The thesis contains the introduction, four chapters, conclusions, reference list, and one appendix.

Form of the thesis: dissertation in Computer Science, subfield of Software and Systems Engineering.

Supervisor:

Dr. sc. comp. Inguna Skadiņa

Institute of Mathematics and Computer Science, University of Latvia

Reviewers:

1) Guntis Bārzdiņš, Dr.sc.comp., Professor, Institute of Mathematics and Computer Science of the University of Latvia, Riga, Latvia
2) Mark Fishel, Dr.sc.comp., Associate Professor, University of Tartu, Tartu, Estonia
3) Alexander Gelbukh, Dr.sc.comp., Research Professor, Center for Computing Research of the National Polytechnic Institute, Mexico City, Mexico

The thesis will be defended at the public session of the Doctoral Committee of Computer Science, University of Latvia, at _____ on _______________, 2015, _____________________ ____.

The thesis and its summary are available at the Library of the University of Latvia, room No. 203, Raina blvd. 19, Riga.

Chairman of the Doctoral Committee /Jānis Bārzdiņš/

Secretary of the Doctoral Committee /Ruta Ikauniece/

ABSTRACT

The aim of this doctoral thesis is to research methods and develop tools that allow successfully integrating bilingual terminology into statistical machine translation systems so that the translation quality of terminology would increase and that the overall translation quality of the source text would increase. The author presents novel methods for terminology integration in SMT systems during training (through static integration) and during translation (through dynamic integration). The work focusses not only on the SMT integration techniques, but also on methods for acquisition of linguistic resources necessary for different tasks involved in workflows for terminology integration in SMT systems.

The thesis describes and evaluates methods designed and implemented by the author for: 1) monolingual term identification in SMT system training data as well as documents submitted for translation, 2) term normalisation for acquisition of canonical forms of terms from terms in different inflected forms, 3) cross-lingual term mapping in parallel and comparable corpora collected from the Web, 4) probabilistic dictionary filtering in order to acquire resources for cross-lingual term mapping, 5) development of character-based SMT transliteration systems from probabilistic dictionaries, 6) inflected form generation for terms through rule-based morphological synthesis or monolingual corpus look-up, and other methods involved in the workflows for static and dynamic terminology integration in SMT systems.

The terminology integration methods have been evaluated using the *Moses* SMT system and the *LetsMT* platform. The evaluation efforts show that the methods for monolingual term identification and cross-lingual term mapping allow achieving state-of-the-art performance, which has been also validated by third party (independent) evaluation efforts. The static terminology integration methods allow achieving a cumulative SMT quality improvement by up to 28.1% (or 3.56 absolute BLEU points) over an initial baseline system for the English-Latvian language pair. However, the most impressive achievement of the author's work is the dynamic terminology integration method in SMT systems using a source text pre-processing workflow. In almost all experiments performed in the scope of the thesis the methods allowed achieving SMT quality improvements. Automatic evaluation for four investigated language pairs in the automotive domain shows SMT quality improvements by up to 26.9% (or 3.41 absolute BLEU points) over baseline systems. Manual comparative evaluation performed for seven language pairs in the information technology domain shows that the proportion of correctly translated terms increases for all language pairs by up to +52.6%.

ACKNOWLEDGEMENTS

First of all, I have to express my gratitude to my parents and my wife. Without their moral support I would not be able to achieve as much as I have. Their support (even keeping fingers crossed when necessary) has helped me motivating myself to strive for better results, to not lose hope when experiments failed to produce positive results, and to finalise the research in order to create this thesis.

Then, I would like to thank my supervisor, Dr.sc.comp. Inguna Skadiņa, for guiding me through the research process, constructively criticising my research decisions, and providing important feedback for the thesis and the involved publications.

Special thanks have to be expressed to the Tilde company and my colleagues in Tilde: 1) to Andrejs Vasiļjevs, Raivis Skadiņš and Tatiana Gornostay for giving me the possibility to participate in the European research projects and for giving advice from their professional experience and also constructively criticising wrong decisions made along the research and development process of the methods described in the thesis 2) to the colleagues of the Tilde's Localisation Department for their contribution to annotation and evaluation efforts.

I would like to acknowledge also the colleagues (research partners) within the research projects ACCURAT and TaaS (especially Ahmet Aker and Monica Lestari Paramita from the University of Sheffield, Nikola Ljubešić from the Faculty of Humanities and Social Sciences of the University of Zagreb, Dan Ştefănescu from the Research Institute for Artificial Intelligence of the Romanian Academy, and Fangzhong Su from the Centre for Translation Studies of the University of Leeds) who have developed methods and tools used in my research (e.g., *Terminology Aligner*, *DictMetric*, *CollTerm*, *Focussed Monolingual Crawler*, etc.), participated in evaluation efforts and performed independent evaluation efforts involving tools developed by me, and have kept a challenging atmosphere in the air that has motivated me to develop more innovative methods for the different natural language processing tasks described in the thesis.

The research and development of the methods described in this thesis has been also supported by two European research projects within the European Union Seventh Framework Programme (FP7/2007-2013): ACCURAT (grant agreement No. 248347) and TaaS (grant agreement No. 296312). A significant motivation for reaching for higher quality and more significant achievements has been gained because of the support from the European Social Fund within the project «Support for Doctoral Studies at University of Latvia». Without the support from these projects, the goals of the thesis would be very difficult (if not impossible) to achieve.

CONTENTS

INTRODUCTION

Machine translation (MT) is "*the use of computers to automate translation from one language to another*" (Jurafsky & Martin, 2009). Machine translation solutions have many different applications. Three of the most important applications (in the author's opinion) are:

1) To provide access to information written in a language that is unknown to the consumer of the information. For instance, MT systems can provide access to information on the Web (news, blog articles, product descriptions, product or service reviews, etc.) that is written in many different languages. Popular publicly available MT services, such as, the Google Translate[1] and the Bing Translator[2] are widely used for automated translation of such information.
2) To lower the language barriers that may not allow people effectively communicating (or communicating at all) between each other if a common language is not known. For instance, for a tourist who gets lost in a foreign country SMT can provide a possibility to communicate with local people.
3) To increase productivity of professional translators. In recent years, due to the rapid development of MT technologies, automated translation services have been also introduced in professional translation workflows. Language service providers as well as leading software developers (for instance, Flournoy & Duran, 2009; Schmidtke, 2008; Skadiņš et al., 2014, and many others) have shown that SMT integration into translation workflows allows to significantly improve the translation productivity of translators, which in turn may result in cost savings and higher competitiveness of the translator and the localisation service provider in the localisation industry.

The first two application areas do not have very high quality requirements as the task of MT is to allow understanding the contents rather than to provide perfect translations. Whereas the third application area requires MT systems to be able to provide precise translations of very high quality, otherwise, the usage of MT services may not be economically justifiable. However, all three scenarios require that the translations would contain correct terminology, because incorrectly used terms may not even allow understanding the correct meaning of the translated text.

[1] Google Translate is an SMT service available online at: https://translate.google.com.
[2] Bing Translator is an SMT service, which is available online at: http://www.bing.com/translator/.

From the theoretical point of view, SMT systems in the translation process try to solve the following problem:

$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|S) = \underset{T}{\operatorname{argmax}} \frac{P(S|T)P(T)}{\underbrace{P(S)}_{\text{Constant}}} = \underset{T}{\operatorname{argmax}}\, \overbrace{P(S|T)}^{\text{Translation model}}\overbrace{P(T)}^{\text{Language model}} \tag{1}$$

That is, for each source sentence S they try to find the target sentence $\hat{T}$, which is the most likely translation of the source sentence S. Most commonly, the problem (solving the $\underset{T}{\operatorname{argmax}}(T|S)$) is decomposed using the Bayes theorem into a noisy-channel model (Brown et al., 1993) ($\underset{T}{\operatorname{argmax}}\, P(S|T)P(T)$), which allows solving two separate problems: 1) we want to identify target language sentences (hypotheses), which are possible translation equivalents of the source sentence, and 2) we want to make sure that the translation hypothesis that we generate are correct sentences in the target language. The first problem we solve with a translation model that is trained on a large parallel sentence corpus and the second problem we solve with a language model that is trained using a large monolingual sentence corpus (often even much larger than the parallel corpus) in the target language.

Current SMT phrase-based models, including the popular *Moses* SMT system (Koehn et al., 2007), do not explicitly handle terminology translation. Although domain adaptation can be performed using additional in-domain training data (Koehn and Schroeder, 2007), such an approach is very resource demanding as it requires gathering of the resources (parallel and monolingual corpora) for each individual domain and for smaller projects or for languages with limited resources this is not feasible. This makes terminology integration with the standard approaches expensive (in terms of time) and for less resourced languages in many cases also impossible (due to lack of parallel or monolingual in-domain corpora). Therefore, this work addresses an unsolved problem in SMT of how to integrate (bilingual) terminology into SMT systems so that the terminology translation quality would be increased and the SMT systems would be easily adaptable to different domains with the help of bilingual terminology.

Relevance of the Research Problem

In professional translation services the quality of translations is evaluated using quality assessment (QA) forms that are based on a specific QA model. For instance, the QA form for translations in the Baltic localisation service provider Tilde[3] is based on the QA Model[4] of the Localization Industry Standards Association (LISA). The QA form requires to identify different types of error classes in translated texts, which are grouped in four main categories: accuracy (e.g., whether the translation contains omissions or unnecessary additions, whether the translation is comprehensible, etc.), language quality (e.g., grammar, punctuation, and spelling mistakes), style (e.g., the word order, whether the translator followed style guidelines, etc.), and terminology (consistency and adherence to a pre-defined collection).

Because terminology is one of key elements that is assessed when performing manual quality evaluation of translations in professional translation and localisation services, it is important that any automated translation solution (if it is to be introduced in professional translation services) provides support for correct handling of terminology by assuring the two main quality requirements:

1. **Terminology has to be used correctly** (i.e., if a term collection is provided, the translations for terms have to be selected only from the provided collection).
2. **Terminology has to be used consistently** (i.e., if a term appears multiple times in a document, only one translation should be used for the translation of the term).

For MT systems, the first requirement is difficult to achieve, because the context (or more precisely, the lack of enough context) may not always allow identifying the correct translations of terms. The second requirement challenges statistical[5] MT (SMT) systems more than rule-based MT systems as the statistics of large amounts of data are difficult to control if not constrained by means of, e.g., bilingual term collections or translation model or language model domain adaptation techniques. If SMT systems are not developed and "*taught*" to understand terminology, ambiguous or unknown contexts in the parallel training data may result in the selection of an incorrect translation hypothesis because of higher contextual likelihood. Therefore, methods are needed to integrate domain-specific term collections into SMT systems in order to perform domain adaptation and produce better quality translations. This necessity has been the driving force of this work with the main goal to develop methods and tools that allow to perform successful integration of terminology into SMT systems in order to improve the quality of terminology translation and at the same time also improving the overall translation quality of the source text.

[3] More information about Tilde can be found on the company's Web site at: http://www.tilde.lv.

[4] A description of the LISA QA model in comparison with other QA models is given by Mateo (2014).

[5] The difference between a rule-based MT system and an SMT system is that for a rule-based system the system's developer has to write many (often thousands) different rules (direct translation examples, morpho-syntactic transfer rules, etc.), however in an SMT system translations are automatically learned (inferred) from a large parallel corpus without the need for hand-crafted rules.

The main issues of terminology translation without explicit support for terminology integration within SMT systems are as follows:

- **Terms may be translated using incorrect translation equivalents**. That is, the translation equivalents may be: 1) from a different domain, 2) from obsolete variants of terms, 3) in abbreviated or non-abbreviated forms (contrary to the terms' form in the source language), 4) used by a client's competitor, etc. For example, the term „*tablet*" is ambiguous – it can refer to a popular consumer electronics product (a tablet computer), a number of sheets of paper fastened together along one edge (according to WordNet 3.1[6]), a pill used in medicine, and others. Because SMT system translation and language models are built by analysing word and phrase frequencies in parallel and monolingual corpora, the correct term translations may occur less frequently than different domain term translations in specific contexts. This may result in selection of the incorrect translations due to higher probabilities assigned by the SMT system's translation and language models. The sentence "*Has anyone seen my tablet?*" translated with three popular English-Latvian SMT services (Google Translate, Bing Translator, and Tilde Translator[7]) perfectly illustrates this issue (see Table 1). If the correct translation would be "*planšete*" (translated in English as a "*tablet [computer]*") then all SMT services would have failed to translate the sentence correctly.

Table 1. Translations of the English sentence "*Has anyone seen my **tablet**?*" into Latvian with three publicly available SMT services[8]

SMT service	Translation
Google Translate	*Vai kāds ir redzējis manu **tableti**?*
Bing Translator	*Vai kāds ir redzējis manu **tablet***
Tilde Translator	*Vai kāds ir redzējis manu **tablete**?*

- **Terms may be missing in the SMT system's models**, which means that out-of-vocabulary terms would not be translated. Table 2 shows the term "*vārsta siltumatstarpe*" (translated in English as "*valve clearance*") in a context translated with three popular SMT services. The word "*siltumatstarpe*" has not been recognised by any of the SMT services, therefore, it has been passed through to the English translations without translation.

[6] More information on WordNet can be found online at: http://wordnet.princeton.edu/.
[7] Tilde Translator is an SMT service available online at: http://translate.tilde.com/.
[8] Note that the translations may have changed already since online SMT services are dynamically improved by the developers.

Table 2. Translations of the Latvian sentence "***Vārstu siltumatstarpe** ir nepieciešama vārstu atvēršanai un aizvēršanai pareizā laikā*." into English with three publicly accessible SMT services

SMT service	Translation
Google Translate	***Siltumatstarpe valve** is required to open and close the valve at the right time.*
Bing Translator	***Valve siltumatstarp** is required for opening and closing of the valve in the correct time.*
Tilde Translator	***Valve siltumatstarpe** requires the valve opening and closing time.*

- **Multi-word terms may be split into several parts during translation**. This problem may occur because SMT systems (including the *Moses* system) use reordering (also known as distortion) models that allow to reorder translated phrases in the target language. Different languages may have different word ordering paradigms, e.g., English has a subject-verb-object (SVO) word ordering (Lehmann, 1978), Dutch has a subject-object-verb (SOV) word ordering (Koster, 1975), Latvian has a relatively free word ordering, however it is considered that most commonly it has SVO word ordering (Lokmane, 2010). Although, in general, the introduction of reordering models has shown to improve SMT quality (Vogel, 2003), as shown in Table 3 it may also cause issues for terminology translation. The example shows that the term "*attribute filter*" from the information technology (IT) domain has not been translated as a non-breakable phrase in one of the SMT systems.

Table 3. Translations of the English sentence "*Using the new **attribute filter** functionality.*" into Latvian with three publicly available SMT services[9] (the correct translation is underlined)

SMT service	Translation
Google Translate	*Izmantojot jauno **atribūts filtra** funkcijas.*
Bing Translator	*Jaunu **atribūtu filtru** funkcionalitātes izmantošana.*
Tilde Translator	*Izmantojot jauno **filtrēšanas** funkcionalitātes **atribūts**.*

- **Multi-word terms may also be translated by breaking morpho-syntactic agreements between constituents of the terms**. It is very important to model morpho-syntactic agreements when translating into morphologically rich languages (e.g., Latvian, Estonian, Czech, etc.). When translating into Latvian, for instance, adjectives need to be generated in same gender, number, and case as the head noun in the immediate noun phrase the adjectives belong to. Table 4 shows an example where two of the SMT services (except the Bing Translator) failed to model the agreement in the noun phrase "*modern home*". The first service failed to create agreement in number and the third service failed to create agreement in case. If terminology integration would be supported, the agreement between the term's constituents could be explicitly modelled, thereby solving such mistakes. However, the translation of the second

[9] Note that the translations might have changed already since online SMT services are dynamically improved by the developers.

service is also not completely correct, because the source noun phrase was given in a singular form.

Table 4. Translations of the English sentence "*A* ***modern home*** *in a valley.*" into Latvian with three publicly accessible SMT services.

SMT service	Translation
Google Translate	***Moderna mājas*** *ielejā.*
Bing Translator	***Mūsdienu mājas*** *ielejā.*
Tilde Translator	***Mūsdienīgs mājas*** *ielejā.*

- **When localising software or translating documents for specific clients, the clients may request the usage of their specific terminology**. SMT systems rely on statistics when translating terms and not on pre-defined term dictionaries. Therefore, the translations may be wrong and also inconsistent (depending on context, different translations may be selected for one term). For instance, if we have a term collection from a client that specifies that the term "*web service*" has to be translated as "*tīmekļa pakalpe*", then the SMT system should be able to support such a user request. However, current SMT solutions do not offer this kind of a functionality. Table 5 shows that none of the publicly accessible SMT services translates the term as required by the client.

Table 5. Translations of the English sentence "*The web service is operational.*" into Latvian with three publicly accessible SMT services.

SMT service	Translation
Google Translate	***Web pakalpojums*** *darbojas.*
Bing Translator	***Web pakalpojums*** *darbojas.*
Tilde Translator	***Tīmekļa pakalpojums*** *darbojas.*

Research Methods

The following are the main contemporary research methods used by the author:

- **Scientific literature review -** to identify the current state-of-the-art methods related to the author's research topics (term identification, cross-lingual term mapping, terminology integration in SMT, etc.) and to identify unsolved gaps and deficiencies of related research, the author analysed publications from the main natural language processing conferences, scientific projects, and journals that cover topics investigated by the author.
- **Implementation of algorithms** – the tools developed by the author have been designed and implemented in an iterative manner, which allows analysing different types of algorithms and improving the author's methods.

- **Controlled experiments** – in order to empirically prove that the author's methods perform better than baseline methods and methods designed in relateds work, the author performed numerous controlled experiments (Wohlin et al., 2003; Wasterbrook et al., 2008) for cross-lingual term mapping and terminology integration methods in SMT.
- **Automatic evaluation** – standard automatic sequence labelling and information extraction (for term identification and cross-lingual term mapping), and machine translation (for terminology integration in SMT and character-based SMT transliteration) evaluation methods (Jurafsky & Martin, 2009) have been used to evaluate the different methods developed in the scope of the thesis.
- **Manual evaluation** – where applicable (and necessary) manual evaluation experiments (for instance, comparative human evaluation for the evaluation of terminology integration in SMT) were performed to validate automatic evaluation results.
- **Error analysis** – where necessary, the author has performed manual error analysis for the developed methods in order to identify possible areas of future improvements.

Object of Research

The object of research of this work are **methods and algorithms for terminology integration in statistical machine translation**. The lack of support for terminology integration in current SMT models affect (as explained in the previous section) terminology translation quality and consistency. Therefore, in this thesis the author proposes novel and effective methods for terminology integration in SMT systems.

Research Hypotheses

Taking into account the limitations of current SMT systems with respect to terminology translation quality and consistency, the author states the following hypothesis: **terminology translation quality as well as text translation quality in SMT systems can be improved by performing static and dynamic terminology integration in SMT systems.** As manual creation of term collections is an expensive and time consuming process, automatic (or at least semi-automatic) methods for term collection creation are necessary. Therefore, to the author states the following second hypothesis: **in situations when authoritative term collections are not available, automatic term identification in comparable corpora and cross-lingual term**

mapping are effective methods to acquire bilingual term collections for the integration in SMT systems. Both research hypotheses in the thesis are proved using experimental methods.

Aim and Objectives

To prove the research hypothesis, effective methods are needed to provide means for terminology integration in SMT systems. Therefore, the aim of this doctoral thesis has been to research methods and develop tools that allow successfully integrating terminology into SMT systems so that the translation quality of terminology and the overall translation quality of the source text would increase. To reach the aim, the research and development activities were split into following objectives:

- To research methods and develop tools for static terminology integration in SMT systems that allow: 1) to adapt SMT systems to the required domain with the help of in-domain terminology, and 2) to increase translation quality.
- To research methods and develop tool for dynamic terminology integration in SMT systems during the translation phase that: 1) do not require re-training of SMT systems, and 2) allow to increase translation quality.
- To research and develop methods for term identification in:
    - SMT system training data (for static terminology integration in SMT systems).
    - Text documents intended for translation (for dynamic terminology integration in SMT systems).
    - Text documents intended for monolingual term candidate extraction with a goal to create monolingual or bilingual term collections usable for integration in SMT systems.
- To research and develop methods for cross-lingual term mapping. In situations where in-domain terminology is not available, however, there exists at least some parallel data (two/three thousand or even less sentence pairs) or a comparable corpus, cross-lingual term mapping methods can be used to automatically create term collections.
- To research and develop methods that address the previous objectives in particular for languages with complex morphologies and little (or no) parallel resources in domains in which terminology integration has to be performed.
- To evaluate the developed methods for the English-Latvian language pair and where applicable also other European languages in order to prove that the methods are general and language independent (not considering language specific resources that the methods may require).

The methods researched by the author are in general applicable to any phrase-based SMT platform. However, to evaluate the methods, the author focusses on the *Moses* SMT system (Koehn et al., 2007) and the *LetsMT* platform (Vasiļjevs et al., 2012).

Scientific Novelty

In the scope of the thesis, the author has researched and developed the following innovative methods:

- A linguistically, statistically, and reference corpus motivated term identification method for semi-automatic creation of term collections. The method has been implemented in the tool *Tilde's Wrapper System for CollTerm*, which allows performing monolingual term identification in translatable documents or documents used for automatic or semi-automatic term collection creation.
- A novel method for term identification in SMT system training data and documents submitted for translation: the *Fast Term Identification*. The method has been specifically designed for the purpose of terminology integration in SMT and allows to perform term identification in parallel and monolingual corpora (used in static integration methods) and the source text (documents, translation segments, sentences, etc.) that is sent to the SMT system for translation.
- A context independent cross-lingual term mapping method. The method uses probabilistic dictionaries and character-based SMT transliteration systems when performing term mapping. It allows mapping multi-word terms and terms with different number of tokens in the source and target languages – two term mapping scenarios that have not been sufficiently addressed by previous research. The method has been implemented in the tool *MPAligner*.
- A novel method for probabilistic dictionary[10] filtering using character-based SMT transliteration systems. As far as the author knows, transliteration systems have not been used in related research to filter probabilistic dictionaries.
- A novel method for SMT-based transliteration system creation using transliteration dictionaries that have been automatically extracted from probabilistic dictionaries using a bootstrapping method (first of a kind).
- A novel method for static terminology integration in SMT systems. The method proposes to transform SMT system phrase tables into term-aware phrase tables by

[10] A probabilistic dictionary is a statistical resource acquired by performing automated word alignment in parallel corpora. Popular word alignment tools are, for instance, *Giza++* (Och and Ney, 2003), *FastAlign* (Dyer et al., 2013), *Anymalign* (Lardilleux et al., 2012), and many other tools.

identifying bilingual terminology in all phrase pairs found in the phrase tables. An important difference from other methods is the method's ability to identify terms in different inflected forms.

- A novel multi-dimensional method for dynamic terminology integration in SMT systems. The method proposes a source text pre-processing workflow that can be directly integrated into the Moses SMT system. The workflow includes the term identification methods and a novel component for rule-based inflected form generation for multi-word terms.

Practical Significance of Work

The author has created a set of tools and resources that are necessary in various tasks related to terminology integration in SMT. The tools and resources can be beneficial also in other research areas of natural language processing.

The most important tools are:

- Tools for term identification:
  - The *Tilde's Wrapper System for CollTerm* (TWSC) for linguistically and statistically motivated identification of terms in documents. *TWSC* can be acquired as part of the *ACCURAT Toolkit*[11]. It is used in the *Terminology as a Service*[12] (*TaaS*) platform (Pinnis et al., 2013) for monolingual term identification when creating monolingual and bilingual term collections (see section 2.2).
  - The *Fast Term Identification* tool for processing of large data sets (e.g., SMT training data) using existing term collections (see section 2.4). The tool is integrated in the *LetsMT*[13] platform (Vasijevs et al., 2012).
  - The *Pattern-Based Term Identification* tool for linguistically motivated term identification in the source text that is submitted for translation (see section 2.3). The tool is used in the *TaaS* platform for term candidate identification in parallel and comparable corpora. The tool has been used to extract term candidates that are integrated in the *Statistical Data Base* (SDB) of the *TaaS* platform (TaaS, 2014a).
- A term normalisation tool for Latvian (see section 2.5). The tool is used in the *TaaS* platform for term candidate normalisation after monolingual term identification.
- A cross-lingual term mapper (the *MPAligner*[14]) that supports term mapping for 25 European languages (i.e., all official languages of the European Union and Russian;

[11] The *ACCURAT Toolkit* can be acquired online at: http://accurat-project.eu/.
[12] The *TaaS* platform is available online at: https://term.tilde.com/.
[13] The *LetsMT* platform is available online at: https://www.letsmt.eu.
[14] The *MPAligner* can be acquired online at: https://github.com/pmarcis/mp-aligner.

see section 3.2). *MPAligner* is used in the *TaaS* platform to perform bilingual term extraction from term-tagged parallel and comparable corpora. It has been also used to extract bilingual terminology for the largest (as far as the author knows) statistical resource of bilingual terminology – the *TaaS* platform's *SDB.*

- A toolkit for terminology integration in SMT systems that is able to perform both static integration (see section 4) and dynamic terminology integration (see section 5) tasks. The toolkit is integrated in the *LetsMT* platform and it can be used by SMT system developers to integrate terminology in SMT systems.

The most important linguistic resources are:

- A multilingual transliteration dictionary[15] for 24 European languages (see section 3.4) that consists of 1,246,908 transliteration pairs. As far as the author knows, this is the first multilingual transliteration dictionary that has been publically released.
- Bilingual terminology automatically extracted from *Wikipedia* and other comparable and parallel data sources using the *MPAligner*. The resource contains over twenty million unique inflected form pairs of terms distributed over 45 subject fields and 26 language pairs (see section 3.5). The resource is the main source of bilingual term pair candidates of the *TaaS platform's SDB.* As far as the author knows, this is the largest currently available resource of automatically extracted bilingual terminology.

Main Results

The scientific and practical results of the thesis have been already described in the sections "Scientific Work" and "Practical Significance of Work". However, this section names three results that the author thinks are the main results of the work:

1) The author's designed and developed **toolkit for static and dynamic terminology integration in SMT systems.** The toolkit has shown to increase overall translation and term translation quality in both automatic and manual evaluation experiments.
2) The author's designed and developed **tool for linguistically, statistically, and reference corpora motivated term identification - *Tilde's Wrapper System* for *CollTerm***.
3) The author's designed and developed **tool for context-independent cross-lingual term mapping – *MPAligner***. *MPAligner* in combination with *TWSC* have been used to create the largest resource of automatically extracted bilingual terminology.

[15] The multilingual transliteration dictionary can be acquired online at: https://github.com/pmarcis/dict-filtering.

Approbation and Publication of the Author’s Work

The author’s work (relevant to the thesis) has been published in **17 publications** – 10 publications the thesis is based on and 7 publications relevant to the topics discussed in the thesis.

The thesis is based on the author’s contributions to the following 10 publications:

- 6 publications in peer-reviewed conference proceedings recognised by the Latvian Council of Science:
  - Aker, A., Pinnis, M., Paramita, M. L., & Gaizauskas, R. (2014b). Bilingual Dictionaries for All EU Languages. *In Proceedings of LREC 2014* (pp. 2839–2845). Reykjavik, Iceland. Indexed in Web of Science. The author’s contributions to the paper are: 1) the transliteration-based dictionary filtering method, and 2) the quantitative analysis of evaluation results (the total contribution is approximately 20%).
  - Pinnis, M. (2013). Context Independent Term Mapper for European Languages. *In Proceedings of RANLP 2013* (pp. 562–570). Hissar, Bulgaria. Indexed in Scopus. The author’s contribution to the paper is 100%.
  - Pinnis, M. (2014). Bootstrapping of a Multilingual Transliteration Dictionary for European Languages. *In Proceedings of Baltic HLT 2014.* Kaunas, Lithuania: IOS Press. Indexed in Web of Science. The author’s contribution to the paper is 100%.
  - Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. *In Proceedings of TKE 2012* (pp. 193–208). Madrid. Indexed in Scopus. The author’s main contributions to the paper are: 1) the role of the leading author, 2) the section about term tagging (excluding the evaluation subsection for Croatian), 3) the section about the real world scenario, 4) example figures for Latvian and Lithuanian, and 5) conclusions of the paper (the total contribution is approximately 40%).
  - Pinnis, M., & 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). Tartu, Estonia, Estonia: IOS Press. Indexed in Scopus. The author’s main contributions to the paper are: 1) the role of the leading author, and 2) all sections except the introduction (the total contribution is approximately 90%).
  - Vasiļjevs, A., Pinnis, M., & Gornostay, T. (2014). Service Model for Semi-Automatic Generation of Multilingual Terminology Resources. *In Proceedings of TKE 2014 (pp. 67–76).* Berlin, Germany. Indexed in Scopus. The author’s main

contributions to the paper are: 1) the section about term candidate identification, and 2) the section about translation equivalent retrieval from the Web (the total contribution is approximately 60%).

- 2 publications in other peer-reviewed conference proceedings:
  - Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation. *In Proceedings of EAMT 2015* (pp. 89–96). Antalya, Turkey. The author’s contribution to the paper is 100%.
  - Skadiņš, R., Pinnis, M., Gornostay, T., & Vasiļjevs, A. (2013). Application of Online Terminology Services in Statistical Machine Translation. *In Proceedings of the XIV Machine Translation Summit* (pp. 281–286). Nice, France. The author’s main contributions to the paper are: 1) the section about related work, and 2) the section about the proposed solution (the total contribution is approximately 60%).
- 2 other publications:
  - Pinnis, M., Skadiņš, R., & Vasiļjevs, A. (2014). Real-world challenges in application of MT for localization: The Baltic case. *In Proceedings of AMTA 2014, vol. 2: MT Users* (pp. 66–79). Vancouver, BC Canada. The author’s main contributions to the paper are: 1) the section about terminology translation, and 2) the section about translation productivity analysis (the total contribution is approximately 40%).
  - Vasiļjevs, A., Kalniņš, R., Pinnis, M., & Skadiņš, R. (2014). Machine Translation for e-Government - the Baltic Case. *In Proceedings of AMTA 2014, vol. 2: MT Users* (pp. 181–193). Vancouver, BC Canada. The author’s main contribution to the paper is the section about term translation (the total contribution is approximately 20%).

The 7 publications relevant to the topics discussed in the thesis are:

- 4 publications in peer-reviewed conference proceedings recognised by the Latvian Council of Science:
  - Pinnis, M. (2012). Latvian and Lithuanian Named Entity Recognition with TildeNER. In *Proceedings of LREC 2012* (pp. 1258–1265). Istanbul, Turkey. Indexed in *Web of Science*. The author’s contribution to the paper is 100%.
  - Pinnis, M., & Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In *Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology* (pp. 14–22). Zurich, Switzerland: Springer Berlin Heidelberg. Indexed in *Scopus*. The author’s main contributions to the paper are: 1) the role of the leading author, 2) the section about

the morphological tagset, 3) the section about training data, 4) the section about feature selection, and 5) the section about results and error analysis (the total contribution is approximately 60%).

  - Pinnis, M., Skadiņa, I., & Vasiļjevs, A. (2013). Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English Latvian IT Localisation. In *Proceedings of CICLING 2013* (pp. 224–235). Samos, Greece: Springer Berlin Heidelberg. Indexed in *Scopus*. The paper received the *Best Student Paper Award* at the conference. The author's main contributions to the paper are: 1) the role of the leading author, 2) the section about collecting and processing comparable corpora, 3) the section about SMT systems, 4) the section about automatic and comparative evaluation, and 5) the results subsection of the evaluation in localisation (the total contribution is approximately 60%).
  - Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiș, D., Verlic, 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 LREC 2012* (pp. 438–445). Istanbul, Turkey. Indexed in *Web of Science*. The author's main contribution to the paper is the section about named entity and term extraction (the total contribution is approximately 8%).
- 3 publications in other peer-reviewed conference proceedings:
  - Pinnis, M., Gornostay, T., Skadiņš, R., & Vasiļjevs, A. (2013). Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. In *Proceedings of eLex 2013* (pp. 122–131). Tallinn, Estonia. The author's main contributions to the paper are: 1) the role of the leading author, 2) the section about the workflow for the creation of a bilingual term collection, and 3) the section about computer-assisted translation tool and machine translation system interfaces (the total contribution is approximately 30%).
  - Pinnis, M., Ion, R., Ştefãnescu, 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). South Korea. The author's main contributions to the paper are: 1) the role of the leading author, 2) the section about the overview of the workflows, 3) the section about named entity extraction and mapping, and 4) the evaluation results for English-Latvian (the total contribution is approximately 25%).

- Skadiņš, R., Pinnis, M., Vasiļjevs, A., Skadiņa, I., & Hudík, T. (2014). Application of Machine Translation in Localization into Low-resourced Languages. In *Proceedings of EAMT 2014* (pp. 209–216). The author's main contributions to the paper are: 1) the section about the result analysis of the first experiment, and 2) the section about the second experiment (the total contribution is approximately 30%).

The research topics covered by the thesis as well as research related to the thesis has been presented in 10 scientific conferences and 3 workshops:

- *The 18th Annual Conference of the European Association for Machine Translation (EAMT 2015),* Antalya, Turkey – poster presentation of the paper "Dynamic Terminology Integration Methods in Statistical Machine Translation", May, 2015.
- *The 6th International Conference Baltic HLT 2014,* Kaunas, Lithuania – oral presentation of the paper "Bootstrapping of a Multilingual Transliteration Dictionary for European Languages", September, 2014.
- *The 9th International Conference on Language Resources and Evaluation,* Reykjavik, Iceland – *poster* presentation of the paper "Bilingual Dictionaries for All EU Languages", May, 2014.
- *The 9th International Conference on Recent Advances in Natural Language Processing,* Hisarya, Bulgaria – oral presentation of the paper "Context Independent Term Mapper for European Languages", September, 2013.
- *The 14th International Conference on Intelligent Text Processing and Computational Linguistics,* Samos, Greece – oral presentation of the paper "Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English Latvian IT Localisation", March, 2013.
- *The W3C Workshop – Making the Multilingual Web Work,* Rome, Italy – presentation of the posters "The Next Step in Translation Automation: Online Terminology Services for Human and Machine Translation" and "ITS 2.0 Enriched Terminology Annotation Use Case", March, 2013.
- *The 3rd International Conference on Terminology – Current Trends in Terminology Theory and Practice*, Riga, Latvia – oral presentation "Improving Machine Translation with Terminology", October, 2012.
- *Human Language Technologies – The Baltic Perspective,* Tartu, Estonia – oral presentation of the paper „MT Adaptation for Under-Resourced Domains – What Works and What Not", October, 2012.
- *The 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, South Korea – system demonstration and poster presentation of the paper „ACCURAT

Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora", July, 2012.

- *The Second Workshop on Creation, Harmonization and Application of Terminology Resources*, Madrid, Spain – oral presentation „Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora", June, 2012.
- *The 10th Conference on Terminology and Knowledge Engineering*, Madrid, Spain – oral presentation of the paper „Term extraction, tagging, and mapping tools for under-resourced languages", June, 2012.
- *The 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey – poster presentation of the paper „Latvian and Lithuanian Named Entity Recognition with TildeNER", May, 2012.
- *The 2nd Workshop on Systems and Frameworks for Computational Morphology*, Zurich, Switzerland – oral presentation of the paper „Maximum Entropy Model for Disambiguation of Rich Morphological Tags", August, 2011.

The most important research projects the work has been approbated in are:

1) Project *Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation* (ACCURAT) - funded by the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 248347 (2010-2012).
2) Project *Terminology as a Service* (TaaS) - funded by the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n° 296312 (2012-2014).
3) Project "*2.6. Multilingual Machine Translation*" – funded by the ICT Competence Centre (www.itkc.lv), contract No. L-KC-11-0003.

The terminology integration methods have been integrated in the LetsMT platform. The public administration SMT solution hugo.lv[16] that is based on the LetsMT platform uses author's methods for terminology integration in SMT. The semi-automatic term collection creation methods have been integrated in the TaaS platform and allow its users to create bilingual term collections from users' documents.

Outline

In order to reach the aim and objectives of the thesis and to prove the hypothesis, the thesis has been organised in sections that describe different steps (or processes) defined in a

[16] hugo.lv hosts also the SMT systems of the Latvia's presidency of the Council of the European Union (Translate 2015); it is accessible online at: https://hugo.lv/.

workflow for terminology integration in SMT systems. The workflow is described in Section 1. The different steps of the workflow are described in the following sections:

1) Section 2 describes methods that the author has designed for automated term identification. The section is based on the publications Pinnis et al. (2012), Vasiļjevs et al. (2014b), and TaaS (2014b).
2) Section 3 presents the state-of-the-art context independent cross-lingual term mapping tool *MPAligner* and the methods for probabilistic dictionary filtering and SMT-based transliteration system creation. The section is based on the publications Aker et al. (2014b), Pinnis (2013), and Pinnis (2014).
3) Section 4 describes research efforts carried out by the author on static terminology integration in SMT systems. The section is based on the publications Pinnis & Skadiņš (2012), Pinnis et al. (2014), Skadiņš et al. (2013), and TaaS (2014b).
4) Section 5 presents a novel multi-dimensional method designed by the author for dynamic terminology integration in SMT systems during translation. The section is based on the publications Pinnis (2015), Pinnis & Skadiņš (2012), Skadiņš et al. (2013), TaaS (2014b), and Vasiļjevs et al. (2014a).

1. WORKFLOW FOR TERMINOLOGY INTEGRATION IN SMT SYSTEMS

In order to perform terminology integration into SMT systems, a term collection is required. A term collection has to be created or acquired from existing sources (e.g., term banks, pre-existing electronic term collections, etc.). The creation can be a completely manual process where a terminologist creates bilingual term pairs using, for instance, a spreadsheet. However, there exist also semi-automated and fully automated methods for bilingual term collection creation. Because a bilingual term collection is a key resource in the author's research, the thesis presents a workflow that allows acquiring bilingual term collections in a semi-automated (or even fully automated) manner from comparable corpora collected from the Web, and integrating the acquired term collections in SMT systems. The conceptual design of the workflow is depicted in Figure 1.

The workflow consists of two main steps: 1) term collection acquisition, and 2) terminology integration in SMT systems. The first step can be skipped completely if a term collection in a required domain is already available. However, if such a term collection is not available, the first step specifies two tasks:

1) Using a comparable corpus, terms have to be identified in the comparable corpus. The identified terms have to be also normalised (the canonical forms have to be generated for terms in different inflected forms) to reduce data redundancy.

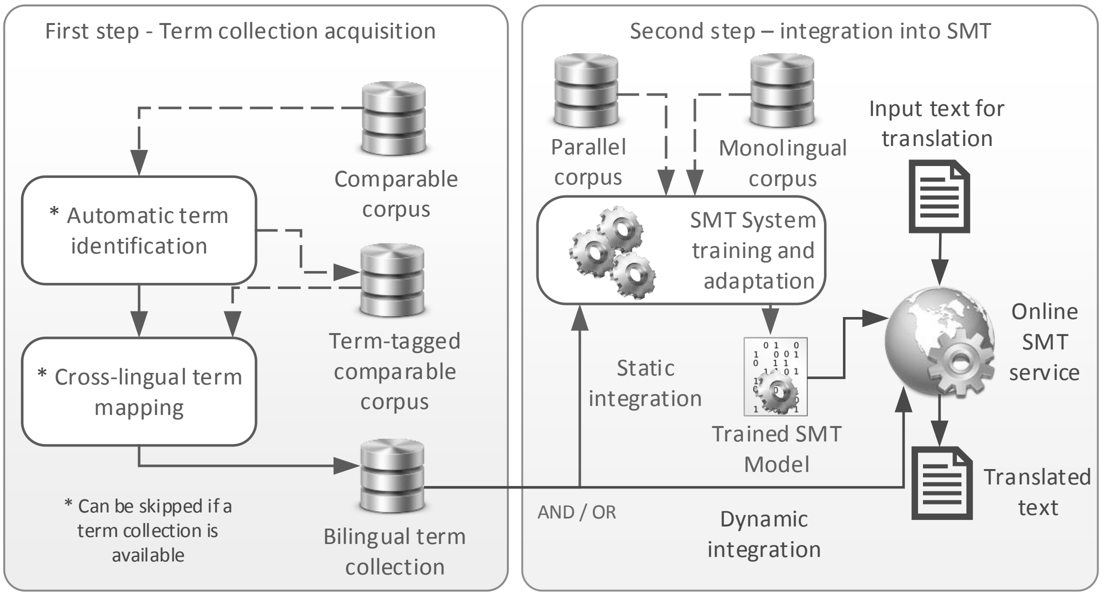

First step - Term collection acquisition
Comparable corpus
* Automatic term identification
Term-tagged comparable corpus
* Cross-lingual term mapping
* Can be skipped if a term collection is available
Bilingual term collection
Second step – integration into SMT
Parallel corpus
Monolingual corpus
Input text for translation
SMT System training and adaptation
Static integration
Trained SMT Model
Online SMT service
Translated text
AND / OR
Dynamic integration

Figure 1. The conceptual design of the workflow for terminology integration in SMT systems as it is presented in this thesis

2) When performing fully automated bilingual term extraction from parallel or comparable corpora, monolingual terms in the source language after term

identification have to be somehow mapped (or aligned) with their translations in the target language. This task is performed by automatic cross-lingual term mapping tools. These tools analyse the source and target terms identified in corpora and search for term pairs that can be considered to be reciprocal translation equivalents. After cross-lingual term mapping, the bilingual term collection can be used in SMT integration experiments as is or it can be manually filtered by a translator to remove noise (i.e., incorrect term pairs), which is introduced by the term mappers.

When the data (i.e., term collections) have been acquired, work on terminology integration in SMT systems can be started. In general, there are two conceptually different approaches to SMT system adaptation and integration of user-specific term collections into SMT systems (both methods are depicted in Figure 1):

- Training level (static) integration. Static integration means that an SMT system is adapted for a specific term collection during the training phase of the SMT system. If a user (an SMT system developer) decides to modify the term collection by deleting, adding, or editing terms (or even replacing the term collection with a new term collection), the whole SMT system has to be re-trained to adjust to the changes made by the user (hence the name "*static*").
- Translation level (dynamic) integration. As static integration cannot be always performed (for instance, for small translation tasks for which re-training of SMT systems may not be economically justifiable), dynamic integration can be a beneficial alternative. Dynamic integration is performed during translation using a pre-trained SMT system. The integration is performed by enriching the source text instead of modifying the pre-trained SMT models. This means that the dynamic integration allows exchanging term collections without having the need to re-train the SMT system (hence the name *"dynamic*").

In the author's work all SMT integration experiments have been carried out within the LetsMT SMT platform (Vasiļjevs et al., 2012), which is based on the Moses SMT system (Koehn et al., 2007). All SMT and terminology integration in SMT experiments are performed using phrase-based SMT systems (Koehn et al., 2003).

1.1. Types of Term Collections

The single most important linguistic resource required to perform terminology integration in SMT systems is a bilingual term collection. A bilingual term collection consists of one to many entries, where each entry describes a bilingual term pair in a specific domain. Each bilingual term pair may consist of a definition, a source term and possibly other information describing the source term, and the target term and possibly other information describing the target term. The "*other information*" may be, for instance, linguistic information, provenance information, disambiguation information, ontological information, etc.). However, the set of the available data describing a term pair may be even limited to just the source term and the target term. This means that methods for terminology integration into SMT systems have to ensure that they can work even with the limited data set. In order to better understand what types of term collections this thesis focusses on, let us look at several types of term collections in terms of origins of the term collections.

The first and possibly the most known type of term collections are ***authoritative term collections***, that is, term collections stored in official or publicly available term data bases (or term banks), for instance, the *Interactive Terminology for Europe*[17] (IATE), the *EuroTermBank*[18] (Vasiļjevs et al., 2008), and many others. Such term collections often contain multilingual term entries where each term entry may contain multiple synonymous term phrases in source and target languages; usually in their canonical forms. As these term entries are usually created by terminologists (maybe even representatives of an authoritative institution for standardisation of terminology for a certain language), the only additional information that is usually attached to the terms is the definition and the domain (i.e., the subject field) the term entry belongs to; further linguistic (e.g., morphological, syntactic, etc.) information is usually not provided. Because such term entries are in general intended for human use, they may also contain dictionary type mark-up, for instance, the Latvian term "*kravietilpība*"[19] from *EuroTermBank* has the English equivalent "*[cargo-]carrying capacity*" specified. It is evident that the optional constituent "*[cargo-]*" of the English equivalent is intuitive for humans, however, it may not be readable by automated processes.

A different type of term collections are ***professional translator created term collections***. Translators in their professional duties often use custom term collections for specific translation tasks or for specific translation domains (even for specific customers). Differently from authoritative term collections, these term collections often contain only pairs of term translations (without any attached additional information) in their canonical forms. As the term collections are usually relatively focussed, they are less ambiguous than the authoritative term collections. That is, each source term in most cases has just one translation equivalent in the target language. This, of course, is much better suited for automated processes as the search

[17] The Interactive Terminology for Europe can be accessed online at: http://iate.europa.eu.

[18] EuroTermBank can be accessed online at: http://www.eurotermbank.com/.

[19] The term entry can be found online at: http://www.eurotermbank.com/GetEntryDetailed.aspx?id=0&more=1&item=519766&resource=0.

space is smaller and the translator usually uses a computer-assisted translation environment, which requires term collections to be defined in machine readable formats.

A completely different type of term collections (because of no human intervention) are ***automatically created term collections***. Using bilingual term extraction methods, term collections can be extracted automatically from parallel or comparable corpora. Because the extraction processes may involve linguistic processing (e.g., morpho-syntactic tagging, lemmatisation, etc.), the extracted term pairs may contain structured and machine readable meta-data describing the terms. When working with automatically created term collections it is important to understand that they will contain also wrongly identified and mapped term pairs (the so called "*noise*" in the data). However, differently from the first two types, automatically created term collections can contain term pairs in different inflected forms. This characteristic is important when translating into morphologically rich languages (i.e., terms are not always translated using canonical forms). For morphologically rich languages in many contexts specific inflected forms are required and automatically created term collections can capture the necessary morphological variations.

Because the automatically created term collections usually contain a certain amount of *noise* and very ambiguous term pairs (e.g., terms from the general language, which may correspond to many different equivalents in the target language), manual revision of the term collection may be necessary. Such ***manually revised term collections*** consequently contain less (or no) noise and less ambiguous terms, however, the linguistic characteristics that were present in the automatically created term collection are kept.

There are other types of term collections, of which the most known are the ***ontology-based term collections***. However, in many professional translation scenarios costs for creation of such term collections do not satisfy the potential benefits. Therefore, the main focus of the author's work is on bilingual term collections that contain independent term entries in which term pairs are described by term phrases in two languages (i.e., term pairs) for which not necessarily additional linguistic information is available.

2. AUTOMATIC TERM IDENTIFICATION

To integrate terminology into SMT systems, a term collection is required. In some situations the term collections may be available (i.e., created by an authoritative source, a translator, a terminologist, an automated process, etc.). If a term collection does not exist, it has to be created. This section, describes two types of term identification methods: 1) a method for automatic term identification using automatic term extraction methods (see section 2.2) that can be used for semi-automatic (Pinnis et al., 2013) and fully automatic (Pinnis et al., 2012) creation of term collections, and 2) two methods for automatic term identification using existing term collections (see sections 2.3 and 2.4) designed for application in machine translation.

When creating a term collection, we may want to include in the collection terms in their canonical (or dictionary) forms. However, terms in text corpora in a general scenario are not in their canonical forms. This means that a term normalisation process is necessary that transforms the terms from their inflected forms (as found in the corpora) into their respective canonical forms. To address this issue, section 2.5 presents a rule-based method for term normalisation developed by the author.

This section is based on research results published in the papers by Pinnis et al. (2012) and Vasiļjevs et al. (2014b) and the *TaaS* project's public deliverable D4.4 "*Integration in SMT Systems*" (*TaaS*, 2014b).

2.1. Related Work on Term Extraction

Term extraction[20] methods in a general scenario analyse text data (a sentence, a paragraph, a document, or even a corpus) and for each phrase (a single-word or multi-word unit) try to identify whether it can be a term candidate. The identification is performed by estimating (or validating depending on the method): 1) the term unithood (i.e., the phrase boundaries of terms), and 2) the term termhood (i.e., how likely a phrase is a term or how specific a phrase is within a corpus or a document). There has been extensive related research done by other researchers on term extraction that focusses on three types of term extraction methods:

- **Linguistically motivated** term extraction methods. Terms are identified using patterns, which are usually morpho-syntactic (or simply part of speech (POS)) regular

[20] Note that different authors use different names for the tasks of "*term extraction*", "*term recognition*", "*term tagging*", and "*term identification*". In this work the author by "*term extraction*" means the process of extracting monolingual lists of term candidates from documents (i.e., just the extraction process). "*Term recognition*" means the process of finding term occurrences in text using existing lists of terms. "*Term tagging*" is the process of tagging a document with term identifying tags (e.g., XML tags) using monolingual term candidate lists generated by the term extraction component. Finally, "term identification" is the whole workflow of processing a document, extracting term candidates (which may be also an optional step if an existing term collection has to be used for term recognition), and tagging terms in the documents.

expressions (also known as term grammars) for single-word or multi-word units that potentially comprise term candidates. The patterns can be hand-crafted by a language specialist or extracted from term-tagged corpora using semi-automated methods. As terms are mostly noun phrases that correspond to syntactic chunks (Bourigault, 1992; Justeson & Katz, 1995), linguistic filtering using morpho-syntactic patterns allows extracting term candidates that are syntactically sound (phrases that could be terms, but not necessarily are terms in a given context). Term extraction methods that perform linguistically motivated term extraction have been investigated by Bourigault (1992) in the *LEXTER* term extractor for French, Justeson & Katz (1995), Dagan & Church (1994) in the *Termight* term extractor for English, Jacquemin et al. (1997) for French, and many others.

- **Statistically motivated** term extraction methods. Terms are identified by performing statistical analysis of words and phrases within a large corpus (Pantel & Lin, 2001). In a general scenario, statistical methods perform minimum frequency filtering in order to filter out rarely occurring phrases, rank phrases using different co-occurrence measures (in literature also named as association measures). Co-occurrence measures allow identifying multi-word phrases that are more likely to be found together than as individual words or shorter phrases. There are many different co-occurrence measures, which have been proposed in related research. Several popular co-occurrence measures are, for instance, the Dice coefficient (Dice, 1945), pointwise mutual information (Church & Hanks, 1990), log-likelihood ratio (Dunning, 1993), the t-score statistic (Church et al., 1991), C-value (Frantzi & Ananiadou, 1997), Q-value (Merkel & Foo, 2007), and many other methods. The methods have been successfully applied in collocation and term extraction tasks (directly or in customised forms) by Pantel & Lin (2001), Pazienza et al. (2005), Wermter & Hahn (2006), Vu et al. (2008), Wong et al. (2008), Bouma (2009), Petrović et al. (2010), and many other researchers. An extensive overview of such measures (over 80 in total) is given by Pecina (2005) and Pecina & Schlesinger (2006). The last step for statistically motivated term extraction methods is a statistical cut-off that allows to decide whether a phrase can be considered a term candidate or not based on the term rankings. The cut-off can also be performed by extracting just the top *N* highest ranked term candidates (Delač et al., 2009). Statistical methods are useful when they are executed on large data sets from which they can draw reliable statistical measures. However, if the corpus is not large enough, the term extraction results may be poor (Grigonyte et al., 2011). As there is no linguistic analysis involved, purely statistical methods wrongly identify phrases starting or ending with stop-words as terms. Therefore, the methods are often enriched with stop-word filters (Petrović et al., 2010) that filter out phrases that start or end with stop-words, e.g., "*and*", "*or*", "*the*", "*in*", etc. This allows improving term unithood identification quality. Purely linguistically motivated methods do not perform term termhood analysis (except a simple sorting according to term frequency in a corpus) as it is assumed that phrases that are valid according to the morpho-syntactic term patterns are valid term candidates. Statistical methods, on the other hand allow identifying terms, which are more or less specific within a corpus.

- **Reference corpus motivated** term extraction methods. Terms are identified using statistics extracted from a broad domain corpus in combination with statistics from the document (or corpus) that is being analysed. As documents for translation may be very short (e.g., even just a sentence long), it can be impossible to obtain reliable statistics using statistical methods for the given documents. A large broad domain corpus can be used to identify words and phrases, which are more general and which are more specific. A commonly used measure is the inverse document frequency (IDF), which assigns lower scores to more general words in a corpus (Spärck Jones, 1972). This method is used together with other methods, because it acts as a filter that allows filtering out too general term candidates (Foo, 2012).

Above mentioned methods are often combined into hybrid methods that incorporate two or all three methods for term extraction (Daille, 1994; Dagan & Church, 1994; Dias et al., 2000; Hong et al., 2001, and many others). Extensive overviews of term extraction methods and techniques have been published by Pazienza et al. (2005), Wermter (2008), and Foo, (2012).

Term extraction in recent years has been also addressed for the Baltic languages. For instance, Kruglevskis (2010) has shown that for Latvian linguistically motivated term extraction methods allow achieving better term extraction results due to the morphological richness of the language. Grigonyte et al. (2011) have made similar findings for Lithuanian term extraction by comparing linguistically and statistically motivated term extraction methods.

2.2. Tilde's Wrapper System for CollTerm

In this section the author presents a workflow for automatic term identification consisting of automatic term candidate extraction from text documents (for instance, news articles, technical manuals, knowledge base articles, such as *Wikipedia* articles, etc.) and term tagging in the documents.

Usually automatic term extraction methods produce just lists of term candidates, for instance, *TermeX* (Delač et al., 2009), *CollTerm* (developed by Nikola Ljubešić and described in Pinnis et al., 2012), etc. The term candidate lists that the term extraction methods produce can contain overlaps of term candidates with different lengths (Frantzi et al., 2000). Consider the following example: "*A crash course in physics*". Term extraction methods might find two term candidates: a single word term candidate "*crash*" and a bigram term candidate "*crash course*" (both may be correct depending on the context). However, in order to capture a more specific representation of terms in the source document, only one of the term candidates is a valid term, e.g., in the example above, an intuitive selection is "*crash course*" if the document is about education. The same challenge has to be addressed by term recognition systems. For instance, if an existing term collection contains both entries (i.e., "*crash*" and "*crash course*" at the same time), term recognition systems have to be able to identify, which of the terms in the given context is the most probable. The task of selecting the correct term from a term candidate's list in a specific context is performed by term tagging methods. In case of the application-oriented scenario of machine translation, the less specific term may cause an SMT system to produce a wrong translation. In general, SMT quality has shown to be higher using longer phrases (Callison-Burch et al., 2005), because longer phrases allow capturing morpho-syntactic agreements between the different constituents of the phrases. This can be directly transferred to terms, that is, the more specific (the longer) fragment can be identified, the higher is the possibility that morpho-syntactic agreements between the term's constituents will be correctly transferred to the target language.

2.2.1. Term Candidate Extraction

For term candidate extraction, the author uses *CollTerm*. *CollTerm* is a tool for automatic extraction of collocation and term candidates from pre-processed (morpho-syntactically tagged) documents. The tool incorporates all three previously described types of term extraction methods. *CollTerm* filters terms using morpho-syntactic patterns (see Figure 2 for an example excerpt for Latvian and Lithuanian) and stopword lists. Stopword restrictions are specified in the term pattern list.

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

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

For Latvian and Lithuanian the term patterns have been created in a semi-automatic manner. At first, morpho-syntactic tag sequences were automatically extracted from morpho-syntactically tagged texts (Pinnis & Goba, 2011) in which terms were marked by human annotators. Then, the obtained morpho-syntactic tag sequences were manually revised and generalised into patterns. The rules for Latvian and Lithuanian (as it can be seen in Figure 2) are limited to four morphological categories: POS, gender, number, and case. The initial generalisation was performed by Dr. Inguna Skadiņa in the *ACCURAT* project (Pinnis et al., 2012) and an updated generalisation has been performed by the thesis author for the *TaaS* project resulting in 103 patterns.

After linguistic filtering, *CollTerm* performs statistic filtering using a minimum frequency threshold. The remaining term candidates are ranked using co-occurrence or reference corpus statistics. For multi-word term candidate ranking, CollTerm supports five co-occurrence measures (Dice coefficient, modified mutual information (MI), chi-square statistic (CS), log-likelihood (LL), and t-score statistic) or the reference-corpus based IDF (Spärck Jones, 1972) scores of words. Table 6 shows the top 10 lemmatised bigram term candidates extracted from the *Wikipedia* article "*Automobile*" using the t-score statistic with a minimum frequency of three for English and two for Latvian and Lithuanian. The candidates are given as lemma sequences since term candidate extraction over lemmatised data allows to perform better statistical analysis (due to reduced data sparseness).

Table 6. Top 10 normalised English, Latvian, and Lithuanian term candidate lemma sequences consisting of two words and their scores obtained with the t-score statistic

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

The reference corpus for IDF score calculation has to be large enough to represent the language (in terms of stopwords in contrast to words that may be important in term extraction). For instance, the Latvian corpus from which lemma IDF scores have been extracted consists of *Wikipedia* articles (7.6 million tokens) and Web news articles (8.2 million tokens). If the IDF score file is given and a co-occurrence statistic is used for n-gram term candidate ranking, a linear combination of TF-IDF and co-occurrence statistic is computed (Pinnis et al., 2012). Single-word terms are ranked using just the TF-IDF measure.

Finally, after ranking, a cut-off method is applied to filter out low ranked term candidates. The resulting list of term candidates is then exported as a sequence of lemmas for term tagging.

2.2.2. Term Tagging in Documents

CollTerm creates an output document containing a list of term candidates of a fixed length (up to four tokens) where n-grams (phrases) are ranked according to one of the ranking methods. This requires *CollTerm* to be executed multiple times to cover single-word and multi-word term candidate extraction.

Because the term candidate lists contain overlapping phrases, terms in the source document are tagged using the *Tilde's Wrapper System for CollTerm* (*TWSC*). *TWSC* takes as input plaintext or pre-processed tab-separated (broken into sentences, tokenised, and POS or morpho-syntactically tagged) documents. *TWSC* then produces either term tagged plaintext where term candidates are marked with *<TENAME>* tags (see Figure 3 for an example) or tab-separated documents (see Figure 4 for an example) where term candidates are marked with *B-TERM* (for the first token) and *I-TERM* (for the remaining tokens) tags. The plaintext annotation format is similar to the named entity annotation format used in the Message Understanding Conference 7 (MUC-7; Chinchor, 1997) and the BIO annotation scheme that was introduced in the CoNLL 2002 conference (Tjong Kim Sang, 2002) is used to tag terms in tab-separated documents.

```
<TENAME SCORE="0.17" MSD="N-msg---------n-----------f- N-msl---------n---
--------l-" LEMMA="serviss aprīkojums">Servisa aprīkojumā</TENAME>
ietilpst <TENAME SCORE="0.0" MSD="N-fpg---------n-----------l- N-fsg-----
----n-----------l- N-msn---------n-----------l-" LEMMA="bremze pārbaude
stends">bremžu pārbaudes stends</TENAME>, <TENAME SCORE="0.61" MSD="N-msg-
--------n-----------l- N-fsg---------n-----------l- N-fsn---------n------
-----l-" LEMMA="motors diagnostika ierīce">motora diagnostikas
ierīce</TENAME>, <TENAME SCORE="1.0" MSD="N-mpg---------n-----------l- N-
fsg---------n-----------l- N-msn---------n-----------l-" LEMMA="ritenis
balansēšana stends">riteņu balansēšanas stends</TENAME>, <TENAME
SCORE="0.44" MSD="N-mpg---------n-----------l- N-fsg---------n-----------
l- N-msn---------n-----------l-" LEMMA="amortizators pārbaude
stends">amortizatoru pārbaudes stends</TENAME>, <TENAME SCORE="1.0"
MSD="N-mpg---------n-----------l- N-fsg---------n-----------l- N-msn-----
----n-----------l-" LEMMA="ritenis montēšana stends">riteņu montēšanas
stends</TENAME> u.c.
```

Figure 3. Fragment of a term-tagged plaintext document in Latvian

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

Figure 4. Fragment of a term-tagged tab-separated document in Latvian

Within one term candidate list, it is possible to select the term candidate that is ranked higher. However, if the overlap is between candidates of different lists, the selection is not straightforward. Two methods have been applied in order to combine different n-gram term candidate lists into one list. The first approach prioritises longer n-grams, while the second approach combines all lists in one list using linear interpolation of term candidate confidence scores by applying different weights to the different length term candidate lists.

2.2.3. Term Tagging Evaluation for Latvian and Lithuanian

TWSC has been evaluated in multiple term tagging and term extraction scenarios. The following three sub-sections will describe: 1) evaluation of *TWSC* in SMT, 2) evaluation of *TWSC* for creation of term collections, and 3) evaluation performed by third party researchers.

2.2.3.1. Evaluation of TWSC for Term Identification for SMT Purposes

For term identification in SMT we aim at identifying term phrases that are: 1) non-breakable when translated, which means that they have to behave like syntactic chunks, 2) as specific as possible in order to correctly translate the terms into the target language. For this evaluation (published in ACCURAT, 2011 and Pinnis et al., 2012), human annotators were asked to manually annotate texts in the IT domain (software manuals, IT news, software reviews, etc.) for Latvian and Lithuanian languages. The annotators were specifically instructed to prefer longer phrases over shorter phrases as terms whenever in doubt.

The human annotated corpora were split into two parts – a development set and a test set. The former was used for tuning of different parameters of *CollTerm* and *TWSC* including: (a) minimum n-gram frequencies, (b) *CollTerm* confidence score thresholds, and (c) linear interpolation coefficients for the term candidate list combination method. The statistics of the human annotated corpora for Latvian and Lithuanian are given in Table 7.

Table 7. Statistics of the Latvian and Lithuanian human annotated corpora

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

During evaluation parameters were tuned on the development set using an iterative approach. At first the minimum n-gram frequency constraints were tuned using the prioritised list combination method. Also the statistical ranking methods were evaluated to identify, which ranking method allows achieving the highest precision, recall, and F-measure (F1) without application of *CollTerm*'s confidence score thresholds. Then term candidate confidence score thresholds were tuned in order to achieve better performance. Results using various term candidate ranking methods on the Latvian and Lithuanian test sets are given in Table 8.

The results show that for Latvian the best recall was achieved with the log likelihood ranking method (70.66%), the best precision was achieved with the chi square statistic (59.85%), and the best F-measure was achieved with the modified mutual information ranking method (54.05). The difference between the different methods is, however, relatively insignificant. For instance, the best achieved F-measure without confidence score threshold tuning with the log likelihood statistic is 54.26 (54.23 on the development set) and with the Dice coefficient - 54.05 (54.35 on the development set). As the development set for the Lithuanian language is relatively small, all term candidate ranking methods produced identical results. Therefore, for further tuning of parameters for Lithuanian the MI measure was selected.

Table 8. Results of the term tagging evaluation for Latvian and Lithuanian

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

Table 8 also shows that threshold tuning on the Latvian development set improves results (in terms of recall, precision, and F-measure) on the test set as well. Although the evaluation shows an F-measure drop for Lithuanian, the author believes that the size of the tuning corpus needs to be increased in order to reliably tune the parameters.

Finally, the interpolation parameters were tuned in order to achieve better F-measure with the interpolation-based term candidate list combination method. The results in Table 8 suggest that the prioritisation method significantly outperforms the interpolation-based method. Moreover, the tuned parameters suggest that longer n-grams are preferred (even in the interpolation-based method).

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

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

2.2.3.2. Evaluation of TWSC for Term Identification for Terminology Creation Purposes

The second evaluation with a goal to evaluate the performance of *TWSC* for term collection creation purposes was performed in the *TaaS* project (Vasiļjevs et al., 2014b). The evaluation covers four languages (English, German, Hungarian, and Latvian) and two subject fields (information technology and mechanical engineering). Two annotators (language specialists with a focus on terminology) were asked to annotate terms in two documents that the annotators would want to have in term bases. 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 according to the specialists who work with the terminology. Therefore, this evaluation included the individual annotations of both annotators. The results are given in Table 9.

Table 9. Evaluation results of TWSC for term collection creation

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	371	609	60.92%	332	770	43.12%

The results show that on average around 50% of the identified terms are true positives. Although seemingly average, the results are acceptable considering that simultaneous identification of termhood and unithood 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 & Skadiņš, 2012).

To show the linguistically, statistically, and reference corpora motivated method's (i.e., the TWSC tool's) superiority over a standard statistical term extraction tool used by translators, Table 10 shows evaluation results of TWSC in comparison to the term extraction tool integrated in the MemoQ computer assisted translation tool. As previously stated, for statistically motivated methods, the document has to be large enough to draw reliable statistics. However, the documents used in the evaluation contained approximately 2,000 and 2,200 words. Another reason for the significantly lower performance is that Latvian is a morphologically rich language and it requires (for better performance) the term extraction methods to be linguistically motivated.

Table 10. Evaluation results of TWSC and the term extraction method in MemoQ for term collection creation

Tool	Information Technology				Mechanical Engineering			
	Term candidates	**Precision**	**Recall**	**F1**	**Term candidates**	**Precision**	**Recall**	**F1**
MemoQ	30	33.33%	1.55%	2.95	31	45.94%	2.37%	4.48
TWSC	609	60.92%	57.34%	59.08	770	43.12%	60.47%	50.34

2.2.3.3. Third Party Evaluation of TWSC

TWSC has been also evaluated by other researchers in the *ACCURAT* and *TaaS* projects as well as by independent (not related to the author) researchers. This section summarises the main findings of these evaluation efforts.

Nikola Ljubešić has evaluated *TWSC* for Croatian in Pinnis et al. (2012). He performed a comparison of purely statistical term identification and linguistically motivated term identification with *TWSC*. The results showed that the linguistic filtering allows improving term identification precision from 4.5% up to 41% on the Croatian test set. This proves the superiority of the linguistically motivated term identification method over statistical methods.

Aker et al. (2014a) have evaluated *TWSC* in a term extraction task for 21 languages using *EuroVoc* (Steinberger et al., 2002) terminology identified in *Wikipedia* documents. They analysed the impact of different POS taggers and manually and automatically created patterns for term extraction. They showed that automatically created patterns for projected POS taggers perform similarly to POS taggers trained on language specific part-of-speech tagged corpora and manually (or semi-automatically) created patterns. Their evaluation efforts proves the applicability of TWSC for morphologically rich languages and also under-resourced languages.

Very recently, an evaluation performed by Arcan et al. (2014a) of the *TaaS* platform showed that *TWSC* achieves similar results to an extended version of the *Wiki Machine* (Arcan et al., 2014a) for English and Italian. Out of four data sets it achieved higher results for three data sets in terms of F-measure. Their evaluation shows that *TWSC* allows identifying significantly more (from 2.5 to even 16 times more) multi-word term candidates than the Wiki Machine. This is due to the fact that *TWSC* in general prefers longer term phrases over shorter term phrases. Their evaluation further shows that integrating such longer terms into SMT systems, allows achieving higher SMT quality.

2.2.4. Application of TWSC in Machine Translation

The application of *TWSC* for term identification in SMT has the following benefits:

- The identified term candidates are linguistically motivated (e.g., we can ensure that verbs are not translated as nouns, stop-words are not treated as terms, etc.).
- The termhood of the identified terms is strengthened by the statistical analysis that is performed in *TWSC* (i.e., we can distinguish domain specific terms from general domain phrases).
- For languages, for which lemmatisation support can be ensured, *TWSC* allows to identify terms in all inflected forms the terms can appear in a text. This is very important for short documents where terms may occur multiple times, but in different inflected forms.

However, *TWSC* has also several issues, which have motivated to investigate different term identification methods that could be better suited for integration in SMT systems:

- In the case if *TWSC* wrongly identifies the term unithood (i.e., specifies wrong term boundaries) for some terms and the correct terms are included in the term collection, the correct terms will not be identified. For instance, imagine that we have in our term collection the term "*crash course*" and we need to pre-process the following sentence:

"*A crash course in physics*". As an output *TWSC* might find the term "*crash*", which according to our term collection would not be correct. Because we did not find "*crash course*" with *TWSC*, but only "*crash*", *TWSC* would not process this example correctly.

- Because of the statistical analysis performed by *CollTerm*, *TWSC* cannot be executed on very large (e.g., more than 5MB) plaintext documents. When processing documents, *CollTerm* has to read the whole POS-tagged document into memory and this can influence system stability.
- For longer documents, the statistical analysis that is performed by *CollTerm* can be very time consuming (up to several minutes). As in professional translation speed is very important, the processing time can result in higher translation costs.
- Because of the limitations of *CollTerm*, *TWSC* is able to identify only terms that consist of up to four tokens. Longer terms that contain, e.g., conjunctions often cannot be identified because of this limitation.

2.3. Pattern-Based Term Identification

The main application of *TWSC* is term identification for semi-automated creation of term collections. However, due to its limitations, TWSC is not suited for terminology integration in SMT. When translating a document, a user usually has an existing term collection available. This means that it is not necessary to perform statistical analysis of the text.

This section proposes to perform linguistic term phrase filtering using morpho-syntactic patterns to identify terms. The method performs term identification in the following steps:

- At first, part-of-speech or morpho-syntactic tagging of the source text is performed in order to tokenise the content and enrich it with linguistic information.
- Then, the morpho-syntactic patterns from *TWSC* are used to identify linguistically valid term candidates. At this point, the identified phrases may overlap. For instance, for the following sentence: "*Do I need a computer mouse?*" we can identify the following valid term phrases: "*computer mouse*" (two nouns), "*computer*" (a noun), and "*mouse*" (a noun).
- Finally, the identified term phrases are cross-referenced with the bilingual term collection. Either the lemma sequences with POS categories or stemmed phrases can be used in this process depending on the linguistic support for a language and the richness of linguistic information in the bilingual term collection. The identified terms are annotated in a left-to-right manner preferring longer term phrases wherever possible. Imagine that we have a term collection, which contains the terms "*computer*"

and “*computer mouse*”. In the example “*Do I need a computer mouse?*” this method would identify “*computer mouse*” as a term.

The strength of this method is the ability to filter out morphologically and for morphologically rich languages also syntactically invalid term candidates. For instance, if in our term collection there is a term “*can*” (i.e., a cylinder type object) this method can effectively deal with the example: “*I can can a can*”. That is, the first “*can*” would be identified as a modal verb, the second “*can*” would be identified as a verb, and only the third “*can*” would be tagged as a term (because it is a noun). Of course, the quality of term identification depends heavily on the quality of the POS or morpho-syntactic tagger that is used to pre-process the source text.

For Latvian, the term patterns in TWSC define also morpho-syntactic agreements between constituents of a term. The linguistic filtering allows filtering out phrases that do not satisfy the agreement requirements. Imagine that we have the term “*datora pele*” (“*computer mouse*”) in a term collection. The linguistic filtering using patterns allows identifying that the sentence “*Datoram pele ir svarīgs aksesuārs*.” (“*For a computer, a mouse is an important accessory*.”) does not contain the term “*datora pele*”, because the phrase “*datoram pele*” does not satisfy the agreement requirement for phrases consisting of two nouns. That is, the first noun has to be in a genitive case, while “*datoram*” is in the dative case.

As this method is not used to create term collections, but rather to identify terms when translating text, it is evaluated in SMT integration scenarios (see section 5.5).

2.4. Fast Term Identification

TWSC and the *Pattern-Based Term Identification* methods both require language dependent linguistic tools and resources in order to identify terms. However, there are languages for which such resources (for instance, POS taggers) are not openly available (e.g., Irish). The application of POS (or morpho-syntactic) taggers is also a very time and resource consuming process, which in some scenarios may not be applicable due to the necessity to provide almost instantaneous results (for instance, term identification in translation segments during translation). For such applications a simpler term identification method was investigated (the *Fast Term Identification* method). The goal of this method is to: 1) achieve fast processing of both – the source text (e.g., real time or close to real time) and SMT system training data (e.g., this method can be used to identify bilingual terminology in *Moses* phrase tables when performing translation model adaptation (see section 4.4)), and 2) to ensure that terms from the bilingual term collection would be identified regardless of their specificity (i.e., without statistical analysis). In this method terms are treated as multi-word sequences and no morpho-syntactic restrictions are applied on the identified phrases.

The *Fast Term Identification* method performs term identification using a left-to-right search over stemmed tokens of the source text (and stemmed terms from the bilingual term collection). Term identification prioritises longer terms over shorter terms. Stemming is performed in order to capture morphological variations of terms, i.e., terms in different inflected forms. For instance, if we have to pre-process the following text:

"Vai man ir vajadzīgs peles paliktnis? Datoram peles paliktnis ir svarīgs aksesuārs."

and we have an English-Latvian term collection (Table 11)

Table 11 Example English-Latvian term collection

English	Latvian
computer	dators
computer mouse	datora pele
mouse pad	peles paliktnis
mouse	pele

the following terms would be identified:

"Vai man ir vajadzīgs **[peles paliktnis]***?* **[Datoram peles]** *paliktnis ir svarīgs aksesuārs."*

The example shows that the term "*peles paliktnis*" was prioritised over "*pele*", because the *Fast Term Identification* method also prioritises longer term phrases. However, because of the prioritisation, the method can also identify incorrect phrases. For example, the phrase "*datoram peles*" does not represent a term from the term collection (the first word is in the dative case instead of a genitive case) as the syntactically correct term would have been "*peles paliktnis*". Because the *Fast Term Identification* method does not perform any linguistic analysis it can create such mistakes. In a different example "*He planted a tree near the power plant.*" a possible mistake would be to identify "*planted*" as a term if a term collection would contain "*plant*" (a noun) as a term. In this example, the verb "*to plant*" would be mistakenly identified as the noun "*plant*". Similarly to the previous method, this method is intended only for term identification in the source text using existing term collections. Therefore, it is evaluated in SMT integration scenarios (see section 5.5).

2.5. Term Normalisation

For users (translators, terminologists, etc.) who work on morphologically rich languages and want to create term collections using semi-automated methods, *TWSC* may produce redundant term candidates. For example, in Czech, Latvian, Estonian, etc., nouns, verbs, adjectives (and other types of words) may have numerous different inflected forms. Therefore, terms have to be normalised for inclusion in term collections. Term normalisation is a process of transforming terms from their inflected forms into their corresponding canonical forms (i.e., dictionary forms). This section describes a rule-based method for term normalisation for the

Latvian language, however, it can be extended to other languages easily. Term normalisation for Latvian has previously been investigated by Vancāne & Krugļevskis (2003). The author builds on the idea by Vancāne & Krugļevskis (2003) and proposes a workflow for term normalisation that can be used for terms identified with the term patterns from TWSC.

The method works as follows:

- At first, terms are identified in a document using *TWSC*.
- Then, for each term we identify the corresponding term pattern that matches the term's morpho-syntactic tag sequence.
- Next, for each pattern a transformation rule for the normalisation is selected from a *pattern transformation table*.
- Finally, term normalisation is performed by synthesising the required inflected forms for each word of a term using a morphological synthesiser. For Latvian, the author uses the morphological synthesiser developed by Deksne (2013).

The *pattern transformation table* is a manually created tab-separated document in which each line specifies a separate morpho-syntactic transformation rule (i.e., how to normalise a term that matches a specific pattern). For single-word terms the normalised forms often correspond to the 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*". For Latvian there are in total 230 different normalisation rules implemented in the term normaliser. An example excerpt of the *pattern transformation table* for Latvian is given in Figure 5.

```
A        ^A.[^ ]*$                                  A*msn*
G        ^G.[^ ]*$                                  G*msn*
N        ^N[^ ]*$                                   LEMMA
A        ^A.fpn[^ ]* N.fpn[^ ]*$                    A*f!n* LEMMA
G        ^G.fpn[^ ]* N...g[^ ]* N.fpn[^ ]*$         G*f!n* TOKEN LEMMA
A        ^A.fpn[^ ]* G.fpn[^ ]* N.fpn[^ ]*$         A*f!n* G*f!n* LEMMA
```

Figure 5. Example excerpt of a pattern transformation table for Latvian

An entry consists of three parts: 1) the POS of the first token, 2) the term pattern, and 3) the transformation rule. The transformations in the transformation rule can be as follows:

- The rule "*TOKEN*" specifies that the token should remain as it is;
- The rule "*LEMMA*" specifies that the token's lemma should be used.
- A positional transformation rule (e.g., "*A*msn**") specifies how to transform particular morpho-syntactic categories in the morpho-syntactic tag of a token to acquire the tag

of the normalised term’s particular token. The morpho-syntactic category transformations can be as follows:

- The symbol “*” denotes that the value of the category has to be kept as it is.
- The symbol “!” denotes that the value of the category has to be taken from the token whose transformation rule is equal to “*LEMMA*”. This is a syntactic rule that requires agreements between two or more tokens of a term.
- Other symbols denote a specific value for a category.

The positional transformation rules are shorter than the actual morpho-syntactic tags (for Latvian, the tag can describe 28 categories). The remaining category values have to be kept as they are (i.e., the rule “*” is applied for the remaining categories).

Figure 6 shows four examples of how morpho-syntactic transformation rules are applied. The examples show data triplets comprising of the original tag sequence, a transformation rule and the transformed morpho-syntactic tag sequence.

```
Tag sequence
      Rule
            Transformed tag sequence

N-msl---------n-----------l-
      LEMMA
            N-msn---------y-----------l-

N-fsg---------n-----------f- N-fsd---------n-----------l-
      TOKEN LEMMA
            N-fsg---------n-----------f- N-fsn---------n-----------l-

A-fsnc-y------------------l-
      A*msn*
            A-msnc-y------------------l-

Gpfpdc-y---p--------------l- N-fpd---------n-----------l-
      G*f!n* LEMMA
            Gpfsnc-y---p--------------l- N-fsn---------n-----------l-
```

Figure 6. Examples of morpho-syntactic tag sequence transformation rules

2.6. Summary of Automatic Term Identification

In this section the author presented novel methods for term identification. In total, three methods were analysed (and implemented): 1) the linguistically and statistically motivated term identification method using *TWSC*, 2) the *Pattern-Based Term Identification* method, which is based on the linguistically motivated part of *TWSC*, and 3) the *Fast Term Identification* method, which is lightly linguistically motivated.

The main usage scenario of *TWSC* is identification of terms in documents for semi-automated creation of term collections. For this scenario *TWSC* has been evaluated by different parties (the author and other researchers) and it has shown to achieve stat-of-the-art term identification performance. It has been also shown by author’s evaluation efforts and third party evaluation efforts that TWSC is applicable for term identification for morphologically rich languages (e.g., Latvian, Estonian, Czech, etc.) and for languages that can be considered under-resourced.

The *Pattern-Based Term Identification* and the *Fast Term Identification* methods were specifically designed for SMT purposes. The methods are evaluated in the context of terminology integration in SMT (see section 5.5).

This section introduced also a rule-based term normalisation method for Latvian that is built on the ideas for Latvian term normalisation by Vancāne & Krugļevskis (2003). The method uses morpho-syntactic transformation rules (a single rule for each morpho-syntactic term pattern defined in *TWSC*) to generate the normalised (or canonical) forms of terms from their inflected forms.

3. CROSS-LINGUAL TERM MAPPING

An important step in automated bilingual term extraction workflows is cross-lingual term mapping. Term mapping is a process that after monolingual term identification performs a cross-lingual analysis over monolingual terms in two languages and identifies term pairs that can be considered reciprocal translations. Such automatically extracted bilingual terminology is a valuable resource not only in human and machine translation, but also in many other fields, for instance, cross-lingual information retrieval, semantic analysis, question answering, and many others.

This section describes a novel method for cross-lingual term mapping that can be used in workflows for the automatic or semi-automatic creation of bilingual term collections for SMT purposes. The method requires a set of linguistic resources, therefore, this section also describes the creation process of the necessary resources, of which the most important resources are bilingual probabilistic dictionaries and transliteration systems. The description of the term mapping method and its evaluation is based on the research paper by Pinnis (2013). The section 3.3 on the probabilistic dictionary filtering is based on the author's contribution to the publication by Aker et al. (2014b) and the section 3.4 on character-based SMT transliteration systems is based on the publication by Pinnis (2014).

3.1. Related Work on Term Mapping

Multi-lingual term collections can be automatically acquired from existing resources (monolingual lists of terms, parallel or comparable corpora, etc.) with the help of term mapping. Term mapping methods according to previous research in the field can be divided in two categories – context dependent methods and context independent methods.

The context dependent methods are applicable in situations when there is enough context from which to draw reliable statistics. The necessary amount of context can differ depending on the method. For instance, for term mapping in parallel data it can be enough to have one parallel document pair or a sentence-aligned parallel corpus (Federmann et al., 2012; Wolf et al., 2011; Lefever et al., 2009; Gaussier et al., 2000).

For under-resourced languages and numerous domains, however, parallel resources are scarce and not always available. Therefore, a more promising resource is comparable corpora, which has recently received much attention in the scientific community for its applicability in MT (Skadiņa et al., 2012). Most of the context-dependent methods designed for term mapping in comparable corpora, however, require relatively large corpora (e.g., hundreds or even thousands of documents) in order to calculate reliable cross-lingual association measures (Fung and Yee, 1998; Rapp, 1999; Shao & Ng, 2004; Morin & Daille, 2010). The proposed methods have been focussed on language pairs with relatively simple morphology (e.g., German-English, French-English), but have not been thoroughly investigated for more complex languages (e.g., Finnish, Latvian, etc.). A recent study in the European Commission financed project TTC (2013) revealed that while the context-dependent methods developed in this project (Morin et al., 2010) perform well for English-French, their applicability for English-Latvian is questionable because of a term mapping precision below 5%. Laroche & Langlais (2010) also reported a relatively low precision (far below 50%) using context-dependent methods for the English-French language pair.

Context independent term mapping methods, on the other hand, are designed for situations when there is no context or the context is not large enough to draw statistics. Recent work on context independent term mapping has been carried out by Ştefănescu (2012) where a cognate similarity measure based on the Levenshtein distance (Levenshtein, 1966) was applied in order to estimate how similar two terms are. The method’s weakness, however, is a very limited term mapping recall.

3.2. *MPAligner* – a Context Independent Term Mapper

Following related research on context independent term mapping, the author has designed a new context independent method for term and term phrase mapping in term-tagged comparable corpora. The method allows mapping multi-word terms and terms with different numbers of tokens in the source and target language parts – two term mapping scenarios that have not been sufficiently addressed by previous research. The mapper has been specifically designed to address term mapping between European languages (including languages with different alphabets that are based on Latin, Cyrillic and Greek alphabets) and it allows integrating linguistic resources to increase recall (while maintaining the same level of precision) of the mapped terms.

The mapper has been evaluated by the author using the *EuroVoc* thesaurus (Steinberger et al., 2002) for 23 language pairs and for the Latvian-English language pair on a medical domain comparable corpus that was collected from the Web. The evaluation shows benefits of having additional linguistic resources (e.g., probabilistic dictionaries, and transliteration support) with respect to having only some of the resources (or none at all) available.

3.2.1. Term Mapping Method

Given two lists of terms (in two different languages) the task of the term mapping system is to identify which terms from the source language contain translation equivalents in the target language. The system (as shown in Figure 7) consists of two main components – monolingual term pre-processing and term mapping. A possible third module that is not discussed in the scope of *MPAligner* is term pair consolidation – a language specific process that performs term pair grouping by identifying different inflected forms of terms and allows increasing term mapping precision by filtering out possible invalid mappings. However, a method for consolidation of *MPAligner* output has been proposed by the author in Vasiļjevs et al. (2014b).

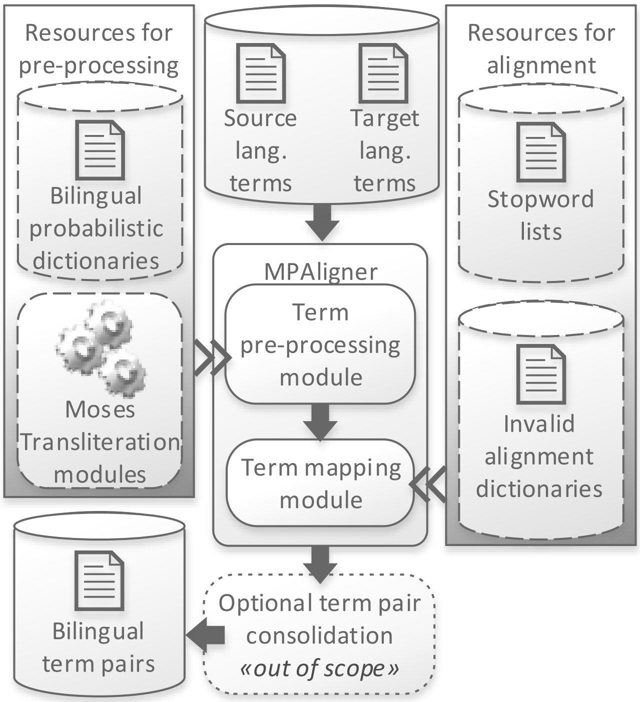

Resources for pre-processing
Bilingual probabilistic dictionaries
Moses Transliteration modules
Source lang. terms
Target lang. terms
MPAligner
Term pre-processing module
Term mapping module
Resources for alignment
Stopword lists
Invalid alignment dictionaries
Bilingual term pairs
Optional term pair consolidation
«out of scope»

Figure 7. The conceptual design of *MPAligner*

3.2.2. Term Pre-processing

Before mapping, all source and target language terms are tokenized and pre-processed using linguistic resources (if such are available). For each token the pre-processing module:

- Rewrites the token using lower-case letters;
- Rewrites the token with letters from the English alphabet (*simple transliteration*); letters that cannot be rewritten (e.g., the Russian softening and hardening marks "*ь*" and "*ъ*") are removed and letters that correspond to multiple letters in the English alphabet are expanded (e.g., the Russian "*ш*" and Latvian "*š*" are rewritten as "*sh*").
- Finds top *N* translation equivalents using a probabilistic dictionary in *Giza++* format (Och & Ney, 2003).
- Finds top *M* transliteration equivalents in the target language using a *Moses* (Koehn et al., 2007) character-based SMT transliteration system.

Table 12 gives an example of a term in Latvian and English languages ("*extensive farming*") that has been pre-processed with direct *source-to-target* and *target-to-source* linguistic resources. If direct resources are not available, English can be used as an *Interlingua* for the dictionary-based look-up and the SMT-based transliteration.

Table 12. Examples of pre-processed terms (a dash means that the particular value could not be acquired)

	Latvian term “*Ekstensīvā lauksaimniecība*”	
Lowercase form	ekstensīvā	lauksaimniecība
Simple transliteration	ekstensiva	lauksaimnieciba
SMT transliteration	extensiva, extensive	lauximnieciba
Translation	-	agriculture, farming
	English term “*Extensive farming*”	
Lowercase form	extensive	farming
Simple transliteration	extensive	farming
SMT transliteration	ekstensīviem, ekstensīvie, ekstensīvai	farmēšana, farmings, farming
Translation	apjomīgam, ekstensīvas, izvērstāku	turēšanas, saimniekošanas, zemkopībā

The system allows limiting the retrieved candidates with confidence score thresholds, therefore, for the Latvian-to-English direction the example shows no more than three transliteration candidates. For translation a limiting factor is also the available number of entries in the probabilistic dictionary.

3.2.3. Term Mapping

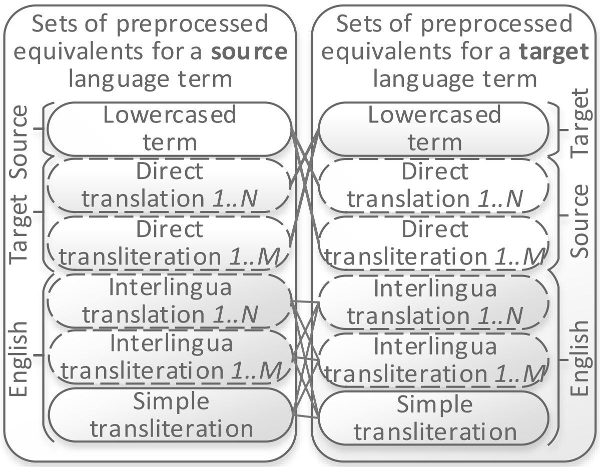

Sets of preprocessed equivalents for a **source** language term
Sets of preprocessed equivalents for a **target** language term
Source
Target
English
Target
Source
English
Lowercased term
Direct translation *1..N*
Direct transliteration *1..M*
Interlingua translation *1..N*
Interlingua transliteration *1..M*
Simple transliteration
Lowercased term
Direct translation *1..N*
Direct transliteration *1..M*
Interlingua translation *1..N*
Interlingua transliteration *1..M*
Simple transliteration

Figure 8. Bi-directional comparison sets for a single pre-processed term pair

After pre-processing, the mapping module performs bi-directional term mapping. As shown in Figure 8, for each token in a term the mapping module operates with a set of constituents - *1* to *N* translation equivalents, *1* to *M* transliteration equivalents, one simple transliteration equivalent and one lowercased equivalent. The set of available constituents depends on the linguistic resources used (e.g., direct dictionaries, interlingua dictionaries, no dictionaries, etc.).

The task of the mapping module is to decide whether a term pair can be mapped or not. The mapping process will be explained with the help of an example – the mapping of the

English term “*dose of chemotherapy*” and its German translation “*chemotherapiedosis*”. The mapping is performed in three steps.

3.2.3.1. Identification of Content Overlaps

At first, for every pre-processed token’s constituent, we identify the *longest common substring* in all pre-processed constituents of the target term’s tokens that are in the same language (in Figure 8 comparison sets of the same language are connected with a bi-directional arrow). For the German-English example, the pre-processing module produced “*chemotherapiedosis*” as a simple transliteration of the German term. As the English lowercased term and the simple transliteration of the German term are within valid comparison sets, the mapper will analyse content overlaps between these constituents.

When identifying the *longest common substring*, the positions of the substring within the constituents are retained. If the length difference between the substring and the full source or target constituents exceeds a threshold (defined in a configuration file), the substring information is kept for the next step.

The results of the first step on the example are given in Figure 9. Two of the three English constituents (“*dose*” and “*chemotherapy*”) can be nested within the German constituent. The third constituent’s (“*of*”) character overlap does not exceed the threshold (0.75 has been empirically selected as an appropriate default value), therefore, the substring information is ignored.

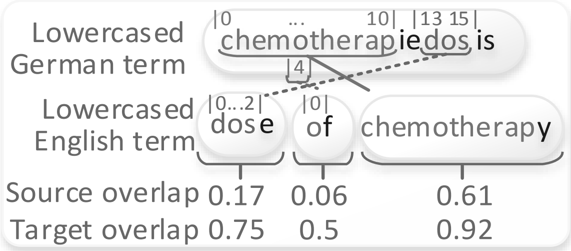


Figure 9. Longest common substring overlaps in German and English candidates

If the longest common substring overlap does not exceed the threshold, the mapper uses a fall-back method based on the *Levenshtein distance* as applied by Ştefănescu (2012). The *Levenshtein distance* metric is transformed to the following similarity metric:

$$Sim(s_1,s_2) = \frac{max(len(s_1),len(s_2))\text{-}LD(s_1,s_2)}{max(len(s_1),len(s_2))} \qquad (2)$$

where *LD* is the *Levenshtein distance* between two strings, and *len* is a string length function. Each deletion, insertion and substitution is equally penalised with one point as in the first version of the *Levenshtein distance* (Levenshtein, 1966).

The motivation behind application of the alternative metric is that the SMT transliteration may introduce additional or different letters in a string and thus the longest common substring-based method can fail. However, the fall-back method has a limitation. That is, it does not allow sub-word level mapping and if the similarity between two strings exceeds a predefined threshold, it is assumed that there is a complete overlap between the two strings. Assuming that the first comparison did not produce satisfactory results, Figure 10 shows the results of the alternative comparison for our example, however, none of the candidate pairs achieves a sufficient content overlap.

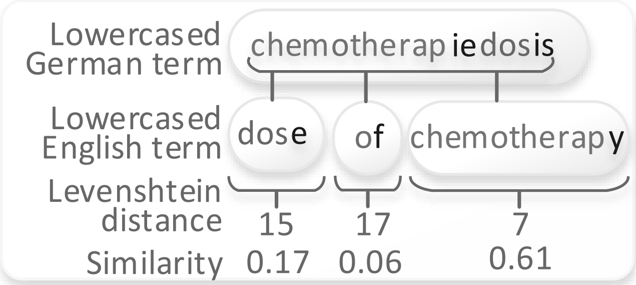


Figure 10. Levenshtein distance-based overlaps in German and English candidates

The result of this step is a list of binary alignment maps for constituent pairs. For instance, the binary alignment maps for "*chemotherapiedosis*" and "*dose*" are "*000000000000011100*" and "*1110*".

3.2.3.2. Maximisation of content overlaps

In the next step the binary alignment lists are used to identify the mapping sequence that maximises the content overlap between the two terms. At first, the system iterates through the source term's tokens and tries to find for each token the constituent that has the highest overlap in a target term's constituent. At the same time the system maintains for each target term's token a binary one-dimensional alignment map that defines what part of the token has been already mapped in order not to allow conflicting and overlapping alignments. The length of the alignment map is determined by the longest constituent of the source and target terms. To find similar mappings from the target language, the iterative process is performed also for each token of the target term.

The example above contained two content overlaps (remember that the overlaps of the constituent "*of*" did not exceed thresholds). The overlap maximisation process in two iterations is shown in Figure 11.

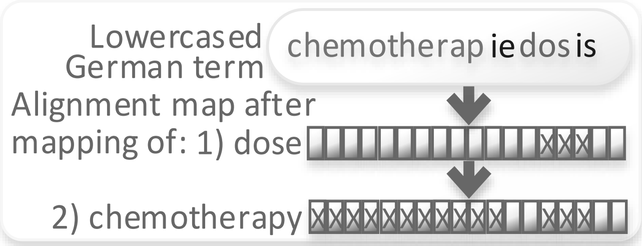


Figure 11. An example of the alignment map generation process for the German-English term pair

The goal of the mapper is to find term mappings that have a content overlap between terms in a way that restricts non-aligned segments (tokens or parts of tokens), but still allows a certain degree of imperfect mappings. For instance, we want the system to be able to decide that "*cost of treatment*" in English can be mapped to "*ārstēšanas izmaksas*" in Latvian (which is a direct translation) although it is evident that the token "*of*" does not have a mapping. On the other hand, we do not want the system to decide that "*β particles*" in English can be mapped to "*daļiņas*" in Latvian (translated as "*particles*") as well as we would not want "*electromagnetic field*" in English to be mapped to "*magnētiskais lauks*" in Latvian (translated as "*magnetic field*"). There is no perfect recipe that allows identifying all good and sufficient mappings from all bad and incomplete mappings in a language independent fashion, however, the mapper allows users to decide whether non-mapped segments at the beginning or the end of terms should be allowed or prohibited. Consequently, the mapper can be executed in order to allow trimmed mappings, but not to limit non-mappings in-between of mapped segments. When trimmed mappings are allowed, it is important to disallow terms starting or ending with stopwords. Therefore, the mapper allows filtering out trimmed term mappings that start or end with stopwords if stopword lists are available.

3.2.3.3. Scoring of consolidated overlaps

In the final step the aligned constituents and their sequence that produced the character alignment map with the maximum content overlap are enrolled in two strings (source and target) in order to score the total overlap. The non-aligned source and target tokens (if there are any) are attached at the end of each string. At the same time, spaces are added to the other string to simulate non-aligned tokens. This allows penalising incomplete overlap segments.

As both the probabilistic dictionaries and the SMT-based transliteration systems provide confidence scores for each candidate, these scores are used as negative multipliers to filter out term pairs that have low confidence and may potentially result in invalid mappings.

The enrolled strings are scored using the *Levenshtein distance*-based similarity metric (described in section 3.2.3.1) multiplied by the negative multipliers. In the example the *Levenshtein distance* between "*chemotherapydoseof*" (representing the English term) and “*chemotherapiedosis$$*” (representing the German term; “*$$*” represent two space symbols) is *6*; the *Levenshtein distance*-based similarity is *0.7*. As the simple transliteration-based pre-processing does not produce a confidence score that could be used as a negative multiplier, the term pair is considered to be mapped if the mapping score (in our example *0.7*) is higher than the threshold.

3.2.4. How to Acquire Linguistic Resources?

MPAligner can benefit (i.e., produce term pairs with higher recall and also quality) from four types of optional linguistic resources: 1) probabilistic dictionaries, 2) external *Moses* SMT-based transliteration modules, 3) invalid mapping dictionaries, and 4) stopword lists.

The first three resources integrated in the term mapper can be created using *Giza++* probabilistic dictionaries that are extracted from the parallel corpora. However, because *Giza++* probabilistic dictionaries are very noisy, that is, the precision of the entries is close to 0% (Aker et al., 2014b), the dictionaries have to be somehow filtered in order to minimise the proportion of wrong translation equivalents. Section 3.3 presents a method that allows effectively filtering probabilistic dictionaries in order to extract good quality probabilistic dictionaries and also invalid mapping dictionaries.

Dictionaries in general (and also probabilistic dictionaries) may contain entries that can be considered to be reciprocal transliterations. Probabilistic dictionaries that have been extracted from large parallel corpora contain also transliterated words in many different inflected forms. Such pairs can be extracted and used to create statistical transliteration systems. Therefore, section 3.4 presents methods how such transliteration entries can be extracted and effectively used in order to create character-based SMT transliteration systems.

The fourth resource, namely a stopword list, is a common resource used in language processing technologies. A stopword list usually consists of functional words (e.g., conjunctions, prepositions, particles, pronouns) and words that appear in almost all documents of a broad domain corpus. Such words rarely start or end terms (if we follow the limitation to noun phrases), therefore they can be used to filter out wrong candidates. In the author’s work, stopwords are also used to filter out potentially wrong term mappings.

3.2.5. Evaluation

MPAligner has been evaluated by the author using two evaluation methods – automated evaluation and manual evaluation. The automated evaluation was performed for language pairs included in the *EuroVoc* thesaurus. It shows the applicability of the method for European languages and allows estimating the upper level of recall that can be expected on comparable

Web corpora. The manual evaluation was performed on terms mapped in a Latvian-English comparable Web corpora in the medical domain. This evaluation allows estimating the expected performance of the method in terms of precision on noisy data.

3.2.5.1. Automatic Evaluation

The automatic evaluation has three goals: 1) to show how additional linguistic resources influence term mapping, 2) to evaluate the performance on European language pairs, and 3) to compare results with previous research using the same evaluation corpus. The *EuroVoc* thesaurus was selected as a suitable test corpus for the automated evaluation because it covers 24 European languages, it contains a relatively large number of terms (at the time of evaluation – 6,797 terms for all languages except Hungarian with 6,790, Italian with 6,643, and Maltese with 987 terms), and in average 65.5% of terms across all languages are multi-word terms.

For each evaluated language pair two monolingual lists of terms were created. Because the mapper sees only two independent lists of terms, the search space for mapping is not 6,797 term pairs, but rather 46.2 million term pairs (e.g., 6,797*6,797 for English-Latvian). In this evaluation the highest matching (i.e., the top one) target term is retrieved for each source term. For the language pairs for which additional resources are available, for every token a maximum of five transliterations and 10 dictionary translations are retrieved.

At first, the mapping performance when using direct (*source-to-target* and *target-to-source*) linguistic resources, Interlingua-based (*source-to-English* and *target-to-English*) resources, and no resources was analysed. Figure 12 shows results (in terms of precision "*P*" and recall "*R*") for the Latvian-Lithuanian language pair. It is evident that direct resources allow achieving significantly higher recall than having Interlingua or no resources.

The results also suggest that the precision is stable at higher thresholds, however, it drops faster when using Interlingua-based resources. This can be explained by the noise that is introduced by the Interlingua-based resources. E.g., the term "*plakne*" (a type of a geometric figure) in Latvian can be wrongly be mapped to "*самолёт*" (a type of an aircraft) in Russian because both translate into English as "*plane*".

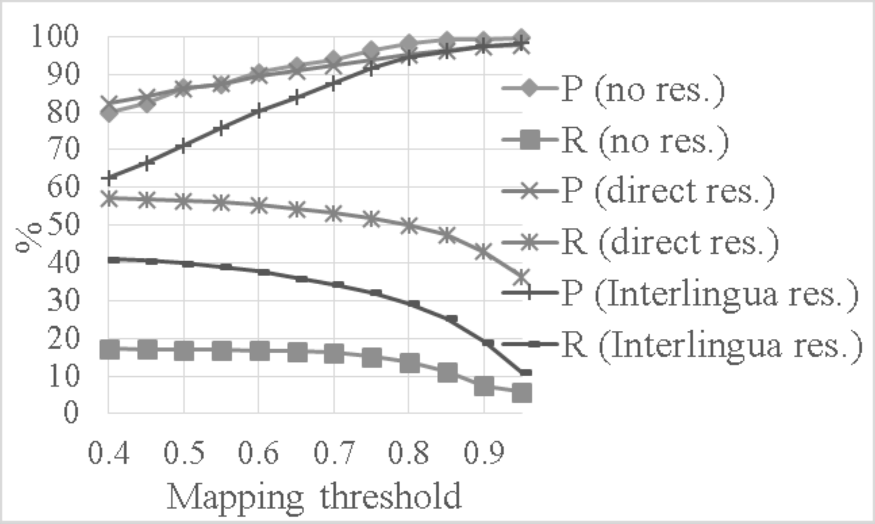


Figure 12. Latvian-Lithuanian evaluation results using direct, Interlingua, and no resources

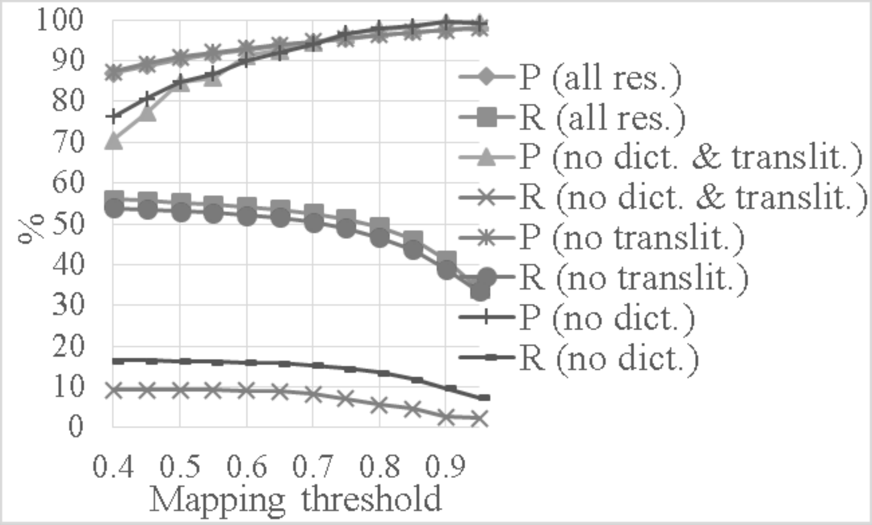


Figure 13. Latvian-English evaluation results using various resource configurations

Further, the benefits of having the probabilistic dictionaries and SMT-based transliteration modules were analysed. Figure 13 gives evaluation results for the Latvian-English language pair. The results show that without linguistic resources the recall is limited. This is due to the small number of terms that can be transliterated with the *simple transliteration* method. An analysis of 100 randomly selected English-Latvian unigram term pairs from the *EuroVoc* thesaurus revealed that 57 pairs were transliterations. 47 out of the 57 pairs were mapped using the *character-based transliteration* module. However, only 24 out of the 57 pairs were mapped using the *simple transliteration* method.

Evidently, adding resources allows significantly increasing the mapped term recall. It is also visible that the best results are achieved by using all linguistic resources.

Finally, term mapping was performed for 22 language pairs of the *EuroVoc* thesaurus with English as the source language. The results are given in Table 13. The evaluation was performed using direct *source-to-target* and *target-to-source* linguistic resources. The resources were built using *Giza++* probabilistic dictionaries extracted from the *DGT-TM* parallel corpus (Steinberger et al., 2012).

Table 13. Evaluation results for *EuroVoc* language pairs with English as the source language (languages are given in the ISO 639-1 format; the results are from experiments carried out in August, 2015 using an updated version of *MPAligner*)

Language pair	Precision	Recall	F1-measure
en-mt	90.3%	72.2%	80.2
en-ro	89.1%	67.5%	76.8
en-es	88.6%	66.8%	76.2
en-pt	88.5%	66.9%	76.2
en-fr	89.9%	64.5%	75.1
en-sk	89.4%	64.6%	75.0
en-lv	91.3%	62.7%	74.3
en-it	87.3%	64.1%	73.9
en-hr	89.5%	59.8%	71.7
en-sl	86.8%	60.8%	71.5
en-bg	85.8%	60.2%	70.8
en-lt	85.4%	58.1%	69.1
en-cs	85.3%	57.3%	68.6
en-pl	85.8%	54.9%	66.9
en-el	82.9%	55.0%	66.2
en-hu	78.7%	46.5%	58.4
en-nl	84.1%	42.9%	56.8
en-sv	82.4%	37.1%	51.2
en-da	83.4%	35.0%	49.4
en-et	74.9%	36.6%	49.2
en-de	77.7%	33.6%	46.9
en-fi	70.7%	31.8%	43.8

The evaluation results show that the author's method significantly outperforms results reported earlier by Ştefãnescu (2012) – an F1 score of 46.3 and 51.1 for English-Latvian and English-Romanian respectively when using the same probabilistic dictionaries. The term mapping method proposed by Ştefãnescu (2012) differs from the author's method in that it maps terms either with the *Levenshtein distance* based similarity metric or dictionary based exact match look-up. The author's proposed method, however, maps term tokens in sub-word level using maximised character alignment maps and applies Levenshtein distance just as a fall-back method and for scoring of the mapped term pairs.

The results suggest that the highest performance is achieved for the English-Maltese language pair, however, it is not comparable to the remaining results as they are based on only 987 term pairs from the *EuroVoc* thesaurus (covering mostly location and organisation named entities, which explains the relatively high recall).

The evaluation results for English as the source language shows that Italic languages (e.g., French and Romanian) achieve the highest results, followed by Slavic and Baltic languages. It is interesting to note that although English is a Germanic language, the results show that Germanic languages achieved considerably worse results than languages from other language families. However, the worse results are achieved with Finno-Ugric (or Uralic) languages.

An important aspect taken into account when designing the mapper was the mapping speed. For the evaluation in Table 13 the mapper required in average 86.8 minutes (which is a speed of 8,868 term pairs per second) for one language pair on an 8 thread (4 core) Windows computer. The speed can be significantly improved by limiting the number of translation and transliteration candidates retrieved from the probabilistic dictionary and the character-based

SMT module. The mapper requires in average less than 7 minutes for a language pair if no linguistic resources are used.

3.2.5.2. Manual Evaluation

The automatic evaluation was performed using terms in their base forms. However, in written documents terms can be found in many different inflected forms (especially for morphologically richer languages). The manual evaluation, therefore, has three goals: 1) to show the methods applicability on Web crawled comparable corpora 2) to show the methods performance in under-resourced conditions (e.g., the medical domain, which is out-of-domain for the *DGT-TM* corpus), and 3) to show that the method can be applied for morphologically rich languages. The manual evaluation was performed for the Latvian-English language pair and for terms in the medical domain.

Following the term mapping workflow proposed by Pinnis et al. (2012), two monolingual corpora were collected from the Web using the *Focussed Monolingual Crawler* (Mastropavlos & Papavassiliou, 2011). The acquired corpora (12,697 Latvian and 21,900 English documents) were then aligned in document level with the *DictMetric* (Su and Babych, 2012) comparability metric (59,600 document pairs were produced). The terms were then tagged in the monolingual documents with *TWSC* (Pinnis et al., 2012). The term tagging step produced a total of 198,401 unique Latvian and 352,934 unique English term candidates. The benefits of document level pre-alignment are evident when considering the full search space for the mapper without document alignment. In order to map 70 billion term pairs the mapper would require over 91 days to complete (using direct linguistic resources). With document alignments the required time can be reduced to less than 2 days.

Finally, terms were bilingually mapped in the 59,600 document pairs. A maximum of three transliteration and translation candidates were retrieved for each token of a term. A total of 24,804 term pairs were produced above a threshold of 0.6 (for each source term only the target language term with the highest confidence score was returned). 1000 randomly selected term pairs were manually evaluated and the results are given in Table 14. The results of the method by Ştefănescu (2012) is given for comparison. It produced on the same data set 2,330 term pairs above a threshold of 0.5. That is, more than ten times less term pairs.

Table 14. Manual evaluation results on the medical domain Latvian-English comparable corpus

Threshold	All terms		Multi-word terms		Single-word terms	
	Pairs	Precision	Pairs	Precision	Pairs	Precision
Author' s method (random 1000/24,804 term pairs):						
1.0	17	88.2%	0	-	17	88.2%
0.9	601	91.3%	111	85.6%	490	92.7%
0.8	724	85.6%	160	73.8%	564	89.0%
0.7	880	74.8%	203	65.0%	677	77.7%
0.6	1000	66.6%	267	50.6%	733	72.4%
Ştefănescu (2012) (random 1000/2,330 term pairs):						
1.0	25	84.0%	2	0.0%	23	91.3%
0.9	44	90.9%	7	71.4%	37	94.6%
0.8	88	93.2%	12	83.3%	76	94.7%
0.7	186	87.6%	46	65.2%	140	95.0%
0.6	387	73.6%	173	49.7%	214	93.0%
0.5	1000	44.8%	697	25.1%	303	90.1%

The results suggest that the author's method performs significantly better for multi-word term mapping, which is the main goal of this method. It is also evident that the majority of true positives are scored with a mapping score of over 0.8.

Another important question left to answer is whether the mapper finds term pairs that are unknown to the linguistic resources integrated in the mapper. The mapping method is only useful if it is able to identify *out-of-vocabulary* (OOV) term pairs. Therefore, the 1000 randomly selected term pairs from the manual evaluation were looked up in the probabilistic dictionary (for the 733 single-word terms) and in a translation model of an SMT system (for the 267 multi-word terms) that was trained on the same parallel corpus from which the probabilistic dictionary was created. The results of the analysis in comparison with the method proposed by Ştefănescu (2012) are given in Table 15.

Table 15 shows that 76.3% of all multi-word term pairs, which were evaluated as "*correct*" during the manual evaluation, could not be found in the translation model of the SMT system. The results also suggest that the probabilistic dictionary introduces mapping errors as 24.75% of the wrongly mapped single-word term pairs were present in the dictionary.

Table 15. OOV analysis of randomly selected Latvian-English term pairs

	Single-word term pairs in the probabilistic dictionary		Multi-word term pairs in the *Moses* phrase table	
	Correct	Wrong	Correct	Wrong
Author' s method:				
Source term OOV rate	13.94%	75.25%	76.30%	97.73%
Target term OOV rate	14.50%	75.66%	75.19%	97.73%
Term pair OOV rate	13.94%	75.25%	76.30%	97.73%
Ştefănescu (2012):				
Source term OOV rate	09.72%	76.00%	63.58%	99.58%
Target term OOV rate	12.09%	80.00%	62.86%	99.62%
Term pair OOV rate	12.09%	80.00%	62.86%	99.62%

3.3. Filtered Probabilistic Dictionaries and Invalid Mapping Dictionaries

3.3.1. Filtering of Probabilistic Dictionaries

Probabilistic dictionaries that are extracted from parallel corpora using automated methods, as explained earlier, contain a lot of noise (wrong alignments, partial alignments, etc.). In order to use a probabilistic dictionary for term mapping purposes, it is advisable to filter the dictionary in order to minimise the noise contained within it. This section describes a method for filtering of probabilistic dictionaries and identifying word pairs, which are similarly written, but are not translation equivalents. Such invalid translation equivalence pairs can be used by *MPAligner* as a linguistic resource that allows detecting incorrectly aligned term constituents.

The main idea of the probabilistic dictionary filtering method using transliteration systems is that when simply applying fixed thresholds we filter out many good translation equivalents from the probabilistic dictionaries, however, identification of translation equivalents that are reciprocal transliterations may allow retaining equivalents that would otherwise be filtered out. In this approach we are also analysing how far in terms of filtering we can get by applying language-specific alphabet filters. The method filters dictionary entries using the following 7 steps:

- The first step performs structural validation of dictionary entries in order to remove obvious noise. At first, we remove all entries that contain invalid character sequences on either source or target side. Character sequences are considered invalid if according to the Unicode[21] character table they contain control symbols, surrogate symbols, or only whitespace symbols. This step also identifies mismatching character sequences by comparing the source and target sides of a dictionary entry. At first it verifies that the source and target token letters are equally capitalised (with an exception of the first letter, which in some languages, e.g., for nouns in German or days of a week in English, is capitalised). Further, it verifies whether the letters contained in the source and target sides belong to the source and target language alphabets and whether both tokens contain equal numbers of digits, punctuation marks, and symbols, and whether they are located in similar positions in the source and target words. As the *Giza++* probabilistic dictionaries are statistical representations of token alignments in a parallel corpus, the alignments contain also easily detectable mistakes, such as, words paired with punctuations, incorrectly tokenized strings paired with words, etc. It is possible to easily filter out such obvious mistakes in the probabilistic dictionaries by applying the character-based validation rules on the source and target language words.

[21] For more information about Unicode refer to the http://www.unicode.org/ Web site.

- The second step identifies dictionary entries that are transliterations. Two different transliteration methods are applied (see section 3.4 for more details): 1) the language independent (however, fixed to the Latin, Greek, and Cyrillic alphabets) rule-based transliteration using Romanisation rules, and 2) the character-based SMT transliteration. While the first transliteration method is fast, it is not able to capture morphological variations in different languages and it treats each character independently of the context. The second method takes context (character n-grams) into account and is able to transliterate words not only into English, but also into other languages, thus transliterated word identification can be performed bi-directionally (from source to target and from target to source languages). To identify transliterated words, the transliterations (e.g., the source word transliterated into the target language) are compared with the other side's word (e.g., the target language word) using the Levenshtein distance-based string similarity metric described in section 3.2.3.1. If the maximum similarity score using any of the transliteration methods and transliteration directions (source-to-target or target-to-source) is higher than 0.7 (identified as an acceptable threshold through empirical analysis) and the source and target words are not equal (because such pairs are often wrong language pairs), we consider the dictionary entry as transliterated and we pass it through to the filtered dictionary (the further filtering steps are skipped).
- In the third step the remaining pairs are analysed using reference corpora based IDF scores (Spärck Jones, 1972) of the source and target words. All pairs that have a difference of word IDF scores greater than 0.9 (also empirically identified) are removed. Such pairs often indicate of functional word (or stopword) miss-alignment with content words (e.g., in the probabilistic dictionaries created by *Giza++* the English "*a*" is usually paired with almost every token of the other language and the IDF-based filter reliably removes such entries).
- In the fourth step the method applies a translation probability value threshold that is differentiated for (source language) words that were already containing transliteration pairs (i.e., if a dictionary entry containing the source word was identified as a transliteration, then all other translation candidates for the source word are required to have a high probability in order to be accepted as translation equivalents).
- Then, the method removes all pairs that partially contain transliterations. For instance, consider the dictionary entry "*monopoly*" (in English) and "*monopols*" (in Latvian). The entry is a transliteration, thus, "*monopolsituācijā*" (translated as "*in the case of a monopoly*") would be filtered out as it contains the whole transliterated part.

- The method applies also several heuristic filters that have shown to remove further noise (e.g., rare words miss-aligned with a probability of one if a source word already contains multiple translation hypotheses, equal source and target words if the source word already contains multiple translation hypotheses, etc.).
- Finally, the pairs that have passed all filter tests are written to the filtered dictionary.

Examples of dictionary entries that were identified using the different filtering steps from the English-Latvian *Giza++* dictionary are given in Table 16.

Table 16. English-Latvian dictionary entries identified according to different filtering steps

Source Token	**Target Token**	***Giza++* Probability**	**Filtering Step**
.	*94/65/ek.*	0.50	Structural validation (1) –- wrong entries
standards	*standarts*	0.02	Transliteration identification (2) –- correct entries
a	*aprobēt*	0.50	IDF score-based filter (3) –- wrong entries
proven	*gazprom*	0.08	Threshold filter (4) –- wrong entries
regulatory	*energoregulatora*	0.50	Partial containment and transliteration filter (5) –- wrong entries
navigational	*dodamos*	1.00	Heuristic filters (6) –- wrong entries

3.3.2. Creation of Invalid Mapping Dictionaries

In order to create the invalid mapping dictionary for *MPAligner*, the filtered dictionary is processed one more time. This time, words from one language are compared with all words from the other language using the Levenshtein distance-based similarity metric without any transliteration. The pairs that have high similarity, but are not defined as translation entries within the filtered dictionary, are included in the invalid mapping dictionary. For instance, "*pants*" in English and "*pants*" in Latvian (translated as "*article*" or "*paragraph*") have a similarity score of 1.0. As such entries would result in obvious misalignment by the term mapper, the inclusion of such similarly written words in the invalid mapping dictionary allows us to reliably filter possible invalid source and target token pairs when performing term mapping.

3.3.3. Evaluation of the Filtered Dictionaries

The probabilistic dictionary filtering method has been evaluated in co-operation with other co-authors of the publication Aker et al. (2014b). The transliteration-based filtering method has been compared with two methods developed by Ahmet Aker: 1) a purely statistical approach, which was first implemented by Munteanu & Marcu (2006) and uses Log Likelihood Ratio (LLR) (Dunning, 1993) in order to test whether two words can be considered translation equivalents or not, and 2) a pivot-based approach, which uses an intermediate language in order to validate translation entries in a dictionary. That is, if we have to filter, for instance, the English-Latvian dictionary, we can use an English-German and German-Latvian dictionary in order to test whether through German as the intermediate language in this example we can acquire the same Latvian translation equivalents as by using the direct English-Latvian dictionary. Each of the methods produces a different set of filtered dictionary entries, therefore, the evaluation analyses individually the entries excluded by all methods, included by just one method, included by two methods, or included by all methods. The evaluation was performed by two language specialists (more precisely, professional translators) per language pair (English-Latvian and English-German dictionaries were evaluated) using a Web-based evaluation platform developed by Monica Lestari Paramita that for 40 randomly selected dictionary entries per test set asked whether it was a complete translation equivalence, whether it was a containment (e.g., for single stem words and compound words). The evaluation results for English-Latvian in Table 17 show that out of all individual methods the author's method (the transliteration-based method) allows acquiring dictionary of higher precision than with the other methods. However, when looking at the results of the various intersections of the different methods, it is evident that much higher precision can be reached by combining the methods. Similar results were identified also for the English-German language pair (Aker et al., 2014b).

Table 17. Results of the English-Latvian manual evaluation by two annotators. The precision figure in each row is computed by dividing the figure in column *Eq.* with the sum of the figures of the columns *Eq.* to *Wrong* of that row (Aker et al., 2014b).

Set name	**All ratings**				**Complete agreement ratings**			
	Eq.	**Cont.**	**Wrong**	**Precision**	**Eq.**	**Cont.**	**Wrong**	**Precision**
All	71	2	7	88.75%	33	0	1	97.06%
Transliteration + LLR	62	4	14	77.50%	25	0	3	89.29%
Transliteration + Pivot	56	7	17	70.00%	25	1	6	78.13%
LLR + Pivot	55	5	20	68.75%	25	2	7	73.53%
Transliteration	49	4	27	61.25%	21	0	11	65.63%
Pivot	34	11	35	42.50%	14	2	14	46.67%
LLR	34	4	42	42.50%	15	1	19	42.86%
Original	5	3	72	6.25%	0	0	32	0.00%

Furthermore, analysis performed by Aker at al. (2014b) has shown that the different methods when applied separately miss out many good translation equivalents. In the term mapping experiments performed by the author only the transliteration-based filtering method's produced dictionaries have been used. Therefore, an important direction for future work is the combination of different filtering methods as well as concatenation of the different dictionaries acquired with the different filtering methods.

3.4. Character-based SMT Transliteration Systems

Transliteration, which is the process of representing words from one language using the writing system of another language (Arbabi et al., 1994; Pouliquen et al., 2005), is a typical method for the translation of named entities and technical terms (Knight & Graehl, 1997) (often applying grapheme-to-phoneme and phoneme-to-grapheme transformation rules in the translation process). Creation of a rule-based system can be very time consuming, and therefore an alternative is to build supervised machine learning based systems (e.g., using statistical machine translation technology; Kirschenbaum, & Wintner, 2010). However, to build supervised transliteration models that could be integrated in machine translation systems, we require a transliteration dictionary. Although there are multilingual named entity dictionaries, e.g., JRC Names (Steinberger & Pouliquen, 2011), HeiNER (Wentland et al., 2008), and others, available, they are not directly applicable for development of transliteration models, because named entities often contain words which are not transliterated. For example, the organisation name "*European Union*" when translated into Latvian ("*Eiropas Savienība*") contains a transliterated and a translated word.

Therefore, to address the necessity of transliteration dictionaries, the following subsections will present a method for transliteration dictionary extraction using a bootstrapping process from existing dictionaries, e.g., automatically extracted probabilistic dictionaries (Aker et al., 2004b) or manually created dictionaries containing words in their canonical (or lemma) forms. The author describes and analyses a large multilingual transliteration dictionary extracted from probabilistic dictionaries for 24 European languages (23 language pairs with English as a source language).

3.4.1. Bootstrapping Method

To create a transliteration dictionary, the author starts with existing *Giza++* (Och & Ney, 2003) probabilistic dictionaries extracted from the *DGT-TM* (Steinberger et al., 2012; for official languages of the European Union) and *MultiUN* (Eisele & Chen, 2010; for English-Russian) parallel corpora. The transliteration dictionaries are bootstrapped from the probabilistic dictionaries in two (or more) steps:

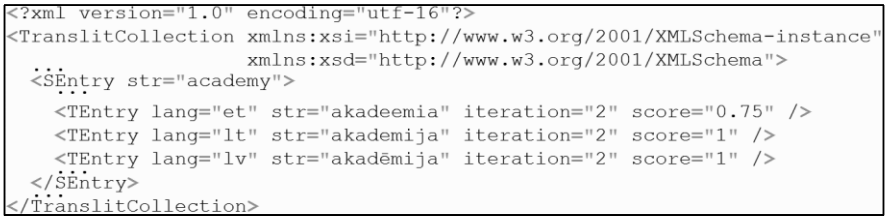
1) In the first step, we apply Romanisation rules (Knight & Graehl, 1997) to all non-English words. The Romanisation rules have been specifically developed for the term mapper *MPAligner* (Pinnis, 2013) and define one-to-one (e.g., the Greek "*β*" and the Bulgarian "*б*" correspond to the English letter "*b*", etc.), one-to-many (e.g., the Greek "*φ*" corresponds to the English "*th*", the Russian "*ч*" corresponds to the English "*ch*", etc.), and one-to-none (e.g., the Russian letters "*ъ*" and "*ь*" are deleted) correspondences of letters from a non-English alphabet into the English alphabet. Then, we compare the English words to the Romanised words with the Levenshtein distance-based similarity metric, which was introduced in section 3.2.3.1 - equation (2). Word pairs exceeding an

empirically set threshold of *0.7* are extracted as reciprocal transliterations for the further bootstrapping steps.

2) In the second step (and further steps if necessary), we use the transliterations identified in the previous step to build character-based statistical machine translation (SMT) systems using the *Moses* SMT toolkit (Koehn et al., 2007). The SMT systems are used to transliterate entries of the initial dictionary. For the experiments presented in this paper, we use the top five SMT transliterations for each non-English word. New transliteration pairs are identified using the same similarity function from Equation 1.

3.4.2. Data Formats

The extracted multilingual transliteration dictionary is stored in an XML document. The dictionary consists of source entries in English (the "*SEntry*" tag in Figure 14). For each source entry, the dictionary provides a list of transliterations in target languages (the "*TEntry*" tags). For each transliteration entry, the dictionary provides the number of the bootstrapping iteration in which the transliteration pair has been identified and the bootstrapping method's confidence score (the Levenshtein distance based similarity). This provides traceability for the data within the transliteration dictionary and allows fine-tuning the dictionary for different application purposes where quality and quantity requirements differ.



```
<?xml version="1.0" encoding="utf-16"?>
<TranslitCollection xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                    xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  ...
  <SEntry str="academy">
    ...
    <TEntry lang="et" str="akadeemia" iteration="2" score="0.75" />
    <TEntry lang="lt" str="akademija" iteration="2" score="1" />
    <TEntry lang="lv" str="akadēmija" iteration="2" score="1" />
    ...
  </SEntry>
  ...
</TranslitCollection>
```

Figure 14. Example of the XML format of the multilingual transliteration dictionary

3.4.3. Statistics of the Multilingual Transliteration Dictionary

To create the multilingual transliteration dictionary, the author performed two bootstrapping iterations. The first bootstrapping iteration produced a total of 598,807 transliteration pairs for 82,454 English words across all 23 language pairs. The second iteration resulted in 1,246,908 transliteration pairs for 104,803 English words.

The quantitative results for English-Latvian (see Table 18) show a significant increase in new transliteration pairs extracted in the second bootstrapping iteration. The increase can be explained by the SMT-based transliteration method's ability to deal with inflectional characteristics of different languages. That is, the SMT translation model learns from parallel data (transliteration equivalents identified in previous steps) to translate language specific word prefixes and suffixes from one language into another. As the rule-based method is not capable of performing such language specific transformations, it cannot identify many good transliteration equivalents.

Table 18. Statistics of new English-Latvian transliteration pairs identified in five bootstrapping iterations

Iteration	New pairs	% increase	New English words	% increase
1	30,879	-	15,598	-
2	41,347	134%	11,992	77%
3	1,704	2%	500	2%
4	469	1%	125	0%
5	961	1%	255	1%
Total	**72,226**		**28,470**	

Table 18 also shows that for English-Latvian, the first two out of five total iterations allow acquiring approximately 97% of all extracted English words. Because the initial dictionaries are exhaustive resources (i.e., they contain a fixed number of entries out of which only a certain amount are potential transliterations) and the first two iterations are able to identify the majority of transliteration equivalents, all further iterations are less productive. The 97% comprise approximately 20% of all 134,146 unique English words present in the initial probabilistic dictionary. Taking into account that English and Latvian are not closely related languages, this is a relatively large number.

As a result, only the first two bootstrapping iterations were performed for the multilingual transliteration dictionary. The statistics of the dictionary for all 23 language pairs with English as the source language are given in Table 19. The extracted pair count for Croatian-English is lower due to a smaller size of the initial probabilistic dictionary.

Table 19. Statistics of the multilingual transliteration dictionary after merging first and second iteration data (languages are given in the ISO 639-1 format)

Target language	Unique English words	Transliteration pairs	Target language	Unique English words	Transliteration pairs
bg	17,567	37,901	lt	25,258	66,243
cs	28,366	58,931	lv	27,590	72,186
da	27,321	51,383	mt	21,217	62,428
de	23,862	41,560	nl	23,673	36,741
el	15,513	31,273	pl	29,723	62,313
es	35,030	64,480	pt	37,666	67,473
et	22,188	48,113	ro	27,295	58,531
fi	18,180	33,860	ru	30,835	71,482
fr	33,367	59,390	sk	31,536	77,607
hr	7,368	14,965	sl	30,364	66,365
hu	26,942	53,664	sv	28,692	53,676
it	31,147	56,343			

A visual example of an entry in the transliteration dictionary for the Baltic languages is given in Figure 15. The light grey to black connectors between English and the target languages indicate low (grey) to high (black) confidence scores assigned to the transliteration pairs by the bootstrapping method.

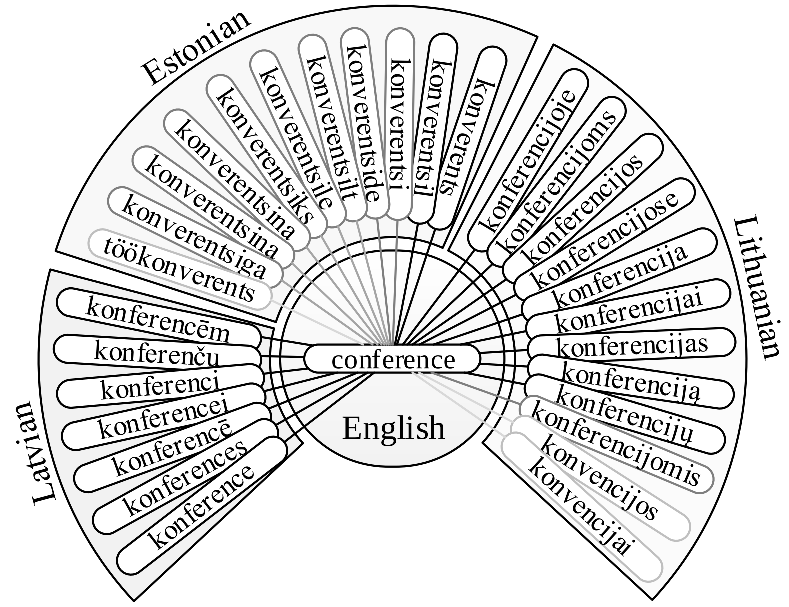


Figure 15. Transliterations of the English word "*conference*" in Estonian, Latvian, and Lithuanian identified in the *Giza++* dictionaries extracted from the *DGT-TM* corpus

3.4.4. Evaluation

The evaluation of the multilingual transliteration dictionary consists of two parts: 1) manual evaluation for the English-Latvian language pair, and 2) automatic evaluation of the transliteration dictionary in an SMT-based transliteration task for 23 language pairs.

3.4.4.1. Manual Evaluation

Manual evaluation of the multilingual transliteration dictionary has been performed for the English-Latvian language pair. The author executed a total of five bootstrapping iterations and extracted only newly identified transliteration pairs from each iteration (the quantitative statistics are given in Table 18). Further, 100 transliteration pairs were randomly selected from the newly extracted transliteration pairs for manual evaluation. A transliteration pair in the manual evaluation is considered correct if:

1) The pair consists of words that are reciprocal translations.
2) The pair qualifies to be a transliteration pair. That is, it has to be possible to acquire from the source word the target word (and vice versa) by performing alphabet specific letter transformations (e.g., the Latvian "*č*" can correspond to the English "*ch*", the Greek "*ρ*" can correspond to the English "*r*", etc.) and language specific prefix and suffix transformations (e.g., the English suffix "*ation*" may correspond to the Latvian "*ācija*", Italian "*azione*", the Bulgarian "*ация*", and other suffixes in many different inflected forms).

The evaluation results are given in Figure 16. The results show that the precision of the transliteration dictionary for English-Latvian is over 90% after the first bootstrapping iteration. Taking into account that the initial probabilistic dictionaries are of very low quality (Aker et al., 2014b), this is a very good result. The figure also shows that the precision of the newly extracted transliteration pairs decreases with each new bootstrapping iteration. Although this was to be expected, the thresholds for different bootstrapping iterations could be differentiated in order to achieve a stable precision of over 90%.

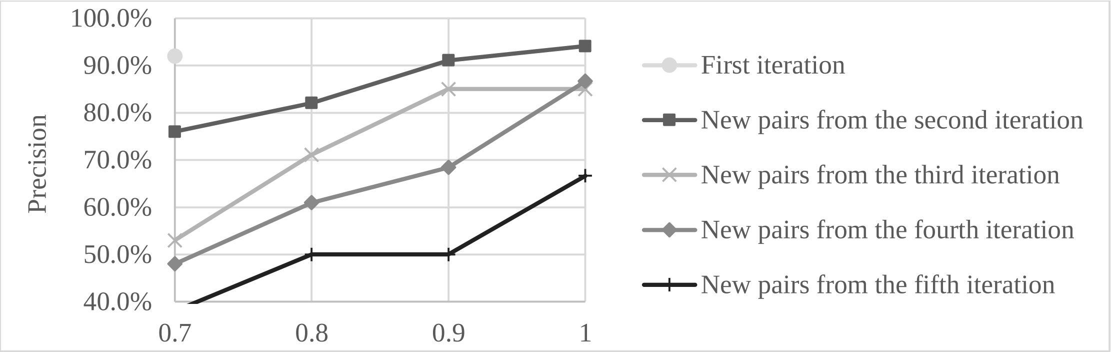


Figure 16. Manual evaluation results for 100 randomly selected transliteration pairs from the English-Latvian transliteration dictionary from different bootstrapping iterations

3.4.4.2. Automatic Evaluation in an SMT-based Transliteration Task

Transliteration dictionaries have shown to be beneficial when integrated into SMT systems (Kirschenbaum & Wintner, 2010). However, they are also used for development of machine transliteration systems (Knight & Graehl, 1997) (e.g., character-based SMT; Finch & Sumita, 2008). In the paper by Pinnis (2013) the author has shown that such systems can be used for cross-lingual term mapping in comparable corpora. In this section, the extracted dictionaries are evaluated in SMT-based transliteration tasks.

After the second bootstrapping iteration, the source-to-English transliteration data was randomly split in 10 data folds. In each data fold, eight parts were used for training, one – for tuning, and one – for evaluation. Then, 10-fold cross validation was performed by measuring character level SMT quality using SMT evaluation metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). The results are given in Table 20. The results are shown with a 99% confidence interval.

Table 20. Character level 10-fold cross-validation results for character-based SMT transliteration (languages are given in the ISO 639-1 format)

Language pair	NIST	BLEU	Language pair	NIST	BLEU
bg-en	11.48±0.04	90.11±0.23	lt-en	11.71±0.02	89.49±0.15
cs-en	12.07±0.03	90.46±0.14	lv-en	11.87±0.03	89.78±0.22
da-en	11.92±0.04	90.37±0.17	mt-en	11.63±0.04	90.35±0.21
de-en	11.89±0.02	90.30±0.17	nl-en	11.68±0.07	89.42±0.29
el-en	10.94±0.04	85.29±0.25	pl-en	11.96±0.02	89.85±0.18
es-en	11.84±0.05	88.20±0.31	pt-en	11.99±0.05	88.83±0.23
et-en	12.13±0.03	91.93±0.20	ro-en	11.67±0.03	88.67±0.13
fi-en	12.10±0.05	92.54±0.47	ru-en	11.23±0.04	83.27±0.18
fr-en	11.99±0.06	88.39±0.27	sk-en	12.15±0.05	90.84±0.18
hr-en	10.53±0.07	87.60±0.31	sl-en	12.02±0.03	89.71±0.12
hu-en	12.17±0.02	91.10±0.13	sv-en	11.92±0.02	89.91±0.14
it-en	11.51±0.05	86.78±0.28			

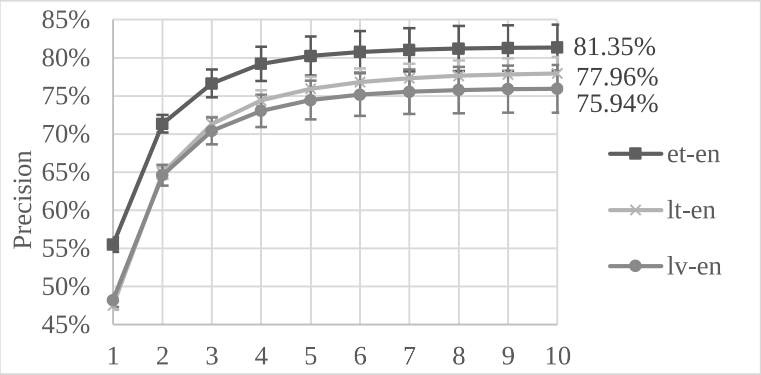


Figure 17. 10-fold cross-validation results for the top *N* SMT transliteration equivalents for Baltic Languages (languages are given in the ISO 639-1 format)

Depending on usage scenarios, an SMT system can be asked to produce one (e.g., for integration of transliteration in machine translation) or many (e.g., for cross-lingual term mapping) transliteration equivalents. Figure 17 shows the precision for up to top ten SMT generated transliteration equivalents for Baltic languages (results for other language pairs are given in Table 21) when transliterated into English. Because of different inflectional forms in transliteration pairs (e.g., singular vs. plural forms, verbs in different tenses, etc.), the results show a significant increase in precision for the top two to top four transliteration equivalents over the results of the top one.

Another reason for the lower precision for the top one transliteration is the ambiguity of different character sequence transformations, which cannot be predicted by analysis of the surrounding context (letters to the left and to the right). For instance, the differences between writing paradigms in American English and British English allow the Latvian "*organizācija*" to be transliterated as "*organi**z**ation*" or "*organi**s**ation*". Another ambiguous (or non-predictive) example is, for instance, the Latvian "*Kuba*" transliterated in English. It can be either the country "*Cuba*" or a three-dimensional figure "*Cube*". Obviously, the top one transliteration will not always be the expected transliteration because of such ambiguities. A list of the most frequent top one transliteration errors for Latvian-English is given in Table 22. Note that the table shows also ambiguous examples, which are not actual errors, e.g., singular vs. plural forms, different verb tenses, etc.

Table 21. 10-fold cross-validation results for the top 1, top 5, and top 10 SMT transliteration equivalents (languages are given in the ISO 639-1 format)

Language pair	**Top 1**	**Top 5**	**Top 10**	**Language pair**	**Top 1**	**Top 5**	**Top 10**
bg-en	51.36±0.6%	78.20±2.2%	79.87±2.8%	lt-en	47.52±0.6%	75.94±1.6%	77.96±2.1%
cs-en	49.93±0.7%	75.15±1.9%	76.13±2.2%	lv-en	48.21±0.9%	74.45±2.5%	75.94±3.1%
da-en	47.37±0.8%	74.94±1.2%	76.38±1.3%	mt-en	53.68±0.9%	75.95±3.2%	76.94±3.6%
de-en	46.01±0.9%	77.02±2.1%	79.12±2.9%	nl-en	45.31±1.0%	63.09±2.2%	63.67±2.3%
el-en	41.89±0.8%	66.10±1.8%	68.06±2.1%	pl-en	47.45±0.5%	75.81±2.0%	77.57±2.6%
es-en	43.94±0.6%	66.42±1.5%	67.37±1.8%	pt-en	46.33±0.8%	66.95±2.7%	67.69±3.1%
et-en	55.49±1.0%	80.24±2.6%	81.35±3.0%	ro-en	44.48±0.6%	73.88±1.9%	75.67±2.4%
fi-en	59.89±1.3%	81.70±1.6%	82.59±1.7%	ru-en	37.95±0.5%	61.68±1.5%	63.79±1.9%
fr-en	42.29±1.0%	63.71±3.3%	64.42±3.6%	sk-en	51.30±0.6%	78.08±1.4%	79.63±2.0%
hr-en	43.12±1.5%	66.02±4.6%	68.38±5.7%	sl-en	48.17±0.5%	76.91±2.5%	78.77±3.1%
hu-en	51.94±0.8%	76.50±2.7%	77.37±3.1%	sv-en	46.64±0.8%	75.55±1.4%	77.37±1.8%
it-en	37.71±0.9%	68.34±2.6%	71.44±3.6%				

Table 22. 15 most frequent character level errors for the Latvian-English SMT-based transliteration system (In the table: Insertions – *Ins.*, Deletions – *Del.*, Substitutions – *Sub.*).

No.	**Error**	**% of all**	**Latvian (in different inflected forms)**	**English**	
				Expected	**Generated**
1	Ins. / Del. *s*	19.79%	zonā	zone**[s]**	zone
			organismus	organism	organism**[s]**
2	Ins. / Del. *e*	6.42%	krese	cress	cress**[e]**
			validēt	validat**[e]**	validat
3	Ins. / Del. *a*	3.82%	komponentā	component	component**[a]**
			memorandu	memorand**[a]**	memorand
4	Ins. / Del. -	3.39%	kvazistatiskas	quasi**[-]**static	quasistatic
			subklīniskas	subclinical	sub**[-]**clinical
5	Ins. / Del. *al*	3.29%	stratēģiskai	strategic	strategic**[al]**
			teorētiskām	theoretic**[al]**	theoretic
6	Sub. *z* ↔ *s*	3.27%	realizējis	reali**[z]**ed	reali**[s]**ed
			organizē	organi**[s]**e	organi**[z]**e
7	Ins. / Del. *o*	2.66%	luksemburga	luxemburg	luxemb**[o]**urg
			fosforu	phosphor**[o]**us	phosphorus
8	Ins. / Del. *d*	2.38%	koncentrētos	concentrate**[d]**	concentrate
			neitralizētu	neutralise	neutralise**[d]**
9	Ins. / Del. *h*	2.37%	homeopātiskas	homeopat**[h]**ic	homeopatic
			metrīta	metritis	met**[h]**ritis
10	Sub. *i* ↔ *y*	2.01%	iridovīrusa	[i]ridovirus	**[y]**ridovirus
			elektrolīts	electrol**[y]**te	electrol**[i]**te

Further, for the Latvian-English transliteration direction, Figure 18 depicts the SMT-based transliteration quality for systems trained on data from the first and second bootstrapping iterations. Although the manual evaluation results show that the overall quality of the data after the second iteration is lower, the SMT evaluation shows that the data from the second iteration allows achieving higher word level precision. The results show that the SMT system is able to build a more generalised translation model by using more data.

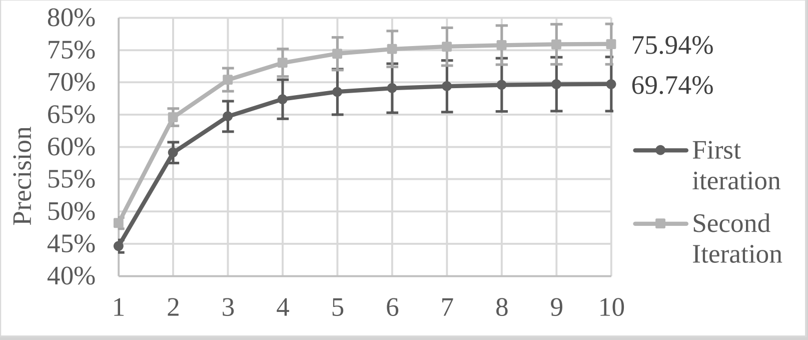


Figure 18. 10-fold cross-validation results for the top N SMT generated transliteration equivalents. The chart compares Latvian-English SMT-based transliteration systems trained on the transliteration dictionaries from the first and second bootstrapping iterations. The error bars indicate a 99% confidence interval.

3.5. *MPAligner* Applied in Practice

The term mapper *MPAligner* has been successfully applied in practice in the *TaaS* project where it has been used to perform cross-lingual term mapping in the *TaaS* platform's *Bilingual Term Extraction System* (*BiTES*) (TaaS, 2014a). The *BiTES* workflows for comparable corpora (depicted in Figure 19) are used for acquisition of bilingual terminology for the *Statistical Data Base* (*SDB*) of the *TaaS* platform (*TaaS*, 2014a).

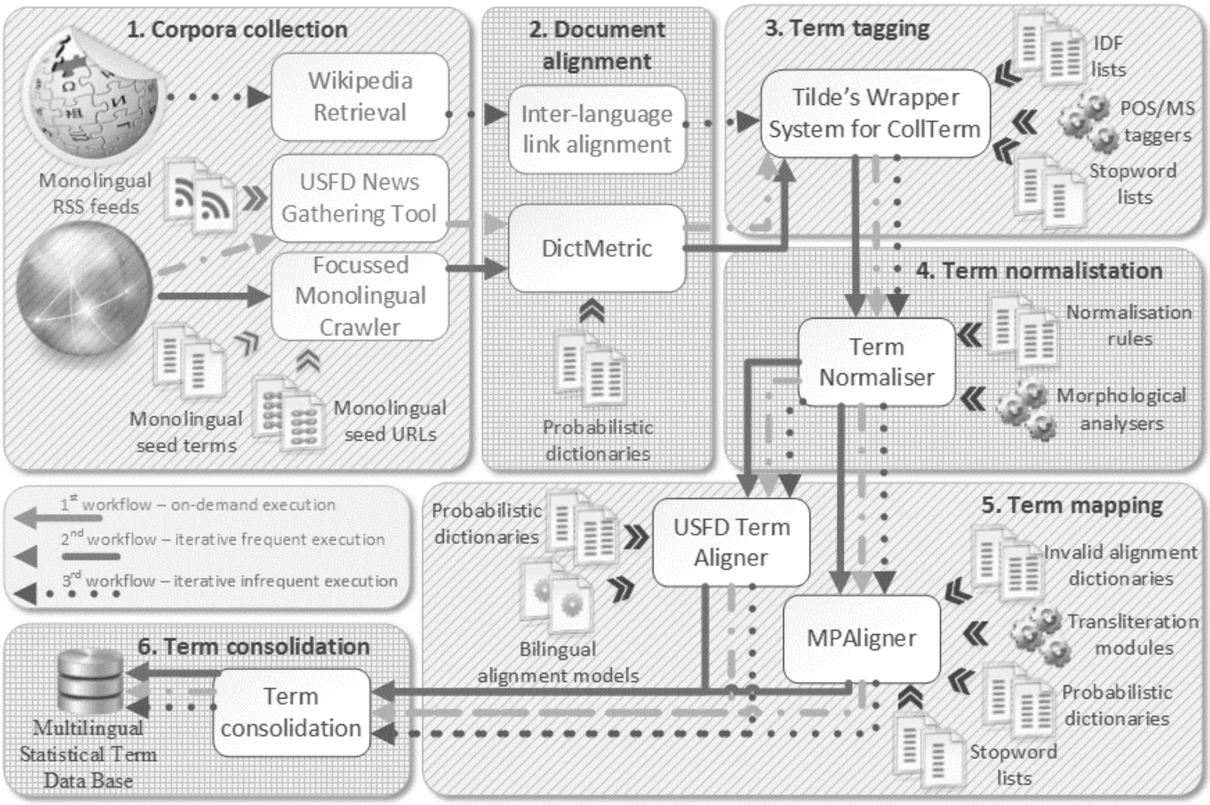


Figure 19. The design of the consolidated multilingual terminology acquisition workflows of the *TaaS Bilingual Term Extraction System* (Vasiļjevs et al., 2014b; *TaaS*, 2014a)

When performing bilingual term mapping on large corpora (i.e., tens or hundreds of thousands of document pairs) or when term mapping is performed iteratively, repeatedly or separately on multiple corpora, it is important for the solution that uses the term mapper's output to deal with redundancy in the mapped data. Redundancy in the term mapper's output may be: 1) duplicate term entries in term lists, and 2) the same terms, but in different inflected forms, paired together. Therefore, after cross-lingual term mapping with *MPAligner*, the bilingual term pairs are integrated into the *SDB* by simultaneously performing term pair morphological consolidation. Because for different languages different linguistic tools may be available (i.e., POS taggers, morphological analysers, lemmatisers, term normalisers, etc.), term consolidation is performed in three levels:

- For languages, for which lemmatisation of words is not available, however POS taggers can be used, terms are grouped together only by their inflected forms and POS sequences.
- For languages with lemmatisation support, terms are grouped by their lemmatised forms and POS sequences. This consolidation level ensures that for morphologically rich languages redundancy, which is caused by having numerous inflected forms of a single word, can be eliminated. However, this method can also group together inflected forms belonging to different terms. For example, the term candidates "*personālais dators*" and "*personāls dators*" from Figure 20 both have identical lemma sequences. This issue can be solved by the third level.
- For languages with term normalisation support, different inflected forms of terms are grouped by their normalised forms and the normalised form POS sequences. This method ensures that term inflected forms are correctly grouped together in the *SDB*.

The different consolidation levels are used in order to provide the most appropriate term translation equivalents for a term lookup query in the *TaaS* platform (*TaaS*, 2014a). If no translation equivalents are identified in the higher consolidation levels, the data from the lower levels is used, thus ensuring that the *SDB* provides as descriptive information for bilingual terms produced by *BiTES* (and the *MPAligner*) as possible.

Term translation lookup queries for terms integrated into the *SDB* can be organised also with the help of pivot languages. For instance, Figure 20 shows translation candidates for the Lithuanian term "*personalinis kompiuteris*" (transl. "*personal computer*") in Latvian that can be acquired using English as a pivot language.

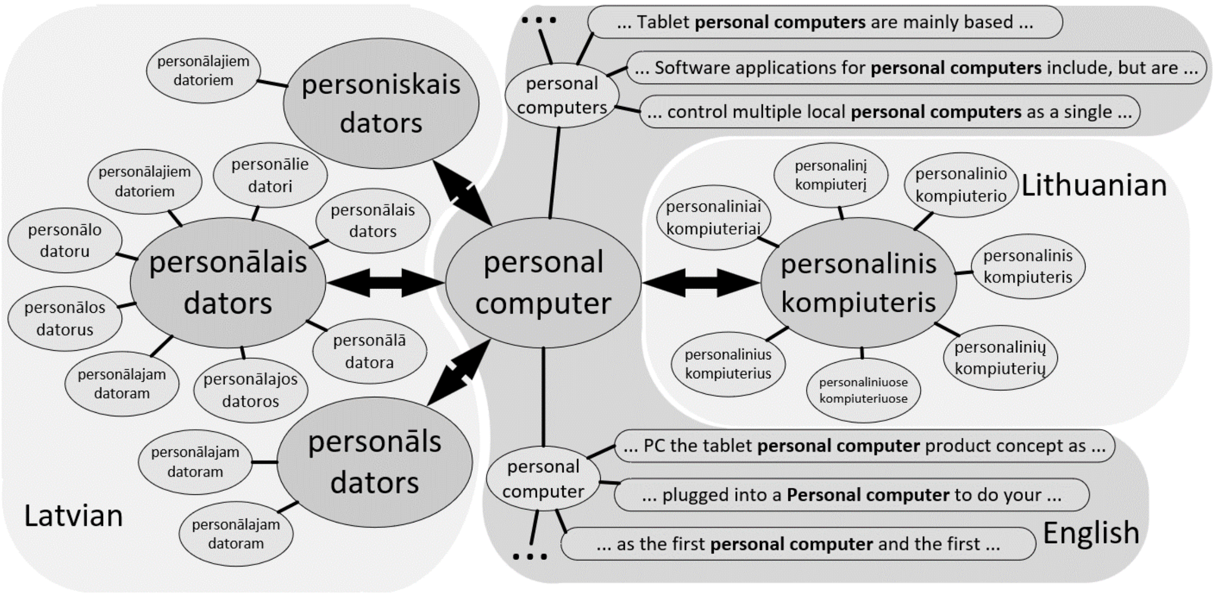


Figure 20. Visualised example of terminological data extracted with *MPAligner* and stored in the *TaaS Statistical Data Base (Vasiļjevs et al., 2014b)*

For the *TaaS* platform, *MPAligner* has been used to cross-lingually map terms in *Wikipedia*, Web news, focussed Web crawled, and parallel corpora. In total, over twenty million unique inflected form pairs of terms distributed over 45 subject fields were integrated into the *TaaS SDB* for 26 language pairs. Statistics for different languages are given in Figure 21. This resource serves as a valuable term translation candidate look-up source in the *TaaS* platform.

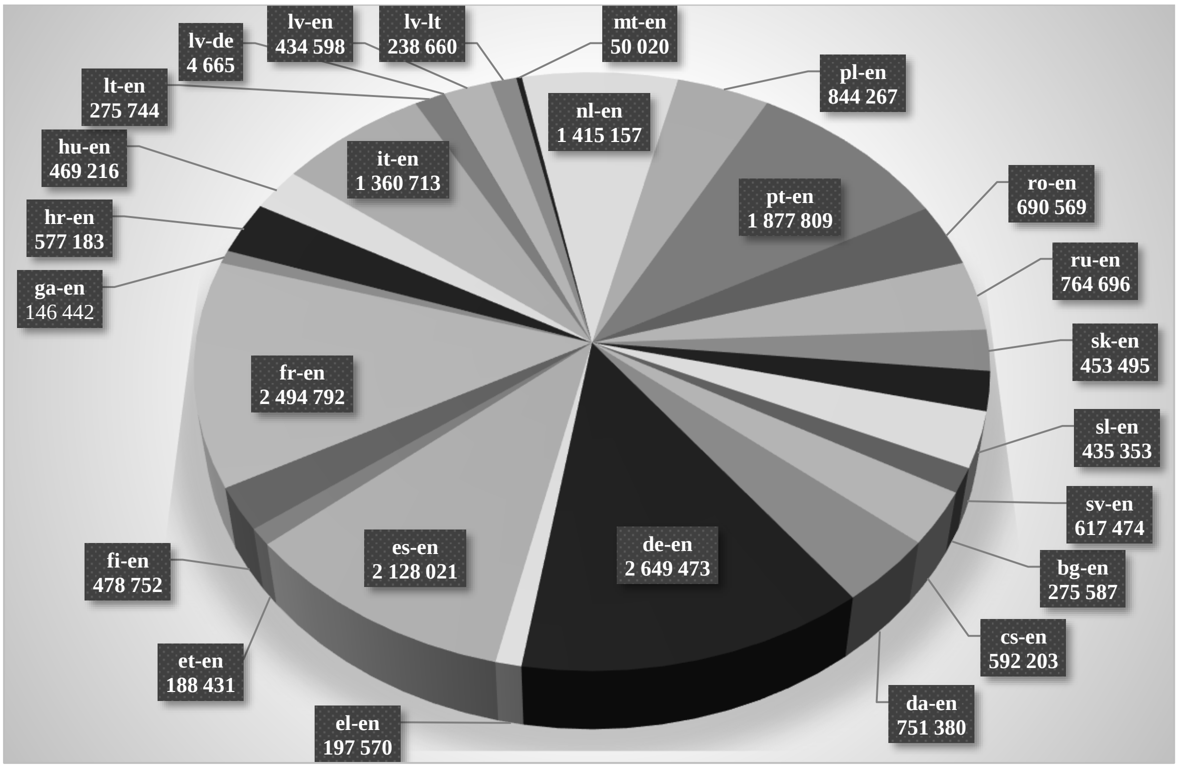


Figure 21. Unique inflected form pairs of terms integrated in the *TaaS SDB* (languages are given in the ISO 639-1 format)

3.6. Summary of Cross-Lingual Term Mapping

In this section, the author presented a new bilingual term mapping method (*MPAligner*) using maximised character alignment maps. The method has been designed to address multi-word term pair as well as compound term pair mapping for European Languages that are based on Latin, Greek and Cyrillic alphabets.

The method has been evaluated 1) automatically using the *EuroVoc* thesaurus for 23 language pairs, and 2) manually on terms mapped in a comparable corpus in the medical domain for the Latvian-English language pair, showing that the mapping method is suitable for handling noisy data collected from the Web. The evaluation also shows that up to 76.3% of the correctly mapped multi-word term pairs are out-of-vocabulary term pairs. The proposed term mapping method is able to find multi-word term pairs with a relatively high precision of up to 85.6%. It should, however, be noted that the scores depend on the corpus processed and may differ between language pairs as seen in the automatic evaluation.

An important resource that has been created with MPAligner (in combination with corpus collection, term identification, term normalisation, and domain classification methods of the *TaaS* platform's *Bilingual Term Extraction System*) in an effort that spanned for more than one year is the *Statistical Data Base* of the TaaS platform. *SDB* contains over 20 million pairs of inflected forms of terms in 25 languages. All pairs have been acquired with the help of *MPAligner*. To the best of the author's knowledge, this is the largest resource of multilingual terminology currently available.

The author also presented a transliteration-based method for filtering raw probabilistic dictionaries extracted from parallel data. The method has been evaluated in comparison with two other methods (LLR-based and pivot language-based) developed by Aker et al. (2014b) and it has shown to produce better quality (in terms of precision) results than the other methods. However, the relatively low recall of all three methods shows that significant improvements could be achieved by combining the three methods. The possible combination, as shown by the evaluation could improve the overall precision as the intersection of all three methods shows to be of much better quality than the quality of all other combinations.

Another method presented in this section was a bootstrapping method for the creation of a multilingual transliteration dictionary from existing probabilistic dictionaries. The multilingual transliteration dictionary generated by the author using probabilistic dictionaries extracted from the *DGT-TM* parallel corpus and the *MultiUN* parallel corpus covers 24 languages and contains a total of 1,246,908 transliteration pairs. To the best of the author's knowledge, the dictionary is the first publicly available multilingual transliteration dictionary. The evaluation has shown that the transliteration dictionary can be effectively applied in SMT-based transliteration tasks and also cross-lingual term mapping by integrating the transliteration systems into *MPAligner*.

The term mapping toolkit together with configuration and evaluation recipes is released under a non-commercial (free to use for scientific purposes) license. The toolkit can be downloaded from https://github.com/pmarcis/mp-aligner. The multi-lingual transliteration dictionary and the tools for creation of the transliteration dictionary as well as tools for filtering probabilistic dictionaries and creating invalid mapping dictionaries are freely downloadable from https://github.com/pmarcis/dict-filtering.

4. STATIC INTEGRATION OF TERMINOLOGY IN SMT SYSTEMS

As already noted in the introduction, terminology integration in SMT systems can be performed in two levels: 1) statically when training SMT systems, and 2) dynamically when translating documents using an already pre-trained SMT system. This section focusses on the research carried out on static terminology integration in SMT systems. The section is based on the author's contributions for the publications by Pinnis et al. (2014), Pinnis & Skadiņš (2012), Skadiņš et al. (2013), and the *TaaS* project's Deliverable D4.4 Terminology Integration in SMT (*TaaS*, 2014b).

All terminology integration experiments reported in this thesis have been performed using the *LetsMT* SMT platform (Vasiļjevs et al., 2012). Just to show the complexity of SMT system training, Figure 22 visually depicts an overall training process of a typical SMT system in the *LetsMT* platform, broken down in many different sub-processes. The figure also shows the processes that combined in workflows train the translation and language models of an SMT system. This section will particularly focus on methods that allow performing efficient domain adaptation of translation and language models using bilingual term collections.

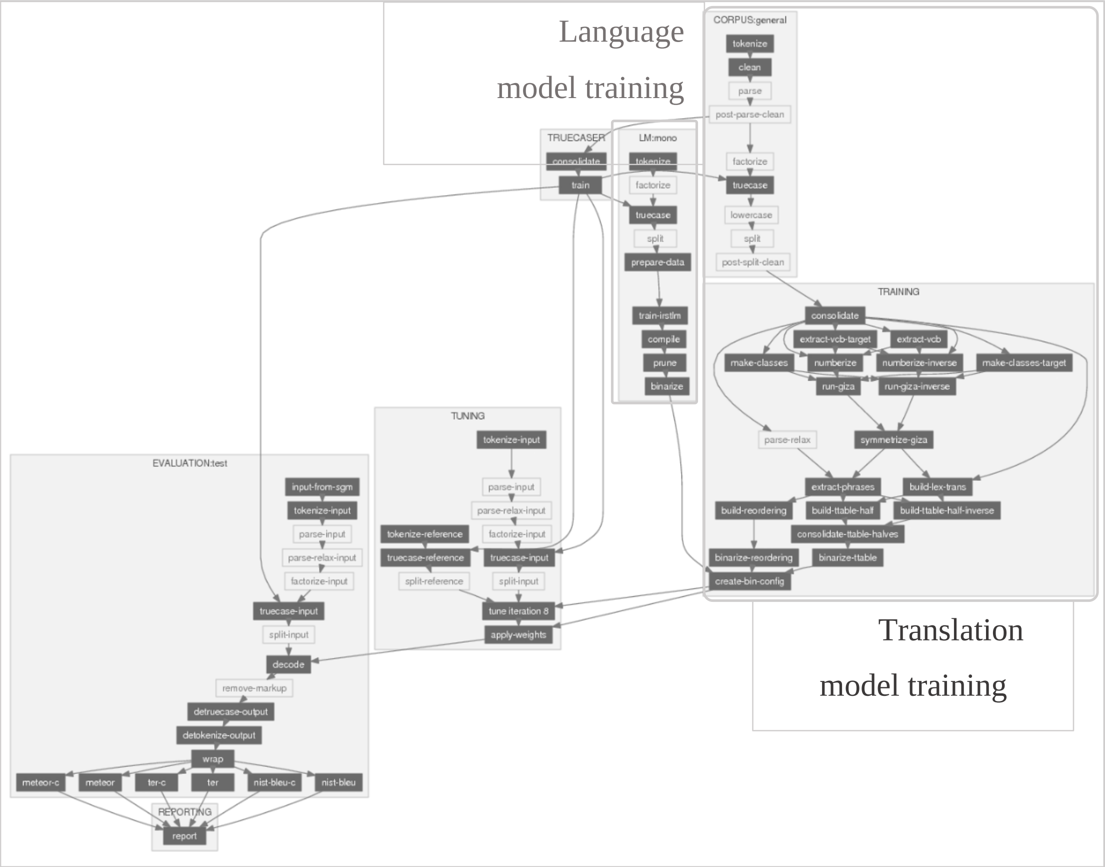


Figure 22. A typical SMT system training process in the *LetsMT* platform

Further, section 4.1 gives insight in related work on static terminology integration in SMT, section 4.3 describes a simple method for terminology integration in SMT (for both the translation models and the language models) that is a prerequisite for more complex static terminology integration methods. The simple method is followed by separately describing methods applied for terminology integration in SMT system translation models (section 4.4) and SMT system language models (section 4.5).

4.1. Related Work on Static Terminology Integration in SMT

There have been numerous research works reporting improvement of translation quality in terms of automatic machine translation evaluation after direct (by using in-domain term collections) and indirect (by tackling the broader challenge of domain adaptation using in-domain parallel or monolingual corpora) integration of terms and term phrases in SMT systems.

Significant research efforts have been spent on using in-domain parallel and monolingual corpora (that contain in-domain terminology) to perform SMT system translation model and language model adaptation to specific domains (to name but a few, Koehn & Schroeder (2007), Bertoldi & Federico (2009), Hildebrand et al. (2005), and many others). The usage of in-domain corpora in combination with out-of-domain corpora, however, is challenging. If all parallel data (in-domain and out-of-domain) is used to train a single translation model, the out-of-domain training data may overwhelm the in-domain data (Koehn & Schroeder, 2007). However, if just the in-domain corpora is used, the trained SMT system may fail generalising general language characteristics, and this can lead to poor translation quality (Thurmair, 2004). A domain specific SMT engine needs to capture the generalisations of an engine trained on large parallel corpora, yet not lose domain specificity. It was shown that to achieve this, the translation model of an SMT system can be trained on all available parallel data including out-of-domain data, however, instead of one language model, the SMT system should utilise two separate language models that are trained on in-domain and out-of-domain sets (Koehn & Schroeder, 2007; Lewis et al., 2010). Although SMT domain adaptation has been an active field in the machine translation research community, the majority of practical SMT applications rely solely on collecting big amounts of domain specific corpora. Moreover, there are not so many more advanced solutions, which would focus on special handling of terminology. It is assumed that training data will contain translations with terminology and correct terminology translation will be learned from the training data. However, it is not usually the case as training data, even if it is in the same domain, can contain contradicting terminology (e.g., industry specific synonyms, product-biased or customer-biased terminology, obsolete terminology, etc.).

Terminology integration has been also indirectly addressed by research on multi-word unit integration in SMT. E.g., Bouamor et al. (2012) showed that for French-English it is enough to simply add multi-word unit pairs to the parallel corpus; however, they observed a limited gain of +0.3 BLEU (Papineni et al., 2002) points. They extracted the multi-word units from the parallel corpus (for each language separately) with a method similar to the linguistically motivated term identification using morpho-syntactic patterns and paired the units using a statistically motivated method by building a vector space model (Salton et al., 1975). They investigated also two translation model adaptation methods: 1) by extending the SMT system's phrase table with new entries (this method did not show significant quality improvements), and 2) by extending the first method further by adding a feature that indicates, which entries have been newly added as term phrases. The third method showed a significant translation quality decrease. Using additional phrase tables and explicit user-specified translations of known phrases is a general practice in SMT for different purposes (e.g., Chen & Eisele (2010) use such methods to create hybrid SMT systems).

Although not directly related to terminology translation, Nikoulina et al. (2012) have proposed a framework for integrating Named Entities within SMT systems by pre-processing parallel corpora and replacing NEs identified in texts with NE category codes. The SMT system is trained using the pre-processed parallel and monolingual data. When performing translation, the source text is always pre-processed with the same techniques that replace NEs with the NE category codes. After the translation with the SMT system, the codes are replaced back with translations of the NEs using NE-specific translation methods. It was shown that the introduced model could lead to +2-3 BLEU point improvement over a baseline system for two different test sets. However, they report results for the translation between languages with little morphological inflection (i.e., from English to French) and the NEs are translated with one to one translation equivalents (i.e., using just the canonical forms), which for translation into morphologically rich languages may not be enough to achieve SMT quality improvement. Because NEs in contexts behave orthogonally to terms (i.e., NEs of the same category often have common contexts, however terms don't have a concept for categories and each term can have different contexts), the NE translation method using replacement is not directly applicable to terminology integration in SMT.

In terms of direct terminology integration, similar work to the author's work that shows significant quality improvements has been recently performed by Arcan et al. (2014a) for the English-Italian language pair (in both translation directions). They use a bilingual term collection to create a "*fill-up*" translation model that consists of a pre-trained SMT system's phrase table merged with a phrase table created from the bilingual terminology. The phrases that are present in both phrase tables get the highest probability scores assigned in the fill-up model. They introduce also a feature in the fill-up model's phrase table that identifies phrase translations coming from the in-domain phrase table. This way, the method allows assigning higher translation probabilities to in-domain term translations over out-of-domain term translations. However, the method does not identify correct term translations that are present in phrases within the initial phrase table, which could potentially have a greater impact on the translation model. Their results show an SMT quality improvement of up to 2 BLEU points.

4.2. Term Collections for SMT Experiments

For experiments on static terminology integration in SMT systems, the author uses three different types of term collections: 1) an English-Latvian term collection created using automatic bilingual term extraction methods from parallel and comparable corpora, 2) a manually filtered version of the automatically acquired term collection, and 3) four term collections for different language pairs (English-Latvian, English-Lithuanian, English-Estonian, and English-German) created by professional translators.

4.2.1. Automatically Created Term Collection

To create the automatically extracted term collection for English-Latvian, a small proprietary parallel corpus of 1,745 sentence pairs in the automotive domain was used. At first, terms and named entities (NE) were monolingually identified in the data. For terms, the methods described in section 2.2 were used and for named entities the *TildeNER* (Pinnis, 2012) named entity recogniser for Latvian and *OpenNLP*[22] for English were used. Then, the monolingually identified terms and named entities (542 unique English and 786 unique Latvian units in total) were cross-lingually mapped using the parallel tuning data and the methodology by Pinnis & Skadiņš (2012). As a result, 783 term and NE phrase pairs were identified. These phrases were then used to collect an in-domain comparable corpus from which additional term pairs were extracted using cross-lingual term mapping methods. The comparable corpora collection procedure is described further.

4.2.1.1. Comparable Corpora Collection

The author performed comparable corpora collection instead of parallel corpora collection because of two reasons: 1) parallel corpora (especially in narrow domains and for under-resourced languages, such as Latvian) is scarce or non-existing (Skadiņa et al., 2012), and 2) comparable corpora is widely available (e.g., domain specific news in different

[22] Apache OpenNLP is available online at: http://opennlp.apache.org/.

languages, open access multi-lingual encyclopaedias, such as *Wikipedia*, localised industry or community created Web sites about specific topics, etc.).

There are many different parallel and bilingual comparable corpora collection tools available. Several of the better known tools are, for instance, *Babouk* (De Groc, 2011), *Bitextor* (Esplá-Gomis, 2009), *Focussed Monolingual Crawler* (*FMC*; Mastropavlos & Papavassiliou, 2011), BootCAT (Baroni & Bernardini, 2004), and many others. For comparable corpora collection, the author uses the *FMC* tool

FMC requires seed terms to collect a domain-specific comparable corpus. For this purpose, the author used the English-Latvian term and named entity pairs extracted in the previous step. As the automatically extracted seed terms can contain also out-of-domain or cross-domain terms and named entities, it is necessary to filter the seed term list so that only domain-specific terms (and as few as possible cross-domain and out-of domain terms) would be included. For instance, if we want to collect a corpus in the automotive domain a natural choice of a term for a seed term list could be "*oil*". But would it really be a good candidate for a seed term? The word "*oil*" is very ambiguous. When talking about "*oil*", we may refer to "*body oil*", "*cooking oil*", "*baby oil*", "*massage oil*", and many other types of "*oil*". When ambiguous seed terms are used to collect a Web corpus, the collected corpora will include texts from different domains. Therefore, the terms were ranked using a term pair specificity estimation method. The method uses reference corpus inverse document frequency (IDF) scores of words calculated on general (broad) domain corpora (in the author's work, the *Wikipedia* and current news corpora) to weigh the specificity of a phrase. Each bilingual phrase was ranked using the following equation:

$$R\left(p_{src}, p_{trg}\right) = min\left(\sum_{i=1}^{|p_{src}|} IDF_{src}\left(p_{src}(i)\right), \sum_{j=1}^{|p_{trg}|} IDF_{trg}\left(p_{trg}(j)\right)\right) \qquad (3)$$

where p_{src} and p_{trg} denote phrases in the source and target languages and IDF_{src} and IDF_{trg} denote the respective language IDF score functions that return an IDF score for a given token. The ranking method has been selected through a heuristic analysis process so that specific in-domain term and named entity phrases would be ranked higher than broad-domain or cross-domain phrases. The method allows filtering out phrase pairs where a phrase may have a more general meaning in one language, but a specific meaning in the other language. Thereby, the method ensures that the bilingual terminology is domain specific and that the collected corpus will be in the required domain. After applying a threshold on the term pair ranks, 614 phrase pairs remained in the seed term list for corpora collection.

Additionally to the seed terms *FMC* requires seed URLs. Therefore, 55 English and 14 Latvian URLs for Web resources focussed on the automotive domain were manually collected.

When the seed terms and seed URLs were acquired, a 48 hour focussed monolingual web crawl was initiated for both languages. The statistics of the corpora are given in Table 23.

Table 23. Monolingual automotive domain corpora statistics

Language	Unique Documents	Sentences	Tokens	Unique Sentences	Tokens in Unique Sentences
English	34,540	8,743,701	58,526,502	**1,481,331**	20,134,075
Latvian	6,155	1,664,403	15,776,967	**271,327**	4,290,213

To perform bilingual term extraction from two monolingual corpora, the corpora have to be aligned in document level (i.e., documents from one corpus have to be paired with documents from the second corpus). For document alignment the author used a cross-lingual comparability metric, more specifically, the *DictMetric* (Su & Babych, 2012) tool. The main task of a comparability metric is to estimate how much content of two documents overlaps (translated phrases, sentences, whole paragraphs, etc.). *DictMetric* scores document pair comparability (the higher the comparability, the more content overlap is present) and aligns document pairs that exceed a specified comparability score threshold. In total, 81,373 document pairs were produced in this step. The final comparable corpus statistics are given in Table 24. The Latvian part of the comparable corpus has been also used in further SMT experiments.

Table 24. English-Latvian automotive domain comparable corpus statistics

Language	Unique Documents	Unique Sentences	Tokens in Unique Sentences
English	24,124	1,114,609	15,660,911
Latvian	5,461	247,846	3,939,921

Once the corpus was collected, bilingual terminology was extracted from the corpus using the *Terminology Aligner* (TEA; Ştefănescu, 2012). In total, both automatic bilingual term extraction methods produced 979 term and named entity pairs. In the SMT experiments described further, this term collection is named as the "***non-filtered***" term collection.

4.2.2. Filtered Term Collection

Because the *non-filtered* term pairs contain noise that is created in the automatic cross-lingual term mapping process, the term collection was further manually filtered in order to remove noise and too general and too ambiguous term pairs. However, note that this is a filtering process and not a term collection creation process, where the terms would be transformed from their inflected forms to their canonical forms. After filtering, the "***filtered***" term collection consisted of 845 term pairs.

4.2.3. Professional Term Collection

In parallel to the automatic bilingual term extraction and manual filtering, a professional translator was asked to create a term collection from the parallel automotive domain corpus from which the automatically extracted bilingual terms were acquired. The translator produced a term collection consisting of 644 term pairs (all terms are given in their canonical forms). This term collection is further referred to as the "***professional***" term collection.

4.3. Terminology as a Corpus

The simplest method for terminology integration in SMT systems that is applied also by related works on terminology integration in SMT systems (Bouamor et al., 2012) is to add the in-domain bilingual term collections to the parallel corpus, which is used for translation model training, and the target language terms to the monolingual corpus, which is used for language model training. This method, although being very simple, is quite efficient, because it ensures that the terms that are not covered by both the parallel corpus and the monolingual corpus (i.e., terms that can be considered as out-of-vocabulary terms) will have a larger possibility of having at least one translation hypothesis. The conceptual design of this method is depicted in Figure 23.

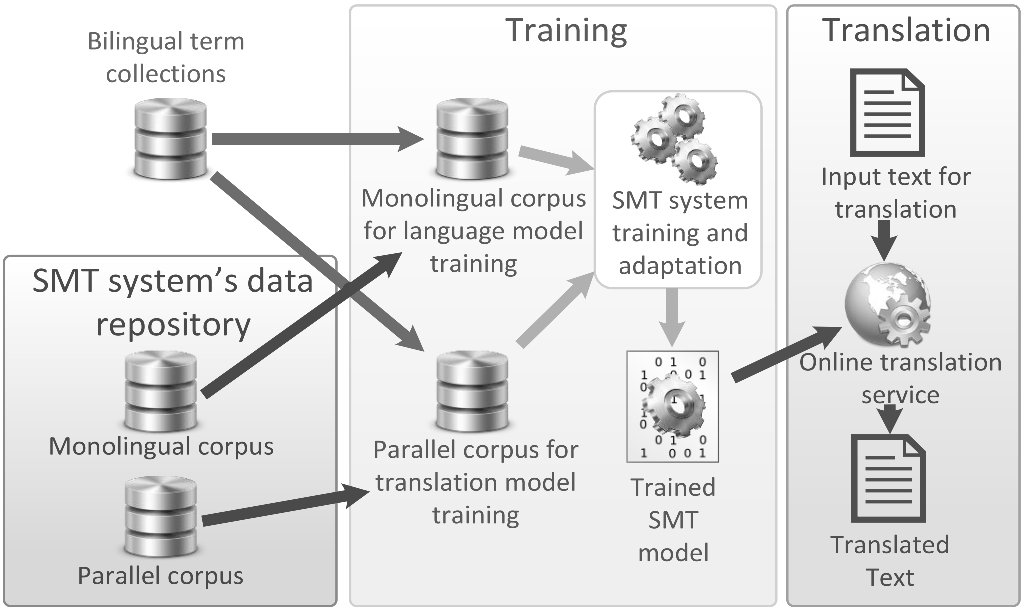

Bilingual term collections
Training
Translation
Monolingual corpus for language model training
SMT system training and adaptation
Input text for translation
SMT system's data repository
Monolingual corpus
Parallel corpus
Parallel corpus for translation model training
Trained SMT model
Online translation service
Translated Text

Figure 23. The conceptual design of the "*Terminology as a Corpus*" method

A requirement for this method to work is that terminology is added to both the parallel corpus and the monolingual corpus. Such a requirement is set, because when translating a sentence, the translation model is responsible for generating translation hypotheses and the language model is responsible for estimating how well the generated hypotheses represent (or are likely to belong to) the target language. Therefore, if a term is not present in the parallel corpus, it cannot be present in the translation hypotheses and if the term is not present in the monolingual target language's corpus, the hypotheses containing the term will receive a low score from the language model.

However, this method has a limitation. Because terms in the term collections that are acquired from term banks, e.g., *EuroTermBank*, *IATE*, the *TaaS* platform, etc., are usually stored in their canonical forms (or base forms), for languages that feature rich morphologies where words can be morphologically inflected, this method won't allow the identification of translation equivalents for terms that in contexts appear in inflected forms different from their canonical forms. Nevertheless, this method can be efficient in the following three scenarios:

- When translating from and to languages with little morphological inflection (e.g., from or to English, German, French, etc.), terms in contexts are often equal to their canonical forms. Consequently, the recall and the effectiveness of the method is higher than for morphologically richer languages. E.g., for Latvian, which is a morphologically rich language, even when translating from English, as shown by Pinnis and Skadiņš (2012), the method does not show quality improvements when using a term collection from an authoritative source (the *EuroTermBank*) because of two main reasons: 1) in Latvian terms appear in many different inflected forms, and 2) many of the terms in the authoritative data base are ambiguous (they may have multiple translation equivalents listed, which all may represent the same terminological concept), thus the addition of new term pairs causes more statistical uncertainty for the SMT system. However, it also does not show a quality decrease, which for the method in general is a positive result.
- When acquiring term collections in an automatic process from, e.g., parallel data or comparable data, the bilingual terms are already stored in inflected forms that are common in different contexts. These bilingual term pairs are better suited as possible translation hypotheses in different contexts than the canonical forms (for which the usage in different contexts may be very limited). For more details on this scenario see section 4.3.1.
- Even if the bilingual terminology is provided by the term data bases in a canonical form, it can still be beneficial in the SMT system training process. More specifically, by adding the bilingual terminology to the parallel corpus, we indirectly provide the word alignment processes (e.g., the *Giza++* tool in *Moses*) and further also the phrase extraction process in the SMT training system with a list of valid term alignments (single word and multi-word alignments), which can help word alignment and phrase extraction processes to produce word and phrase alignments with a higher precision.

4.3.1. Evaluation Scenarios and Results

The static terminology integration methods are all evaluated for the English-Latvian language pair using evaluation data from the automotive domain. For baseline systems, the publicly available *DGT-TM* parallel corpus (Steinberger et al., 2012) as the general language corpus is used throughout the experiments. More specifically, the *DGT-TM* releases of 2007, 2011 and 2012 have been used. The total amount of parallel sentences in the corpus is 3’159,459 before noise filters (duplicate filters, corrupt sentence filters, etc.) of the *LetsMT* platform and 1’954,740 sentence pairs after the filters. The target language side of the parallel corpus is used for language modelling. After noise filtering, the monolingual corpus consisted of 1’887,304 sentences. For tuning of the English-Latvian system, a small in-domain parallel corpus of 2,617 sentence pairs was used (the same corpus used to create the automatically extracted term collection in section 4.2). The corpus was randomly split into a tuning set (1745 sentence pairs) and an evaluation set (872 sentence pairs). The tuning set and the evaluation set are static throughout the whole English-Latvian experiments presented in this thesis. For the English-Latvian experiments for static terminology integration in SMT systems three types of term collections (described in section 4.2) were used:

- The automatically extracted bilingual term collection consisting of 979 term pairs (“*non-filtered*” in the results below).
- The manually revised version of the automatically extracted bilingual term collection consisting of 845 term pairs (“*filtered*” in the results below).
- The bilingual term collection created by a professional translator consisting of 644 term pairs (“*professional*” in the results below).

Using the publicly available corpus and the tuning data, a baseline system was trained within the *LetsMT* platform. The automatic evaluation results are given in Table 25. Then, the non-filtered term collection was added to the parallel and monolingual corpora and the system was re-trained. The results show that there is a significant increase over the baseline system in translation quality (from 12.68 to 15.51 BLEU points). When training an SMT system using the filtered term collection, it is evident that the results are lower than with the non-filtered terms. This may be explained with the fact that the automatic alignments were acquired from very precise in-domain data with respect to the evaluation data and even though the aligned pairs were noisy and ambiguous, they represented the in-domain data better. Therefore, the noisy data allows achieving a higher result. Finally, a system was trained using the term collection created by the professional translator. The results are lower than with the automatically extracted term collection and the manually post-processed (filtered) term collection. Adding the professionally created term collection to the parallel and monolingual corpora did not yield a better result than the automatically extracted term collection, because of two main reasons: 1) the terms were in their base forms, which when translating into Latvian often are not the required inflected forms, and 2) the professional term collection contains terms, which in different contexts may be ambiguous and provides just one translation candidate (e.g., "*cover*" may be a noun "*pārsegs*" or a verb "*nosegt*", "*fill*" may be a noun "*uzpilde*" or a verb "*uzpildīt*"/"*aizpildīt*" depending on the context, etc.). The automatically extracted term collection is able to provide multiple translation equivalents for each term also in different inflected forms (as found in the corpus from which the bilingual term collection is extracted).

Table 25. Terminology as a Corpus evaluation results

Scenario	BLEU (C)	BLEU	NIST (C)	NIST	METEOR (C)	METEOR	TER (C)	TER
Baseline	12.00	12.68	4.1361	4.2644	0.1439	0.1849	0.7893	0.7801
Non-filtered	**14.60**	**15.51**	**4.4756**	**4.6301**	**0.1599**	**0.2011**	**0.7660**	**0.7531**
Filtered	13.94	14.76	4.4010	4.5376	0.1580	0.1985	0.7719	0.7604
Professional	12.97	13.62	4.3422	4.4792	0.1513	0.1941	0.7697	0.7586

4.4. Translation Model Adaptation

As described in the introduction, the task of the translation model is to generate translation hypotheses for source language sentences. Therefore, the goal of terminology integration in an SMT system's translation model is to either make the translation model prefer in-domain translation hypotheses for terms over out-of-domain translation hypotheses in as many in-domain contexts as possible (i.e., generate in-domain translation hypotheses with higher translation likelihood scores than out-of-domain translation hypotheses) or to allow only in-domain translation hypotheses of terms. The conceptual design of the translation model adaptation methods using bilingual term collections is depicted in Figure 24.

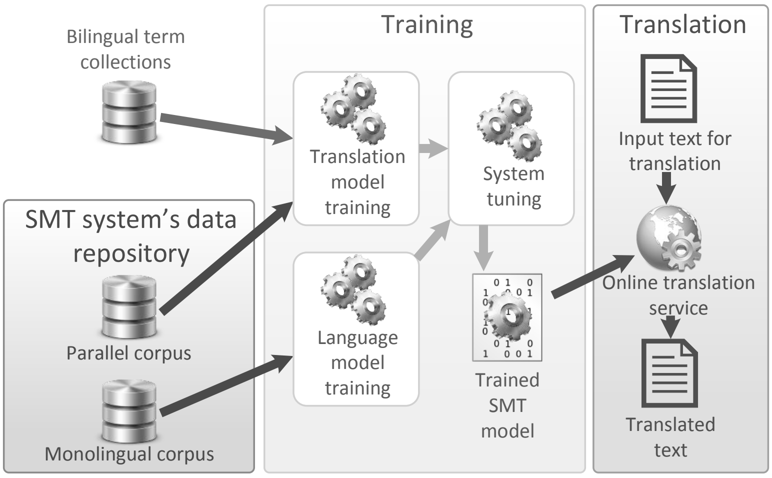

Bilingual term collections
Training
Translation
Translation model training
System tuning
Input text for translation
SMT system's data repository
Parallel corpus
Monolingual corpus
Language model training
Trained SMT model
Online translation service
Translated text

Figure 24. The conceptual design of the "Translation model adaptation" methods

Further subsections will describe two methods that allow performing translation model adaptation using bilingual term collections.

4.4.1. Phrase Table Adaptation

In this method, following the methodology published in Pinnis and Skadiņš (2012), the *Moses* phrase table of the translation model is transformed into an in-domain term-aware phrase table. This is performed by adding a new feature to the default features that are used in *Moses* phrase tables. Figure 25 shows that the phrase table adaptation is performed immediately after a phrase table is created in the SMT system's training process (the "*consolidate-ttable-halves*" process in the *LetsMT* platform).

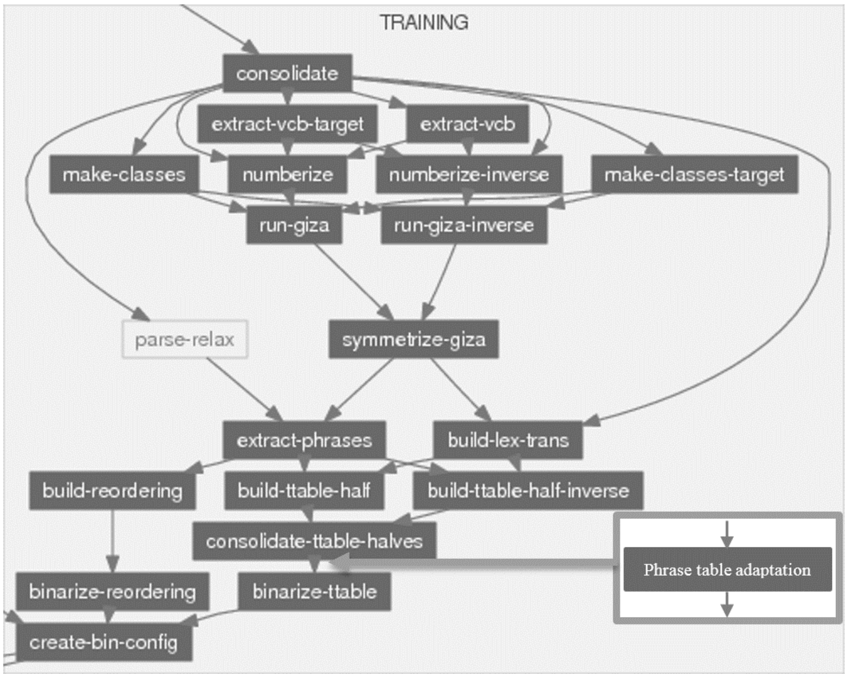
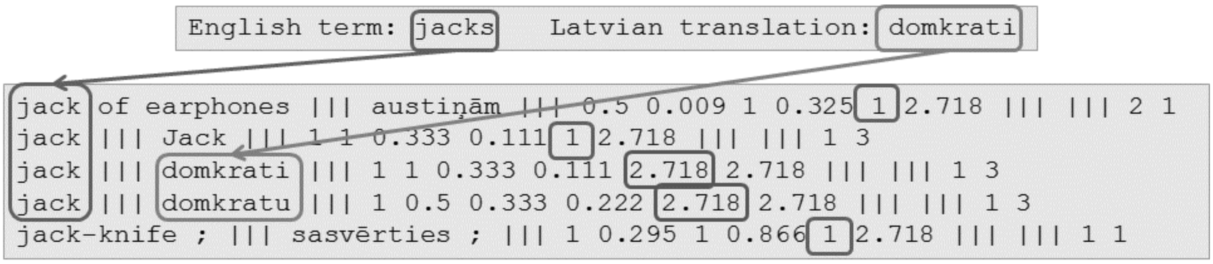


Figure 25. Phrase table adaptation as a step in the translation model training workflow in the LetsMT platform

The term identifying feature receives the following values:

- "*1*" if a phrase on both sides (in both languages) does not contain a term pair from a bilingual term list. If a phrase contains a term only on one side (in one language), but not on the other, it receives the value "1" as such situations indicate about possible out-of-domain (wrong) translation candidates.
- "*2.718*" if a phrase on both sides (in both languages) contains a term pair from the bilingual term collection.

In order to find out whether a phrase in the phrase table contains a given term or not, phrases and terms are stemmed prior to comparison. This allows finding inflected forms of term phrases even if those are not given in the bilingual term list. The new feature identifies phrases containing in-domain term translations and allows assigning higher translation probabilities to in-domain translation hypotheses. Different from the method proposed by Arcan et al. (2014a), this method affects the whole phrase table as it identifies terms that are contained within longer phrases. An example excerpt from an English-Latvian *Moses* phrase table with the term identifying feature is given in Figure 26.



```
English term: jacks   Latvian translation: domkrati

jack of earphones ||| austiņām ||| 0.5 0.009 1 0.325 1 2.718 ||| ||| 2 1
jack ||| Jack ||| 1 1 0.333 0.111 1 2.718 ||| ||| 1 3
jack ||| domkrati ||| 1 1 0.333 0.111 2.718 2.718 ||| ||| 1 3
jack ||| domkratu ||| 1 0.5 0.333 0.222 2.718 2.718 ||| ||| 1 3
jack-knife ; ||| sasvērties ; ||| 1 0.295 1 0.866 1 2.718 ||| ||| 1 1
```

Figure 26. Example excerpt from an English-Latvian *Moses* phrase table with the term identifying feature

When both the translation model and the language model are created, in a typical SMT system training workflow the system is tuned, e.g., with Minimum Error Rate Training (MERT; Bertoldi et al., 2009). The task of tuning is to learn weights for the different features of an SMT model using a representative of the target domain set of parallel sentences – the tuning data. The phrase table of the translation model after the adaptation contains an additional feature that identifies whether a phrase pair contains bilingual terminology. In order for the new feature to be productive, the tuning data has to contain the same terminology that was used to adapt the phrase table, otherwise the tuning process will learn that the new feature is "*useless*" and assign it a negative weight. Therefore, an important aspect for the phrase table adaptation method is the selection of tuning data. In the scenarios in section 4.4.3 the non-filtered and filtered bilingual term collections have been also enriched with terms automatically extracted from the tuning data, thereby ensuring the presence of in-domain terminology in the tuning data. However, tuning data could be also selected in an automated process from parallel data, e.g., by randomly selecting sentence pairs containing 0, 1, 2, etc. bilingual term pairs from the bilingual term collections in the sentence pairs.

4.4.2. Phrase Filtering

When performing phrase table adaptation the SMT system is trained to prefer in-domain translation hypotheses to out-of-domain translation hypotheses. However, in some situations we might want to limit the term translation hypotheses to only those that are present in a term collection and disallow all out-of-domain translations at all. Such a scenario could be beneficial for the translation of, e.g., named entities (which are not terms, but nonetheless), terms that have to have a specific translation in domains or use cases that can be considered sensitive (e.g., in a very sensitive scenario we could disallow racist or abusive translations), etc. Thus, this section describes a phrase filtering method that allows implementing the above mentioned restrictions for translation hypotheses selection in SMT system translation models.

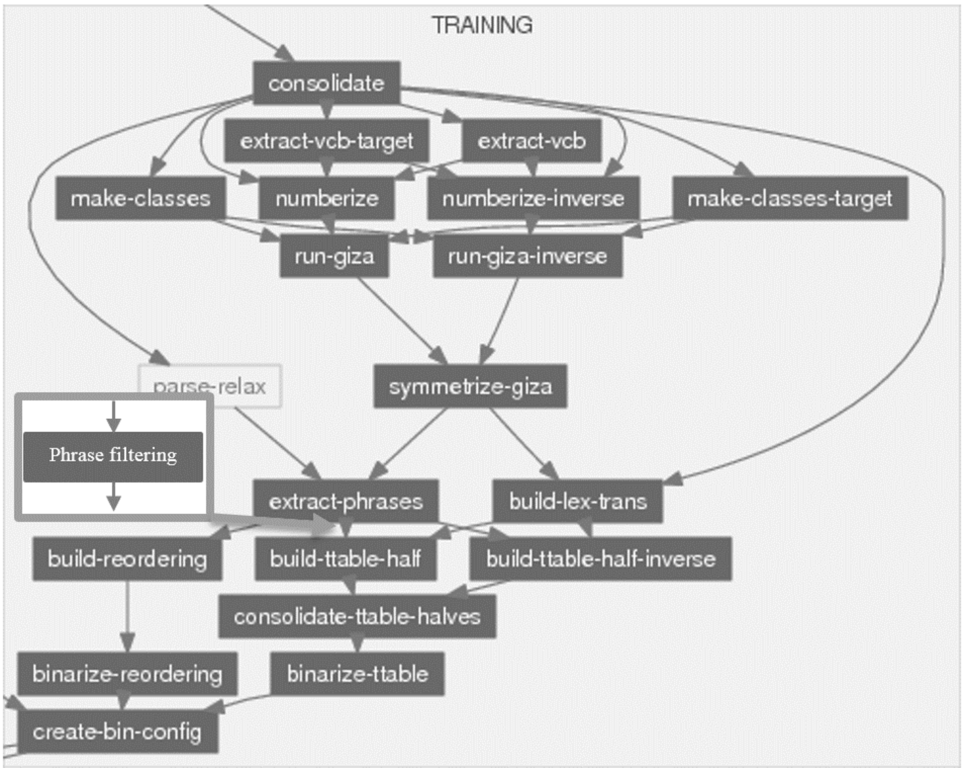


Figure 27. Phrase filtering as a step in the translation model training workflow in the LetsMT platform

As shown in Figure 27, the phrase filtering can be performed immediately after phrase extraction of the *Moses* SMT system (in the *LetsMT* platform the phrase extraction process is named "*extract-phrases*"). The filtering has to be performed before phrase scoring in order to ensure correct calculation of translation probability scores. In order to identify terms in different inflected forms, the filtering process for each word of a term in the first language performs lightweight stemming (i.e., removes only endings). For each corresponding term in the second language the process keeps only the first four letters (however, if a word is shorter than five letters, it is stemmed) of each word. When searching for invalid term pairs that have to be filtered out, such a lightweight and "*rude*" stemming approaches allow limiting the possibility of filtering out many correct term pairs because of high possible morphological variations that the lightweight stemming approach (if applied for both languages) would not be able to capture.

The term filtering method is both effective (it filters out all wrong translation hypotheses) and very risky. That is, if a term collection contains ambiguous terms, that is, phrases that may have multiple meanings and multiple translations also in in-domain texts and not all translations will be defined in the term collection, then the phrase pairs that contain such translation equivalents, regardless of the fact that they are correct translation hypotheses, will be filtered out. Let us go through a small example. Imagine that we have an English-Latvian term collection containing terms from the automotive domain (see Table 26) and we have an example excerpt from phrases extracted by the *LetsMT* platform's SMT system training process and filtered out by the term filtering process (see Table 27), i.e., the filtering process has decided that the term pairs are wrong.

Table 26. An example English-Latvian term collection in the automotive domain

English term	Latvian term	English term	Latvian term
force	*spēks*	*rail*	*sliedes*
production	*ražošana*	*production*	*ražošana*
version	*versija*	*product*	*produkts*
service	*apkope*	*instrument*	*instruments*
service	*serviss*	*transmission*	*transmisija*
rail	*dzelzceļš*		

Table 27. An example of English-Latvian phrase pairs that were filtered out by the phrase filtering process

No.	English phrase	Latvian phrase	Correct	In-domain?
1	**force** majeure	majeure	No	-
2	for the **production**	par tās izpildi	No	-
3	the **production**	tās	No	-
4	the Dutch **version**	holandiešu **tekstā**	Yes	No
5	entry into **service**	nodots **ekspluatācijā**	Yes	Yes
6	**service**	**pakalpojums**	Yes	Yes *
7	gateway in the **rails**	ieeju kuģu **margās**	Yes	No
8	plant protection **products**	augu aizsardzības **līdzekļu**	Yes	No
9	**products**	**izstrādājumu**	Yes	No *
10	control **instruments**	kontroles ierīces	Yes	No
11	**transmission**	**pārnesumkārbas**	No	-

For each phrase pair the Table 27 also gives information whether the phrase pair is correct, i.e., whether such a translation exists regardless the automotive domain constraints and it also shows whether the phrase pair, if correct, can be considered an in-domain phrase pair regardless of its presence in the term collection. That is, the last column should tell us whether there exist term pairs that we have forgotten to include in the term collection. The "*" in the last column indicates that the domain affiliation is ambiguous (meaning, it could and could not belong to the domain).

Table 27 shows that the filtering step is able to filter out pairs that are incorrect in terms of phrase boundaries (the examples 1, 2, 3, and 4 in the table). The method also correctly filters out correct, but out-of-domain phrase pairs. However, it can be seen that the method is not forgiving if the term collection lacks an important translation equivalent. All phrases containing the missing translation equivalent are filtered out. Nevertheless, section 4.4.3 will show that the method can be beneficial if applied wisely and bearing in mind the behaviour of the method; it will be also shown that performing phrase filtering with this method using automatically extracted term pairs from parallel or comparable corpora without manual revision is not recommended.

4.4.3. Evaluation Scenarios and Results

For the evaluation of translation model adaptation with bilingual terminology, a similar data combination as for the evaluation of the Terminology as a Corpus scenario (see section 4.3.1) is used. The difference, however, is that two baseline systems were built and bilingual term collections were integrated in the parallel and monolingual corpora (thus also the BLEU scores are higher for the baseline systems). Furthermore, the baseline systems were built with a second language model – an in-domain language model. The data for the in-domain language model was collected from the Web using the *Focussed Monolingual Crawler* (*FMC*). The in-domain monolingual corpus consists of 1’664,403 sentences before the *LetsMT* platform’s noise filter and 224,639 sentences after noise filtering. As the noise filter removes also duplicate sentences, there is a large size reduction of the in-domain monolingual corpus. More details on the corpus collection process can be found in section 4.2.1.1.

For both baseline scenarios (with *non-filtered* and with *filtered* term collections) the translation models were separately adapted in order to evaluate the translation quality changes. Then, the phrase table filtering method was evaluated in two separate scenarios – source-to-target filtering and target-to-source filtering. Because in the filtering scenario invalid phrase pairs are removed with respect to the first language (irrelevant of the translation direction), we get different filtering results if we consider the source language the first or the target language. The evaluation results in Table 28 show that for English-Latvian the source-to-target filtering achieves a higher result (in terms of translation quality). It is also evident that using the filtered term collection, from which ambiguous terms and too general terms were manually removed, the translation quality exceeds even the baseline system’s translation quality. Whereas the *non-filtered* term collection causes valid phrase pairs to be filtered out from the phrase table. Therefore, the translation quality slightly decreases in comparison to the baseline system. It should be noted that the phrase filtering is a challenging method that can have beneficial effects, however the term collection has to be very complete (either consisting of non-ambiguous terms or all terms that are ambiguous have to have all possible translation equivalents specified in the term collection) in order to achieve a translation quality improvement. However, as shown by the results with the *filtered* term collection, translation quality improvements can be achieved.

Table 28. Evaluation results of terminology integration in SMT systems during training – translation model adaptation

Scenario	BLEU (C)	BLEU	NIST (C)	NIST	METEOR (C)	METEOR	TER (C)	TER
Non-filtered terms								
Baseline	14.96	15.72	4.5095	4.6825	0.1588	0.2026	0.7660	0.7532
Source-to-target filtering	14.95	15.68	4.5329	4.6976	0.1609	0.2041	**0.7626**	**0.7507**
Target-to-source filtering	14.34	15.06	4.4613	4.6249	0.1565	0.2005	0.7745	0.7632
Term identifying feature	**15.21**	**15.96**	**4.5884**	**4.7566**	**0.1623**	**0.2058**	0.7636	0.7514
Filtered terms								
Baseline	13.12	13.87	3.9872	4.1404	0.1385	0.1811	0.7987	0.7874
Source-to-target filtering	**13.42**	**14.21**	4.0753	4.2273	0.1417	0.1839	0.7877	0.7754
Target-to-source filtering	12.31	12.95	3.8403	3.9850	0.1314	0.1730	0.8070	0.7963
Term identifying feature	13.39	14.1	**4.1029**	**4.2458**	**0.1434**	**0.1852**	**0.7857**	**0.7737**

Finally, the phrase table adaptation with the help of an additional feature in the phrase table that identifies bilingual terminology in phrase pairs also achieves a translation quality improvement over the baseline systems for both *filtered* and *non-filtered* term collection scenarios. In addition, it should be noted that this method has not shown a translation quality decrease in the author's experiments.

4.5. Language Model Adaptation

The second area of focus after terminology integration in translation models has been the usage of bilingual terminology to perform language model adaptation. The conceptual design of bilingual terminology integration in SMT system language models is depicted in Figure 28. More specifically, this section investigates methods for monolingual corpora splitting with the help of in-domain terminology into in-domain and out-of-domain sets.

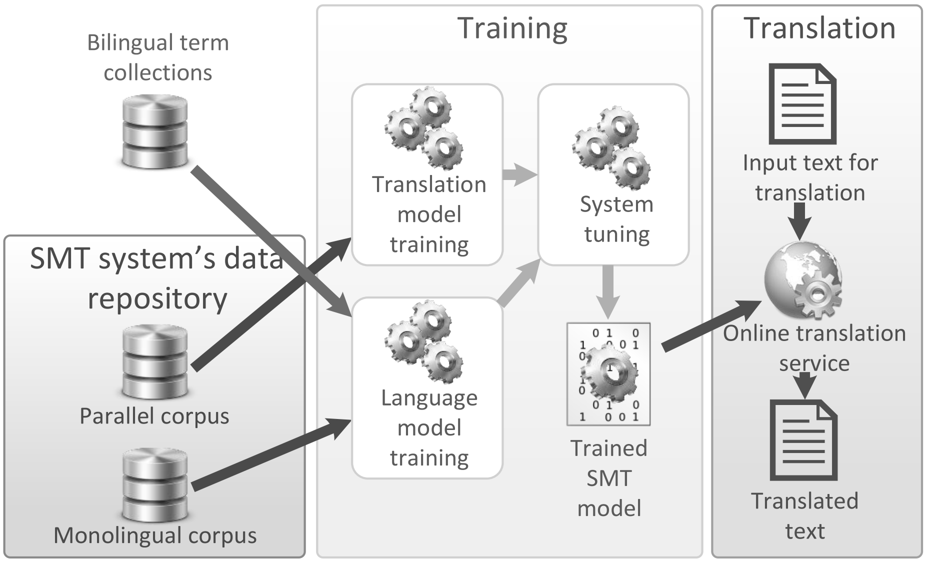

Bilingual term collections
Training
Translation
Translation model training
System tuning
Input text for translation
SMT system's data repository
Parallel corpus
Monolingual corpus
Language model training
Trained SMT model
Online translation service
Translated text

Figure 28. The conceptual design of the "*Language model adaptation*" methods

4.5.1. Monolingual Corpora Splitting

The idea behind monolingual corpora splitting is that if we already have a large monolingual corpus, we could use this large corpus and extract from it sentences that we consider as in-domain sentences using a term collection. Because the in-domain sentences contain in-domain terminology, combined in a corpus they should represent the in-domain texts of the target language better. However, it should be noted that this method is highly experimental. By performing just monolingual analysis the method can also extract sentences that contain the lexical forms of terms, however with different meanings (i.e., if the lexical forms are ambiguous). Nevertheless, several experiments were performed in order to understand how much can be achieved by the monolingual corpora splitting method.

Each monolingual corpus can be split in two parts – an in-domain part and an out-of-domain part. When we have just one monolingual corpus, we can easily split it in the two parts (see Figure 29). However, if we have more than one corpus, we have multiple choices:

- We can split both corpora in two parts and train four language models – two in-domain language models and two out-of-domain language models (see Figure 30).
- We can split both corpora in two parts and then combine the in-domain and out-of-domain parts together so that we would end up having again two corpora – an in-domain and an out-of-domain corpus (see Figure 31).
- We can also create three language models by splitting just one of the corpora or by splitting all of them, but concatenating back just the in-domain or the out-of-domain part.

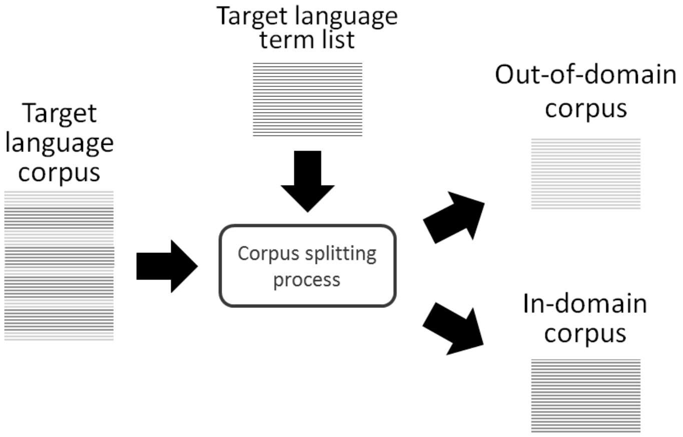


Figure 29. The monolingual corpus splitting method for one target language corpus

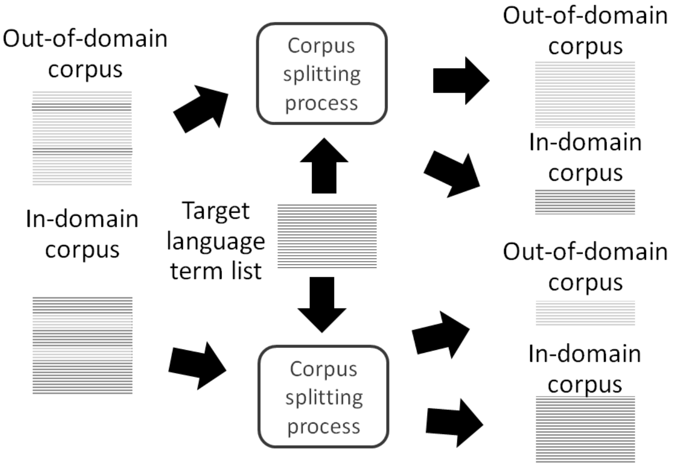


Figure 30. The monolingual corpus splitting method for two target language corpora (a)

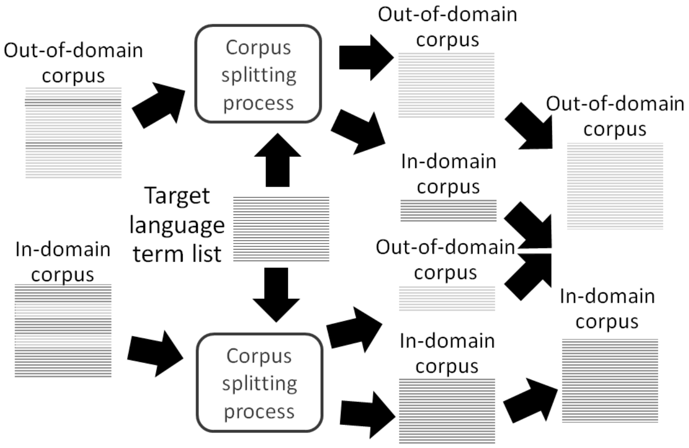


Figure 31. The monolingual corpus splitting method for two target language corpora (b)

4.5.2. Evaluation Scenarios and Results

For the evaluation of language model adaptation with bilingual terminology, similar data set-up as for the evaluation of the *Terminology as a Corpus* scenarios (see section 4.3.1) was used. The difference, however, is that the in-domain monolingual corpus that was collected from the Web (see section 4.2.1.1) was also used to train language models. The baseline system has been trained using two language models – the general language model that is based on the *DGT-TM* corpus and the in-domain language model that is based on the comparable Web corpus. For these experiments only the "*non-filtered*" term collection was used as it allowed achieving the highest results in previous experiments.

Additionally to the baseline system, three experiments were performed with monolingual corpora splitting techniques. In the first experiment, the general domain corpus and the in-domain corpus were split in two parts. The resulting in-domain parts were joined together in a larger in-domain corpus, however, the out-of-domain parts were kept separated. Thus, we trained three language models – one with out-of-domain data (from the *DGT-TM* corpus), one with pseudo-out-of-domain data (the out-of-domain part of the initial in-domain corpus), and one with in-domain data from both initial corpora. For the second experiment all four parts were kept separated (thus having two in-domain and two out-of-domain language models). For the third experiment the in-domain parts and respectively also the out-of-domain parts were concatenated in order to train just two language models.

The Table 29 shows the evaluation results. It is evident in our results that the system with two language models, which were based on the two reorganised monolingual corpora) achieved a significantly higher result than all other systems. However, further analysis is needed in order to verify that the method works also with different corpora and different term collections.

The Table 29 shows also results of phrase table adaptation (the scenarios that use the "*term identifying feature*"). For all scenarios with the adapted phrase table, translation quality improved.

Table 29. Evaluation results of terminology integration in SMT systems during training – language model adaptation

Scenario	**BLEU (C)**	**BLEU**	**NIST (C)**	**NIST**	**METEOR (C)**	**METEOR**	**TER (C)**	**TER**
Baseline	13.41	14.03	4.0188	4.1510	0.1390	0.1795	0.7991	0.7881
+ term identifying feature	13.77	14.43	4.0963	4.2284	0.1424	0.1823	0.7838	0.7735
3 mono corpora	13.79	14.45	4.1979	4.3170	0.1497	0.1903	0.7797	0.7691
+ term identifying feature	14.03	14.7	4.2493	4.3825	0.1492	0.1913	0.7753	0.7636
4 mono corpora	13.49	14.3	4.0986	4.2610	0.1418	0.1848	0.7913	0.7778
+ term identifying feature	13.91	14.69	4.1278	4.2795	0.1470	0.1881	0.7823	0.7695
2 reorganised mono corpora	14.21	15.06	4.2406	4.3855	0.1501	0.1926	0.7771	0.7667
+ term identifying feature	**15.34**	**16.24**	**4.4966**	**4.6588**	**0.1603**	**0.2053**	**0.7596**	**0.7470**

4.6. Summary of Static Integration of Terminology in SMT Systems

In this section the author presented methods for static bilingual terminology integration in SMT systems. In total, three different types of methods were discussed: 1) terminology injection into SMT system parallel and monolingual data (i.e., the *Terminology as a Corpus* method), 2) methods for SMT system translation model adaptation (i.e., phrase table adaptation and phrase table filtering), and 3) methods for language model adaptation using monolingual corpus splitting techniques.

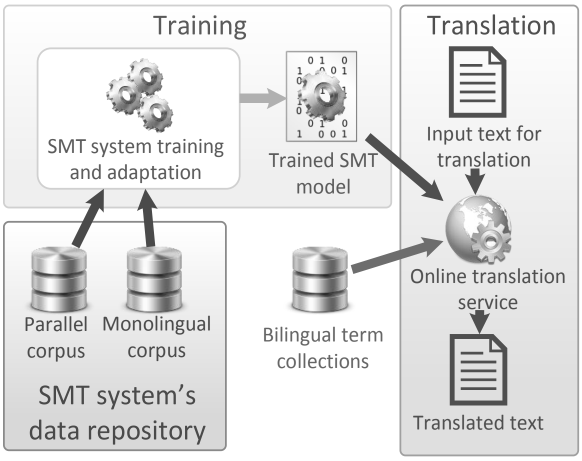
The evaluation was performed for the English-Latvian language pair using automatic SMT evaluation metrics (i.e., BLEU, NIST, TER, and METEOR). The evaluation showed that the *Terminology as a Corpus* method allows significantly boosting SMT quality by up to 22.3% (or 2.83 absolute BLEU points) over the general domain baseline system when using the automatically extracted bilingual term collection (i.e., "*non-filtered*" in the results). This is a significant result as it shows that when translating into morphologically rich languages terminology integration in SMT systems has to take into account the morphological variability of terms (i.e., the different inflected forms of terms). As a reminder, the "*non-filtered*" term collection contains term pairs in inflected forms as they were found in the in-domain parallel corpus used for the SMT system tuning and also in the in-domain comparable corpus collected from the Web.

The translation model adaptation by the introduction of a bilingual terminology identifying feature allowed to improve SMT quality cumulatively (comparing to the initial baseline system) up to 25.9% (or 3.28 absolute BLEU points) over the baseline system. Performing also language model adaptation with corpora splitting techniques allowed to boost the SMT quality improvement up to 28.1% (or 3.56 absolute BLEU points) over the initial baseline system.

The author believes that the most stable methods for static terminology integration in SMT systems are the *Terminology as a Corpus* method and the translation model adaptation with the term identifying feature in the translation model's phrase table. However, it has to be noted that in order for the new feature to be effective, the tuning data used for tuning of the SMT system has to be rich with the in-domain terminology used for adaptation of the translation model. If the tuning data will not contain the in-domain terminology, the tuning process will not be able to identify that the new feature is of any help to the improvement of the translation quality. It may even consider that the new feature has a negative effect on the SMT system's quality.

5. DYNAMIC INTEGRATION OF TERMINOLOGY IN SMT SYSTEMS

Terminology integration in SMT systems during training, as shown by the evaluation results, allows to tailor SMT systems to a required domain, however, it requires to re-train if not the whole SMT system then at least a significant portion of the system (e.g., the translation model, the language model, or even both and the systems have to be also re-tuned in order to adjust weights of the different features used in the SMT system). For many translation tasks (or projects for localisation service providers) re-training of a system could also be uneconomical (for instance, if all you need is to translate a five page document). Furthermore, if we have already trained a relatively good SMT system (let it be a general domain system or a close-domain system to the domain that is needed), why should we spend time on re-training it? We should instead be able to use the same SMT system, but tailor it to the required domain with the help of the right bilingual terminology. This section documents methods developed by the author that allow to perform dynamic terminology integration in SMT systems (conceptually depicted in Figure 32). The section is based on the author's contributions for the publications by Pinnis (2015), Pinnis & Skadiņš (2012), Skadiņš et al. (2013), Vasiļjevs et al. (2014a), and the TaaS project's Deliverable D4.4 Terminology Integration in SMT (TaaS, 2014b).



Training
SMT system training and adaptation
Trained SMT model
Translation
Input text for translation
Online translation service
Translated text
Parallel corpus
Monolingual corpus
SMT system's data repository
Bilingual term collections

Figure 32. The conceptual design of terminology integration in the SMT system translation level

The terminology integration is performed with a source text pre-processing workflow (see Figure 33) that uses the bilingual term collection in order to identify terms in the source text (e.g., sentence, paragraph, even a full document) using term identification methods described in section 1, annotates the content with possible translation hypotheses from the bilingual term collection using XML mark-up[23] that complies with the *Moses* SMT system's XML mark-up

[23] More details on the *Moses* XML mark-up can be found on the *Moses* SMT system's home page at: http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc7.

format, assigns translation likelihood scores for each of the translation hypotheses, and, finally, translates the document taking into account the injected mark-up.

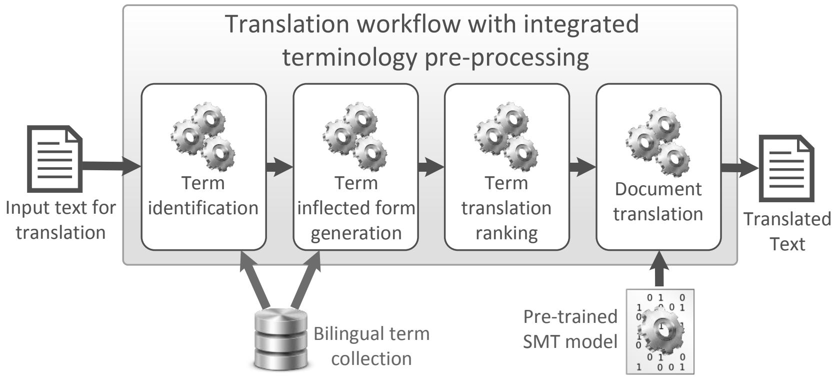

Translation workflow with integrated terminology pre-processing
Input text for translation
Term identification
Term inflected form generation
Term translation ranking
Document translation
Translated Text
Bilingual term collection
Pre-trained SMT model

Figure 33. Source text pre-processing for terminology integration in SMT system translation level

Further, section 5.1 describes related work on dynamic terminology integration in SMT systems, sections 5.2, 5.3, and 5.4 describe the different components involved in the source text pre-processing workflow, section 5.5 describes evaluation efforts and evaluation results, and section 5.6 gives a summary of the methods for dynamic terminology integration in SMT.

5.1. Related Work on Dynamic Terminology Integration in SMT

In recent years considerable research efforts have been spent on methods for integration of term collections in SMT that do not require re-training of SMT systems. This section describes the most relevant topics in related research to the author’s work.

A common terminology translation issue is that terms are often not found in phrase-based SMT translation models or the translation models contain out-of-domain translation equivalents. This issue can be solved if SMT systems provide runtime integration with existing terminology databases or term collections provided by users. Carl & Langlais (2002) in their research used term dictionaries to pre-process the source text and achieved an increase in translation quality for the English-French language pair. Babych & Hartley (2003) showed that inclusion of certain named entities (namely, organisation names) in special “*do-not-translate*” lists allowed to increase translation quality for the English-Russian language pair using a pre-processing method that restricts translation of the identified phrases.

The popular Moses SMT platform also supports input data in a format (the Moses XML format) that can be enriched with externally generated translation candidates for phrases. In a recent work for English-Italian (in both translation directions), Arcan et al. (2014a) identify exactly matched terms and provide translation equivalents from the Wiki Machine[24] using the

[24] The Wiki Machine is available online at: https://bitbucket.org/fbk/thewikimachine.

Moses XML format. If there are multiple translation equivalents for a term, they perform context-based disambiguation using the source context and the relevant Wikipedia documents. Their evaluation results show an SMT quality improvement of up to 2 BLEU points.

Although not directly applicable for terminology integration in SMT, a promising method for NE integration in SMT systems has been proposed by Okuma et al. (2008). They substitute less frequent NEs (e.g., place names, organisations, person names, etc.) with frequent NEs, which are more likely to be present in SMT system phrase tables and translate the substituted text with an SMT engine. After translation they substitute the NEs back with the translation of the less frequent NEs using a one to one dictionary look-up. This method is not applicable to terms, because named entities of an equal NE category can be easily exchanged to different NEs of the same category in contexts, but terms in general are not grouped in categories.

A hot topic due to the increasing popularity of post-editing technologies has recently been the development of dynamic translation and language models for online adaptation of SMT systems (Bertoldi, 2014). Recently Arcan et al. (2014b) have shown that for English-Italian terminology can be successfully integrated in SMT systems using dynamic translation models.

However, these methods have been investigated either for languages that feature limited morphology or phrases that are left untranslated (e.g., many company and organisation names). The study in the FP7 project TTC (2013) showed that for English-Latvian such simplified methods do not yield positive results. Hálek et al. (2011) also showed that the translation performance with on-line pre-processing drops according to BLEU for English-Czech named entity translation. This proves that the method is not stable when translating into morphologically rich languages (e.g., the Baltic and Slavic languages). For such languages, the task of terminology translation requires development of more linguistically rich methods.

5.2. Identification of Terms in the Source Text

The first task that has to be performed when pre-processing the source text using a bilingual term collection is to identify terms in the source text. For this purpose the three methods described in section 1 can be used. The three methods are: 1) the linguistically and statistically motivated term identification using *TWSC* (Pinnis et al., 2012; described in section 2.2), 2) the linguistically motivated term identification using the *Pattern-Based Term Identification* (escribed in section 2.3), and 3) the *Fast Term Identification* using minimal linguistic support (in the form of stemming tools) in order to identify terms in different inflected forms. As explained in the respective sections, all three methods have positive and negative aspects. In the further sections all three methods will be evaluated to identify, which method achieves the best results in the dynamic terminology integration scenario.

5.3. Inflected Form Generation for the Identified Terms

The next pre-processing step after term identification is the generation of inflected forms for the identified terms. Previous research (Nikoulina et al., 2012; Carl & Langlais, 2002; Babych & Hartley, 2003, and others) on source text pre-processing methods has not given special attention to this task, because the bilingual term collections already "*provide*" translation equivalents. However, the issue is that the terms that are provided in the bilingual term collections are usually in their canonical forms and the canonical forms may often not be the required inflected forms in various contexts. Previous research has not seen the need to address these issues, because of the focus on language pairs that do not require (or require very limited) morphological generation (e.g., English-French, English-Italian, etc.). Therefore, to address this issue, the following sub-sections will present several approaches that allow to generate inflected forms for bilingual terms:

- The first method (see section 5.3.1) does not perform morphological generation of inflected forms. It is intended as a baseline for the inflected form generation methods.
- The second method (see section 5.3.2) uses morphological synthesis and language dependent inflected form generation rules to generate inflected forms for terms in canonical forms. Because the rules are language dependent, this method has been investigated just for Latvian and English.
- The third method (see section 5.3.3) uses a monolingual corpus in the target language and identifies inflected forms using stemming tools (similarly to the *Fast Term Identification).*

5.3.1. No Inflected Form Generation

To evaluate the impact of canonical forms on the translation quality when performing source text pre-processing, the first inflected form generation method relies only on the translation equivalents from the bilingual term collections. This method is used as a baseline method to show whether additional inflected form generation can achieve better results.

Because bilingual term collections can also be acquired using automatic bilingual term mapping methods, e.g., the *MPAligner* (Pinnis, 2013; described in section 3), the *USFD Term Aligner* (Aker et al., 2013), or *TEA* (Ştefănescu, 2012), the terms can also be given in the most common inflected forms found in various contexts. The evaluation results in section 5.5 will show that for term collections that have been acquired using bilingual terminology extraction methods, inflected form generation is not necessary. However, for term collections that contain terms in their canonical forms, inflected form generation is important.

5.3.2. Rule-based Morphological Synthesis

The second method (and the first real method) for inflected form generation for terms is based on morphological synthesis. For each target term from the bilingual term collection, the method performs the following steps to generate a list of inflected forms:

- First, we perform morphological analysis of terms, which do not contain morpho-syntactic information (morpho-syntactic information is usually included in term collections that have been automatically extracted from parallel or comparable corpora using the bilingual term extraction methods and tools described in sections 1 and 3). For each token of a term we acquire a list of possible morpho-syntactic tags and lemmas. For instance, the Table 30 shows the morphological information acquired for the Latvian term "*tīmekļa lapu*"[25] ("*Web page*" in English), using Tilde's morphological analyser for Latvian. The term is not in its canonical form, but in an inflected form (the canonical form would be "*tīmekļa lapa*").

Table 30. Morphological information acquired for the Latvian term "*tīmekļa lapu*" from the Tilde's morphological analyser of Latvian

Token	**POS**	**Lemma**	**Morpho-syntactic tag**
tīmekļa	N	tīmeklis	`N-msg---------n-----------l-`
lapu	N	lapa	`N-fpg---------n-----------l-`
lapu	N	lapa	`N-fsa---------n-----------l-`

- Then, based on the morphological analysis all morpho-syntactic term patterns (from *TWSC*) that may correspond to any sequence of the morpho-syntactic tags of the term's tokens are identified. The goal of this step is to identify the morpho-syntactic structure of multi-word terms. Single word terms are usually matched to their part-of-speech containing morpho-syntactic patterns. For the Latvian term "*tīmekļa lapu*" the only matching term pattern from the pattern list is "`^N...g.* ^N.*`". The pattern defines a two-word term consisting of two nouns. The first noun is in a genitive case, but the second noun is allowed to be in any inflected form.
- Next, we identify a morpho-syntactic inflection rule. Each term pattern has to have a manually defined morpho-syntactic inflection rule assigned to it. For the Latvian term "*tīmekļa lapu*" the inflection rule is as follows: "`**************************0* ***00*********************0*`". The morpho-syntactic inflection rule specifies that the first token has to be kept as is (the only change that can be applied is

[25] "*Tīmekļa lapa*" is an information technology and data processing term that can be found in *EuroTermBank*: http://www.eurotermbank.com/search.aspx?text=t%C4%ABmek%C4%BCa%20lapa&langfrom=lv&langto=en&where=etb%20extres&advanced=false#pos=1.

capitalisation) and the second token can be inflected by changing the number and case of the noun.

- Further, all possible inflected forms for each token of the term are generated. Because the lemmas and the parts of speech of the tokens are known, a morphological synthesiser can be used to generate the inflected forms of the tokens (Table 31 shows inflected forms generated for the term "*tīmekļa lapa*").

Table 31. Inflected forms of words "*tīmeklis*" (web) and "*lapa*" (page)
using Tilde's Latvian morphological synthesiser

***tīmeklis* (noun)**		***lapa* (noun)**	
Inflected form	**Morpho-syntactic tag**	**Inflected form**	**Morpho-syntactic tag**
tīmekli	`N-msa---------n-----------l-`	*lapa*	`N-fsn---------n-----------l-`
tīmekli	`N-msv---------n-----------l-`	*lapa*	`N-fsv---------n-----------l-`
tīmeklim	`N-msd---------n-----------l-`	*lapai*	`N-fsd---------n-----------l-`
tīmeklis	`N-msn---------n-----------l-`	*lapas*	`N-fpa---------n-----------l-`
tīmeklī	`N-msl---------n-----------l-`	*lapas*	`N-fpn---------n-----------l-`
tīmeklīt	`N-msv---------y-----------l-`	*lapas*	`N-fpv---------n-----------l-`
tīmeklīti	`N-msa---------y-----------l-`	*lapas*	`N-fsg---------n-----------l-`
tīmeklīti	`N-msv---------y-----------l-`	*lapiņ*	`N-fsv---------y-----------l-`
tīmeklītim	`N-msd---------y-----------l-`	*lapiņa*	`N-fsn---------y-----------l-`
tīmeklītis	`N-msn---------y-----------l-`	*lapiņai*	`N-fsd---------y-----------l-`
tīmeklītī	`N-msl---------y-----------l-`	*lapiņas*	`N-fpa---------y-----------l-`
tīmeklīša	`N-msg---------y-----------l-`	*lapiņas*	`N-fpn---------y-----------l-`
tīmeklīši	`N-mpn---------y-----------l-`	*lapiņas*	`N-fpv---------y-----------l-`
tīmeklīši	`N-mpv---------y-----------l-`	*lapiņas*	`N-fsg---------y-----------l-`
tīmeklīšiem	`N-mpd---------y-----------l-`	*lapiņu*	`N-fpg---------y-----------l-`
tīmeklīšos	`N-mpl---------y-----------l-`	*lapiņu*	`N-fsa---------y-----------l-`
tīmeklīšu	`N-mpg---------y-----------l-`	*lapiņā*	`N-fsl---------y-----------l-`
tīmeklīšus	`N-mpa---------y-----------l-`	*lapiņām*	`N-fpd---------y-----------l-`
tīmekļa	`N-msg---------n-----------l-`	*lapiņās*	`N-fpl---------y-----------l-`
tīmekļi	`N-mpn---------n-----------l-`	*lapu*	`N-fpg---------n-----------l-`
tīmekļi	`N-mpv---------n-----------l-`	*lapu*	`N-fsa---------n-----------l-`
tīmekļiem	`N-mpd---------n-----------l-`	*lapā*	`N-fsl---------n-----------l-`
tīmekļos	`N-mpl---------n-----------l-`	*lapām*	`N-fpd---------n-----------l-`
tīmekļu	`N-mpg---------n-----------l-`	*lapās*	`N-fpl---------n-----------l-`
tīmekļus	`N-mpa---------n-----------l-`		

- Once the morpho-syntactic inflection rule and the inflected forms of the term's tokens are known, all possible combinations of the term's inflected forms can be generated. All valid combinations for the term "*tīmekļa lapu*" are given in Table 32. It is evident that just one inflected form of the first token qualifies while for the second token multiple inflected forms (in which only the number and case differs) qualify.

Table 32. Valid morpho-syntactic combinations for the term “*tīmekļa lapu*”

tīmeklis (noun)		**lapa (noun)**	
Inflected form	**Morpho-syntactic tag**	**Inflected form**	**Morpho-syntactic tag**
~~*tīmekli*~~	~~N-msa---------n-----------l-~~	*lapa*	N-fsn---------n-----------l-
~~*tīmekli*~~	~~N-msv---------n-----------l-~~	*lapa*	N-fsv---------n-----------l-
~~*tīmeklim*~~	~~N-msd---------n-----------l-~~	*lapai*	N-fsd---------n-----------l-
~~*tīmeklis*~~	~~N-msn---------n-----------l-~~	*lapas*	N-fpa---------n-----------l-
~~*tīmeklī*~~	~~N-msl---------n-----------l-~~	*lapas*	N-fpn---------n-----------l-
~~*tīmeklīt*~~	~~N-msv---------y-----------l-~~	*lapas*	N-fpv---------n-----------l-
~~*tīmeklīti*~~	~~N-msa---------y-----------l-~~	*lapas*	N-fsg---------n-----------l-
~~*tīmeklīti*~~	~~N-msv---------y-----------l-~~	~~*lapiņ*~~	~~N-fsv---------y-----------l-~~
~~*tīmeklītim*~~	~~N-msd---------y-----------l-~~	~~*lapiņa*~~	~~N-fsn---------y-----------l-~~
~~*tīmeklītis*~~	~~N-msn---------y-----------l-~~	~~*lapiņai*~~	~~N-fsd---------y-----------l-~~
~~*tīmeklītī*~~	~~N-msl---------y-----------l-~~	~~*lapiņas*~~	~~N-fpa---------y-----------l-~~
~~*tīmeklīša*~~	~~N-msg---------y-----------l-~~	~~*lapiņas*~~	~~N-fpn---------y-----------l-~~
~~*tīmeklīši*~~	~~N-mpn---------y-----------l-~~	~~*lapiņas*~~	~~N-fpv---------y-----------l-~~
~~*tīmeklīši*~~	~~N-mpv---------y-----------l-~~	~~*lapiņas*~~	~~N-fsg---------y-----------l-~~
~~*tīmeklīšiem*~~	~~N-mpd---------y-----------l-~~	~~*lapiņu*~~	~~N-fpg---------y-----------l-~~
~~*tīmeklīšos*~~	~~N-mpl---------y-----------l-~~	~~*lapiņu*~~	~~N-fsa---------y-----------l-~~
~~*tīmeklīšu*~~	~~N-mpg---------y-----------l-~~	~~*lapiņā*~~	~~N-fsl---------y-----------l-~~
~~*tīmeklīšus*~~	~~N-mpa---------y-----------l-~~	~~*lapiņām*~~	~~N-fpd---------y-----------l-~~
tīmekļa	N-msg---------n-----------l-	~~*lapiņās*~~	~~N-fpl---------y-----------l-~~
~~*tīmekļi*~~	~~N-mpn---------n-----------l-~~	*lapu*	N-fpg---------n-----------l-
~~*tīmekļi*~~	~~N-mpv---------n-----------l-~~	*lapu*	N-fsa---------n-----------l-
~~*tīmekļiem*~~	~~N-mpd---------n-----------l-~~	*lapā*	N-fsl---------n-----------l-
~~*tīmekļos*~~	~~N-mpl---------n-----------l-~~	*lapām*	N-fpd---------n-----------l-
~~*tīmekļu*~~	~~N-mpg---------n-----------l-~~	*lapās*	N-fpl---------n-----------l-
~~*tīmekļus*~~	~~N-mpa---------n-----------l-~~		

This method is very language dependent (because it requires a morphological analyser, morphological synthesiser, term patterns, and term morpho-syntactic inflection rules) and requires significant manual efforts in order to provide support for additional languages. Therefore, the next section describes a method that requires much less manual efforts to provide support for a new language.

5.3.3. Monolingual Corpus Look-up

For the languages for which the rule-based morphological synthesis method is not feasible, a language independent method for the acquisition of term translation equivalents (in different inflected forms) has been investigated. The method requires a large monolingual corpus in the target language and it performs a look-up (similarly to the way how terms are identified in the *Fast Term Identification* method described in section 2.4) for inflected forms for all terms in a given term collection. Of course, not all inflected forms for a term will be found, because: 1) the monolingual corpus may not be large enough to contain all inflected forms of infrequent terms, and 2) stemmers cannot substitute high quality lemmatisation and morpho-syntactic tagging tools (which means that not all inflected forms will be found even if they are given in the corpus). However, as the results in section 5.5 will show, it is sufficient to achieve SMT quality improvements.

5.4. Ranking the Translation Equivalents

Now that terms in the source text have been identified and the inflected forms have been generated, translation likelihood scores still have to be assigned to the translation hypotheses (i.e., the inflected forms). It is important to apply the ranking of translation hypotheses, because not all translation hypotheses are well suited in the observed contexts. In addition, some translation hypotheses are in general more common than others. Two methods have been investigated by the author for term translation candidate ranking:

- The first method assigns equal translation likelihood scores to all translation hypotheses of a term. This method is used as a baseline method for translation hypotheses ranking. When assigning equal weights to all translation hypotheses the language model is allowed to select the translation hypotheses. However, relying simply on the language model means that important statistics that come from a translation model (e.g., source to target language transfer information) are lost. We also lose important information from the source language's context as that could help identifying, which translation hypotheses is more likely in a given context.
- The second method uses a large monolingual corpus to rank translation hypotheses. For each translation hypothesis its relative frequency between all the translation hypotheses of a source language term is assigned. Only exact match phrases are counted for the translation hypotheses. This method allows assigning higher scores for more common translation hypotheses.

5.5. Evaluation Scenarios and Results

In total, three different evaluation experiments were performed to evaluate the dynamic terminology integration methods:

- Automatic evaluation in the automotive domain using broad domain SMT systems and standard SMT system evaluation metrics (i.e., NIST (Doddington, 2002), BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), and TER (Snover et al., 2006)). The evaluation is described further in section 5.5.1.
- Manual comparative system evaluation in the automotive domain. In this experiment, professional translators had to compare translations of sentences produced by the same broad domain SMT systems as in the automatic evaluation without integrated terminology and with integrated terminology. The best performing dynamic terminology integration scenario, which was identified in the automatic evaluation,

was used for terminology integration in this evaluation experiment. The evaluation is described further in section 5.5.2.

- Manual comparative system evaluation and term translation quality evaluation in the information technology domain. In this experiment, professional translators had to compare translations of sentences and terms, which were found in the evaluated sentences, produced by in-domain SMT systems without and with integrated terminology. Additionally, translators had to rate the translation quality of each term. The evaluation is described further in section 5.5.3.

5.5.1. Automatic Evaluation

The automatic evaluation experiments were performed for four language pairs: English-Latvian, English-Lithuanian, English-Estonian, and English-German. The baseline systems for all four language pairs have been trained using the *DGT-TM* parallel corpus (the releases of 2007, 2011, and 2012). For English-Latvian an in-domain tuning set of 1,745 sentence pairs was available; for the remaining systems held-out sets of 2,000 sentence pairs were used. For evaluation, 872 sentence pairs were used for each of the language pairs. The evaluation set is comprised of car service manual translation segments. The original data set was available for English-Latvian, therefore, the evaluation data for the remaining three language pairs were prepared by professional translators (i.e., the translators translated the 872 English sentences into the three remaining languages).

Similarly to the evaluation data, the English terms from the *professional* English-Latvian term collection were translated into the three remaining target languages. Due to language specific characteristics and different translators involved in the term collection creation process, the *professional* term collections consisted of 662 term pairs for English-Lithuanian, 619 term pairs for English-Estonian, and 692 term pairs for English-German.

The automatic evaluation experiments are split in seven dimensions depending on the different methods investigated in the pre-processing and SMT integration workflow's sub-processes and depending on the data used for pre-processing of the source text:

- The **term collection** used for pre-processing:
    - *Non-filtered* – a raw bilingual term collection automatically extracted from parallel corpora (tuning data of the SMT system).
    - *Filtered* – the raw bilingual term collection manually revised by deleting general language phrases and wrong translations.
    - *Professional* – a bilingual term collection manually created by a professional translator.

- The **term identification method**:
  - *TWSC* – the *TWSC-based Term Identification* (see section 2.2).
  - *Fast* – the *Fast Term Identification* (see section 2.4).
  - *Pattern* – the *Pattern-based Term Identification* (see section 2.3).
- The **inflected form generation method**:
  - *None* – only the translation equivalents that are present in the bilingual term collection are used as translation equivalents, i.e., we do not generate or acquire any other translation equivalents (see section 5.3.1).
  - *Synthesis* – the *Rule-based Morphological Synthesis* (see section 5.3.2).
  - *Corpus* – the *Monolingual Corpus Look-up* (see section 5.3.3). For the acquisition of inflected forms for terms, the broad domain monolingual corpus of the SMT system was used.
  - *Combined* – the combination of *Synthesis* and *Corpus* methods. Because the *Synthesis* method does not always produce translation equivalents (e.g., for words unknown to the morphological analyser), the combination of the two methods (one that acquires through generation and one that acquires through look-up) we can identify more translation equivalents for the terms.
- The **monolingual corpus** from which inflected forms of terms have been extracted for the *Corpus* and *Combined* methods:
  - *In-domain corpus* – the Web crawled in-domain corpus described in section 4.2.1.1.
  - *Broad domain corpus* – the *DGT-TM* monolingual target language corpus.
  - *In-domain and broad domain corpora combined*.
- The **translation equivalent ranking method** (see section 5.4):
  - *Equal* – every translation equivalent of a source term gets an equal translation likelihood score assigned.
  - *Simple* –translation equivalent translation likelihood scores are assigned based on the translation equivalent relative frequencies in a large monolingual corpus (the broad domain corpus in our experiments).
- The *Moses* SMT platform allows treating translation equivalents in the XML input documents as "*exclusive*" (that is, to select a translation only from the equivalents specified in the XML document) or "*inclusive*" (that is, to allow the translation equivalents specified in the XML document to compete with translation equivalents from the SMT system's phrase table). The "*exclusive*" decoding option can ensure terminology translation consistency and that only in-domain translation hypotheses will be selected. However, if a term collection is ambiguous, then restricting an SMT

system to just the pre-defined set of translation equivalents from a term collection can actually have a negative effect on translation quality. For instance, the English word "*application*" could represent a term from the IT domain (e.g., a computer program), public administration domain (e.g., a formal document signed and submitted to someone for a purpose), it can be a participle describing an action, etc. Therefore, depending on the level of ambiguity in the lexical forms of terms in a term collection, both the "inclusive" or "exclusive" options for translation hypotheses selection can be beneficial (as also shown by the experiment results further).

When terms are identified in the source text the translated content can be POS-tagged and lemmatised to find terms in their different inflected forms. When lemmas are available, the *Fast Term Identification* method is based on searching for matching lemma sequences instead of stemmed inflected form sequences. This allows to identify more and linguistically more reliable inflected forms than with the stemming-based approach (which can also identify forms with spelling mistakes). However, as lemmatisers are usually based on a lexicon, they have a limited vocabulary, which means that the stemming-based approach can identify inflected forms for terms that contain words not covered by the lemmatisers. In the experiment results the scenarios with POS-tagging support have been marked with "*POS*". Note that the "*Pattern*" and "*TWSC*" based term identification methods require POS-tagging to perform term identification.

5.5.1.1. English-Latvian

The first English-Latvian experiments reported in this section were performed using the baseline system that was used in the *Terminology as a Corpus* experiments (see section 4.3.1).

The results are distributed in three tables based on the type of term collection used:

- Table 33 provides results for pre-processing experiments with non-filtered terms.
- Table 34 provides results for pre-processing experiments with filtered terms.
- Table 35 provides results for pre-processing experiments with the term collection created by a professional translator.

Each table provides the results of the baseline system and the different pre-processing experiments from the first five evaluation scenario dimensions (at the time of evaluation, the last two dimensions were not performed, however see below for experiment results with a larger language model where also these dimensions have been taken into account). For each pre-processing scenario, the results provide also a score showing the change over the baseline system, i.e., translation quality increase or decrease according to the BLEU metric.

Table 33. English-Latvian results of dynamic terminology integration in SMT systems using the non-filtered term collection

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	4.2644	12.68	0.1849	0.7801	4.1361	12.00	0.1439	0.7893	
Non-filtered terms and no surface form analysis									
Fast + None + Equal	**4.6386**	**15.34**	**0.2104**	**0.7507**	**4.3407**	**13.21**	**0.1619**	**0.7743**	21.0%
Pattern + None + Equal	4.5031	14.22	0.2014	0.7599	4.2256	12.34	0.1531	0.7815	12.1%
TWSC + None + Equal	4.4821	14.01	0.1966	0.7614	4.2864	12.72	0.1526	0.7765	10.5%
Non-filtered terms and morphological synthesis based surface form analysis									
Fast + Synthesis + Equal	4.4141	13.46	0.1974	0.7686	4.1513	11.99	0.1511	0.7905	6.2%
Fast + Synthesis + Simple	4.4283	13.63	0.1997	0.7674	4.1634	12.18	0.1532	0.7901	7.5%
Pattern + Synthesis + Equal	4.3811	13.45	0.1931	0.7702	4.1319	12.11	0.1469	0.7901	6.1%
Pattern + Synthesis + Simple	4.3964	13.64	0.1946	0.7693	4.1425	12.30	0.1482	0.7898	7.6%
TWSC + Synthesis + Equal	4.3905	13.41	0.1919	0.7683	4.2142	12.48	0.1488	0.7819	5.8%
TWSC + Synthesis + Simple	4.3899	13.44	0.1923	0.7685	4.2130	12.50	0.1493	0.7822	6.0%
Non-filtered terms and in-domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.3121	12.97	0.1908	0.7785	4.0683	11.65	0.1459	0.7981	2.3%
Fast + Combined + Simple	4.4433	13.53	0.1981	0.7682	4.1709	12.01	0.1509	0.7905	6.7%
Fast + Corpus + Equal	4.5067	14.39	0.2002	0.7630	4.2398	12.63	0.1538	0.7846	13.5%
Fast + Corpus + Simple	4.5935	14.72	0.2047	0.7554	4.2998	12.81	0.1568	0.7798	16.1%
Pattern + Combined + Equal	4.2681	12.78	0.1861	0.7810	4.0369	11.55	0.1413	0.7989	0.8%
Pattern + Combined + Simple	4.3733	13.30	0.1924	0.7714	4.1167	11.89	0.1457	0.7918	4.9%
Pattern + Corpus + Equal	4.4246	13.98	0.1933	0.7682	4.1736	12.39	0.1475	0.7877	10.3%
Pattern + Corpus + Simple	4.4971	14.21	0.1975	0.7614	4.2228	12.46	0.1503	0.7834	12.1%
TWSC + Combined + Equal	4.3580	13.22	0.1892	0.7711	4.1908	12.28	0.1470	0.7837	4.3%
TWSC + Combined + Simple	4.4074	13.50	0.1921	0.7678	4.2243	12.45	0.1488	0.7815	6.5%
TWSC + Corpus + Equal	4.4226	13.78	0.1924	0.7659	4.2471	12.59	0.1502	0.7793	8.7%
TWSC + Corpus + Simple	4.4730	14.15	0.1947	0.7622	4.2790	12.84	0.1514	0.7770	11.6%
Non-filtered terms and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.3520	13.16	0.1933	0.7743	4.0915	11.76	0.1478	0.7957	3.8%
Fast + Combined + Simple	4.4379	13.78	0.1982	0.7689	4.1730	12.31	0.1515	0.7911	8.7%
Fast + Corpus + Equal	4.5813	14.95	0.2050	0.7579	4.2891	12.97	0.1568	0.7818	17.9%
Fast + Corpus + Simple	4.6110	15.06	0.2069	0.7563	4.3175	13.07	0.1585	0.7805	18.8%
Pattern + Combined + Equal	4.3528	13.30	0.1911	0.7719	4.1061	11.97	0.1457	0.7911	4.9%
Pattern + Combined + Simple	4.4038	13.58	0.1944	0.7693	4.1495	12.21	0.1478	0.7895	7.1%
Pattern + Corpus + Equal	4.5422	14.58	0.2003	0.7583	4.2640	12.76	0.1526	0.7802	15.0%
Pattern + Corpus + Simple	4.5547	14.57	0.2016	0.7575	4.2735	12.72	0.1536	0.7797	14.9%
TWSC + Combined + Equal	4.3805	13.29	0.1909	0.7686	4.2072	12.37	0.1483	0.7817	4.8%
TWSC + Combined + Simple	4.4016	13.41	0.1922	0.7681	4.2215	12.47	0.1490	0.7817	5.8%
TWSC + Corpus + Equal	4.4591	13.79	0.1946	0.7635	4.2675	12.57	0.1509	0.7783	8.8%
TWSC + Corpus + Simple	4.4636	13.79	0.1951	0.7632	4.2665	12.57	0.1511	0.7786	8.8%
Non-filtered terms and in-domain and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.2685	12.76	0.1884	0.7834	4.0254	11.50	0.1438	0.8029	0.6%
Fast + Combined + Simple	4.4329	13.67	0.1974	0.7699	4.1655	12.22	0.1506	0.7917	7.8%
Fast + Corpus + Equal	4.4466	14.12	0.1970	0.7691	4.1829	12.46	0.1510	0.7905	11.4%
Fast + Corpus + Simple	4.5538	14.65	0.2029	0.7599	4.2732	12.88	0.1558	0.7830	15.5%
Pattern + Combined + Equal	4.2596	12.79	0.1857	0.7818	4.0285	11.56	0.1412	0.7995	0.9%
Pattern + Combined + Simple	4.3725	13.35	0.1924	0.7717	4.1246	12.05	0.1462	0.7911	5.3%
Pattern + Corpus + Equal	4.3984	13.82	0.1921	0.7703	4.1498	12.28	0.1465	0.7895	9.0%
Pattern + Corpus + Simple	4.4682	14.06	0.1964	0.7639	4.2105	12.49	0.1501	0.7843	10.9%
TWSC + Combined + Equal	4.3450	13.13	0.1887	0.7722	4.1780	12.19	0.1467	0.7846	3.5%
TWSC + Combined + Simple	4.3895	13.42	0.1912	0.7697	4.2174	12.45	0.1485	0.7823	5.8%
TWSC + Corpus + Equal	4.4073	13.68	0.1916	0.7670	4.2320	12.49	0.1495	0.7802	7.9%
TWSC + Corpus + Simple	4.4324	13.87	0.1929	0.7659	4.2541	12.69	0.1504	0.7793	9.4%

Table 34. English-Latvian results of dynamic terminology integration in SMT systems using the filtered term collection

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	4.2644	12.68	0.1849	0.7801	4.1361	12.00	0.1439	0.7893	
Filtered terms and no surface form analysis									
Fast + None + Equal	4.5860	15.16	**0.2053**	**0.7495**	4.3307	13.13	**0.1590**	**0.7693**	19.6%
Pattern + None + Equal	4.4836	14.47	0.200507	0.7604	4.2445	12.78	0.15446	0.7787	14.1%
TWSC + None + Equal	4.4686	14.25	0.196788	0.7635	4.2677	12.83	0.15249	0.7791	12.4%
Filtered terms and morphological synthesis based surface form analysis									
Fast + Synthesis + Equal	4.3777	13.49	0.1917	0.7662	4.1856	12.53	0.1486	0.7811	6.4%
Fast + Synthesis + Simple	4.3946	13.69	0.1939	0.7642	4.2000	12.75	0.1506	0.7799	8.0%
Pattern + Synthesis + Equal	4.3376	13.26	0.1896	0.7710	4.1576	12.39	0.1469	0.7846	4.6%
Pattern + Synthesis + Simple	4.3558	13.49	0.1913	0.7691	4.1722	12.62	0.1483	0.7835	6.4%
TWSC + Synthesis + Equal	4.3526	13.22	0.1896	0.7713	4.1982	12.49	0.1478	0.7831	4.3%
TWSC + Synthesis + Simple	4.3520	13.25	0.1902	0.7711	4.1945	12.51	0.1482	0.7835	4.5%
Filtered terms and in-domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.2503	12.83	0.1840	0.7770	4.0836	11.96	0.1428	0.7898	1.2%
Fast + Combined + Simple	4.3569	13.38	0.1915	0.7679	4.1688	12.35	0.1492	0.7827	5.5%
Fast + Corpus + Equal	4.4509	14.65	0.1953	0.7612	4.2484	13.14	0.1523	0.7765	15.5%
Fast + Corpus + Simple	4.5111	14.77	0.1985	0.7562	4.3030	13.29	0.1554	0.7721	16.5%
Pattern + Combined + Equal	4.2022	12.54	0.1817	0.7825	4.0454	11.76	0.1406	0.7941	-1.1%
Pattern + Combined + Simple	4.3005	13.09	0.1883	0.7741	4.1250	12.17	0.1462	0.7874	3.2%
Pattern + Corpus + Equal	4.3859	14.20	0.1914	0.7678	4.1939	12.89	0.1487	0.7819	12.0%
Pattern + Corpus + Simple	4.4253	14.10	0.1945	0.7648	4.2329	12.94	0.1516	0.7794	11.2%
TWSC + Combined + Equal	4.3205	13.05	0.1873	0.7739	4.1711	12.32	0.1461	0.7853	2.9%
TWSC + Combined + Simple	4.3513	13.33	0.1896	0.7726	4.1869	12.43	0.1477	0.7849	5.1%
TWSC + Corpus + Equal	4.4102	14.10	0.1926	0.7670	4.2350	12.86	0.1503	0.7802	11.2%
TWSC + Corpus + Simple	4.4280	14.09	0.1935	0.7668	4.2519	12.96	0.1511	0.7799	11.1%
Filtered terms and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.3447	13.35	0.1896	0.7691	4.1598	12.47	0.1472	0.7834	5.3%
Fast + Combined + Simple	4.4231	13.85	0.1944	0.7620	4.2296	12.90	0.1515	0.7778	9.2%
Fast + Corpus + Equal	4.5646	15.18	0.2014	0.7542	4.3278	13.46	0.1562	0.7726	19.7%
Fast + Corpus + Simple	**4.5879**	**15.28**	0.2032	0.7519	**4.3513**	**13.64**	0.1580	0.7709	20.5%
Pattern + Combined + Equal	4.3084	13.17	0.1878	0.7735	4.1337	12.32	0.1457	0.7865	3.9%
Pattern + Combined + Simple	4.3730	13.60	0.1921	0.7676	4.1893	12.72	0.1494	0.7821	7.3%
Pattern + Corpus + Equal	4.5161	14.86	0.1990	0.7592	4.2885	13.31	0.1540	0.7766	17.2%
Pattern + Corpus + Simple	4.5389	14.95	0.2009	0.7568	4.3118	13.46	0.1558	0.7747	17.9%
TWSC + Combined + Equal	4.3445	13.15	0.1889	0.7723	4.1922	12.42	0.1474	0.7838	3.7%
TWSC + Combined + Simple	4.3645	13.28	0.1906	0.7707	4.2062	12.53	0.1487	0.7829	4.7%
TWSC + Corpus + Equal	4.4486	14.22	0.1947	0.7654	4.2545	12.86	0.1509	0.7805	12.1%
TWSC + Corpus + Simple	4.4588	14.25	0.1956	0.7642	4.2600	12.89	0.1517	0.7798	12.4%
Filtered terms and in-domain and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.2352	12.78	0.1833	0.7789	4.0741	11.97	0.1424	0.7910	0.8%
Fast + Combined + Simple	4.3704	13.61	0.1914	0.7675	4.1819	12.59	0.1492	0.7822	7.3%
Fast + Corpus + Equal	4.4120	14.47	0.1926	0.7647	4.2154	13.01	0.1499	0.7791	14.1%
Fast + Corpus + Simple	4.4762	14.62	0.1970	0.7592	4.2737	13.29	0.1541	0.7747	15.3%
Pattern + Combined + Equal	4.1918	12.55	0.1812	0.7838	4.0381	11.77	0.1404	0.7949	-1.0%
Pattern + Combined + Simple	4.3151	13.32	0.1890	0.7730	4.1389	12.40	0.1471	0.7863	5.0%
Pattern + Corpus + Equal	4.3545	14.05	0.1898	0.7705	4.1663	12.77	0.1473	0.7839	10.8%
Pattern + Corpus + Simple	4.4168	14.21	0.1943	0.7650	4.2265	13.08	0.1517	0.7793	12.1%
TWSC + Combined + Equal	4.3061	12.96	0.1866	0.7757	4.1599	12.24	0.1457	0.7865	2.2%
TWSC + Combined + Simple	4.3484	13.34	0.1895	0.7727	4.1858	12.44	0.1476	0.7847	5.2%
TWSC + Corpus + Equal	4.3941	14.01	0.1916	0.7683	4.2214	12.79	0.1495	0.7810	10.5%
TWSC + Corpus + Simple	4.4033	14.00	0.1926	0.7685	4.2288	12.88	0.1502	0.7813	10.4%

Table 35. English-Latvian results of dynamic terminology integration in SMT systems using the professional term collection

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	4.2644	12.68	0.1849	0.7801	4.1361	12.00	0.1439	0.7893	
Professional terms and no surface form analysis									
Fast + None + Equal	4.6976	15.55	0.21215	0.7415	4.4487	13.55	0.16539	0.7607	22.6%
Pattern + None + Equal	4.5580	14.55	0.205381	0.7547	4.3227	12.78	0.15907	0.7730	14.7%
TWSC + None + Equal	4.4790	14.12	0.197025	0.7594	4.2899	12.67	0.15351	0.7739	11.4%
Professional terms and morphological synthesis based surface form analysis									
Fast + Synthesis + Equal	4.5204	13.67	0.1988	0.7535	4.3398	12.77	0.1552	0.7695	7.8%
Fast + Synthesis + Simple	4.5307	13.66	0.2012	0.7522	4.3480	12.70	0.1575	0.7687	7.7%
Pattern + Synthesis + Equal	4.4546	13.46	0.1955	0.7611	4.2808	12.64	0.1519	0.7762	6.2%
Pattern + Synthesis + Simple	4.4636	13.43	0.1968	0.7602	4.2857	12.56	0.1529	0.7758	5.9%
TWSC + Synthesis + Equal	4.3934	13.34	0.1923	0.7654	4.2435	12.57	0.1500	0.7774	5.2%
TWSC + Synthesis + Simple	4.3892	13.27	0.1928	0.7651	4.2364	12.44	0.1506	0.7777	4.7%
Professional terms and in-domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.4375	13.24	0.1942	0.7599	4.2690	12.43	0.1516	0.7746	4.4%
Fast + Combined + Simple	4.5125	13.81	0.1984	0.7551	4.3374	12.90	0.1554	0.7702	8.9%
Fast + Corpus + Equal	4.6343	14.75	0.2052	0.7458	4.4469	13.76	0.1614	0.7614	16.3%
Fast + Corpus + Simple	4.7188	15.25	0.2091	0.7412	4.5257	14.10	0.1642	0.7568	20.3%
Pattern + Combined + Equal	4.3706	13.01	0.1910	0.7678	4.2077	12.28	0.1485	0.7815	2.6%
Pattern + Combined + Simple	4.4419	13.57	0.1953	0.7632	4.2687	12.74	0.1523	0.7777	7.0%
Pattern + Corpus + Equal	4.5455	14.44	0.2006	0.7558	4.3625	13.53	0.1569	0.7706	13.9%
Pattern + Corpus + Simple	4.6288	14.88	0.2047	0.7512	4.4397	13.86	0.1600	0.7660	17.4%
TWSC + Combined + Equal	4.3571	13.14	0.1902	0.7686	4.2087	12.38	0.1482	0.7801	3.6%
TWSC + Combined + Simple	4.3964	13.40	0.1927	0.7664	4.2416	12.60	0.1501	0.7785	5.7%
TWSC + Corpus + Equal	4.4564	13.89	0.1954	0.7616	4.2934	13.04	0.1524	0.7742	9.5%
TWSC + Corpus + Simple	4.4974	14.00	0.1977	0.7600	4.3293	13.08	0.1538	0.7729	10.4%
Professional terms and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.4834	13.46	0.1967	0.7575	4.3071	12.60	0.1534	0.7729	6.2%
Fast + Combined + Simple	4.5477	13.76	0.2016	0.7534	4.3698	12.85	0.1580	0.7687	8.5%
Fast + Corpus + Equal	4.7929	15.98	0.2125	0.7376	4.5568	14.26	0.1654	0.7563	26.0%
Fast + Corpus + Simple	**4.8187**	**16.09**	**0.2138**	**0.7360**	**4.5876**	**14.42**	**0.1664**	**0.7551**	26.9%
Pattern + Combined + Equal	4.4268	13.29	0.1939	0.7639	4.2571	12.49	0.1508	0.7782	4.8%
Pattern + Combined + Simple	4.4816	13.54	0.1975	0.7612	4.3063	12.70	0.1535	0.7761	6.8%
Pattern + Corpus + Equal	4.6680	15.13	0.2066	0.7500	4.4493	13.61	0.1600	0.7676	19.3%
Pattern + Corpus + Simple	4.6924	15.22	0.2080	0.7483	4.4764	13.73	0.1611	0.7664	20.0%
TWSC + Combined + Equal	4.4050	13.36	0.1924	0.7656	4.2569	12.59	0.1502	0.7771	5.4%
TWSC + Combined + Simple	4.4178	13.29	0.1940	0.7647	4.2652	12.47	0.1511	0.7770	4.8%
TWSC + Corpus + Equal	4.5186	14.23	0.1982	0.7587	4.3302	12.87	0.1539	0.7734	12.2%
TWSC + Corpus + Simple	4.5317	14.24	0.1991	0.7572	4.3430	12.88	0.1547	0.7723	12.3%
Professional terms and in-domain and broad domain monolingual corpus for surface form analysis									
Fast + Combined + Equal	4.4139	13.15	0.1929	0.7622	4.2478	12.37	0.1507	0.7766	3.7%
Fast + Combined + Simple	4.5090	13.79	0.1985	0.7555	4.3385	12.94	0.1554	0.7702	8.8%
Fast + Corpus + Equal	4.5995	14.49	0.2032	0.7490	4.4161	13.62	0.1595	0.7640	14.3%
Fast + Corpus + Simple	4.6852	14.88	0.2075	0.7438	4.4956	13.90	0.1623	0.7592	17.4%
Pattern + Combined + Equal	4.3511	12.95	0.1899	0.7694	4.1902	12.23	0.1478	0.7827	2.1%
Pattern + Combined + Simple	4.4384	13.57	0.1954	0.7636	4.2690	12.79	0.1522	0.7778	7.0%
Pattern + Corpus + Equal	4.5179	14.25	0.1991	0.7583	4.3385	13.43	0.1555	0.7726	12.4%
Pattern + Corpus + Simple	4.5996	14.58	0.2034	0.7535	4.4134	13.70	0.1583	0.7683	15.0%
TWSC + Combined + Equal	4.3546	13.15	0.1898	0.7691	4.2084	12.40	0.1480	0.7802	3.7%
TWSC + Combined + Simple	4.3981	13.41	0.1929	0.7663	4.2441	12.61	0.1503	0.7783	5.8%
TWSC + Corpus + Equal	4.4387	13.76	0.1946	0.7630	4.2789	12.97	0.1518	0.7750	8.5%
TWSC + Corpus + Simple	4.4816	13.88	0.1972	0.7610	4.3184	13.03	0.1535	0.7735	9.5%

The results show that all combinations produced results that exceed the results of the baseline system. The average results for the *non-filtered* term collection in terms of BLEU were lower than for the manually *filtered* term collection, which on average allowed achieving 0.25% higher results over all evaluation scenarios, and much lower than using the *professional* term collection, which on average allowed achieving 1.6% higher results. The best overall results were achieved using the *professional* term collection (on average 1.35% higher results compared to the results achieved with the *filtered* term collection).

It is also evident that the *Fast Term Identification* allows achieving better results than the other term identification methods. This is mainly because it identifies significantly more terms in the translatable content (1,404; compared to 1,261 for the *Pattern-Based Term Identification* and 620 for the *TWSC-Based Term Identification*). This is a very positive result as we can achieve the highest performance (in terms of speed) and still maintain the best quality. However, as noted earlier, this can backfire if the terms in the term collection are morphologically ambiguous (e.g., if a word is translated differently if it is a noun or a verb and the term collection contains only the translation of the noun).

For translation equivalent acquisition four different methods were applied:

- The first method used just the translation equivalents from the term collection. For the evaluation scenarios based on the *non-filtered* term collection, this method achieved the best result. This is because the automatically extracted term collection already contains terms in their potential inflected forms. By generating additional inflected forms we create a larger ambiguity and consequently the translation quality drops. For the *filtered* and *professional* term collection (especially the professional term collection), this method did not achieve the highest results, however the results were still better than for many of the evaluation scenarios. There are multiple reasons for the relatively good performance. For instance, the term collections contain many multi-word term pairs (from over 70% in the *non-filtered* collection to just below 58% in the *professional* collection), for which not all words of the multi-word terms are affected by inflection when generating different inflected forms (often just the head word is inflected). Because the baseline system's score is relatively low (just 12.68 BLEU points), translation quality improves by translating correctly only a part of the multi-word terms. Another reason is related to search space. Having term translations in their canonical forms means that in contexts where the canonical forms are required we will have a 100% precision (of course if the terms have just one translation equivalent specified in the term collection). However, if we generate multiple inflected forms, the SMT system has a higher possibility of selecting an incorrect translation.

- The *Rule-Based Morphological Synthesis* did not achieve the best overall results, however it still exceeded the baseline system's results in all evaluation scenarios. It is evident that when generating all possible inflected forms, it is crucial to be able to rank the translation equivalents by taking the source context into account. It can be beneficial to drop the least likely translation equivalents as the high ambiguity makes it difficult for the SMT system to select the correct form.
- The *Monolingual Corpus Look-up* allowed achieving the best results. For the *filtered* and *professional* term collections, it even outperforms the scenarios without inflected form generation (by 0.54 BLEU points for the *professional* term collection). This proves that by generating inflected forms we can achieve a higher translation quality.
- The *Combined* method achieved the lowest results. The author believes that the low results are caused by the high ambiguity that results when combining the two different inflected form generation methods.

The *Monolingual Corpus Look-up* method for inflected form generation requires a monolingual corpus. In the author's experiments three different corpora were investigated. The best results were achieved with the *DGT-TM* monolingual corpus of the target language. Experiments with the in-domain corpus achieved the lowest results. Consequently, the combination of both corpora achieved better results than using just the in-domain corpus, however, lover results than with the broad domain corpus. Two possible explanations for the lower results are: 1) the in-domain corpus contains many spelling mistakes, because it was collected from the Web without validation, and 2) the corpus due to its relatively small size does not sufficiently represent the different inflected forms of terms.

It is also evident that ranking is a crucial component, because higher scores for more frequent inflected forms in almost all experiments allowed achieving higher results.

The experiments showed that the translation quality of an SMT system on in-domain data can be improved in a source text pre-processing scenario when using an automatically extracted bilingual term collection. The highest achieved score in the experiments was *15.34* BLEU points. However, an automatically extracted term collection requires that an in-domain parallel corpus (e.g., 2000 sentence pairs) is available for the extraction of the term collection. Obviously, this requirement cannot be always satisfied. Therefore, a more significant result is that by using a term collection created by a professional translator it is possible to achieve even better results; as explained earlier, the highest achieved score was 16.09 BLEU points.

In a summary, the results show that the best results (a maximal score of 16.09 BLEU points) were achieved with the combination of the *Professional* term collection, *Fast Term*

Identification, *Monolingual Corpus Look-up* of inflected forms from the *Broad Domain Corpus*, and the *Monolingual Corpus-based Term Translation Equivalent Ranking*.

According to the methodology described in section 4, when performing static terminology integration in SMT systems, terms from a term collection are injected into the monolingual corpus used for training of the SMT system's language model. However, when performing dynamic integration, translation equivalents may not be known to the language model. This means that the language model may not be able to reliably score correct and incorrect translation hypotheses. To address this issue, language models for SMT systems are usually trained on much larger corpora than used in the previously described experiments (up to tens of millions of sentences). To test whether dynamic terminology integration allows achieving translation quality improvements with a much larger language model, the author performed further experiments for the English-Latvian language pair using a new baseline system. The new baseline system is trained on the same parallel data as in the previous experiments, however for language model training, a corpus of 60.9 million unique Latvian sentences was used. The experiments were limited to the pre-processing scenario that achieved the highest results in the previous experiments. The results are given in Table 36.

Table 36. English-Latvian results of dynamic terminology integration in SMT systems using the professional term collection and a language model based on 60 million sentences

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	4.5534	14.42	0.1986	0.7536	4.4180	13.74	0.1552	0.7660	
Professional terms and broad domain monolingual corpus for surface form analysis									
Fast + Corpus + Simple + Exclusive	4.9967	16.79	**0.2237**	0.7181	4.7487	15.11	0.1730	0.7384	16.4%
Fast + Corpus + Simple + Inclusive	**5.0646**	**17.05**	**0.2237**	**0.7081**	**4.8357**	**15.54**	**0.1731**	**0.7263**	18.2%
Fast + Corpus + Simple + POS + Exclusive	4.9513	16.68	0.2220	0.7217	4.7040	15.00	0.1716	0.7420	15.7%
Fast + Corpus + Simple + POS + Inclusive	5.0087	16.91	0.2219	0.7116	4.7822	15.41	0.1718	0.7297	17.3%

It is evident from the results that the improvement is slightly smaller (in relative measures) than when using a smaller language model, however, the source text pre-processing workflow still allows achieving a significant translation quality improvement by up to 2.63 BLEU points over the new baseline system. The experiments also involved the remaining evaluation scenario dimensions – inclusive and exclusive decoding as well as POS tagging for the *Fast Term Identification* method. The results suggest that inclusive decoding for this evaluation scenario performs better. This means that either the term collection is ambiguous or the SMT system's translation model can generate in some contexts better translation hypotheses because of the limitations of the *Monolingual Corpus-Based Look-up* method for inflected form generation.

5.5.1.2. English-Lithuanian

Similarly to results reported for English-Latvian, the Table 37 shows automatic evaluation results for English-Lithuanian.

For English-Lithuanian the results show that the baseline system has a very low automatic evaluation result. Therefore, the evaluation of the baseline system was also performed using a *DGT-TM* based evaluation set (in-domain evaluation set for the SMT system). The results suggest that the automotive domain texts contain a significantly different language (in terms of writing style, terminology, etc.) than in the *DGT-TM* corpus. Although the baseline system shows such a relatively low score, the pre-processing experiments were still performed to identify by how much the low score can be improved.

Table 37. English-Lithuanian automatic evaluation results for dynamic terminology integration in SMT systems

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	*3.1750*	*6.94*	*0.1422*	*0.9068*	*2.9914*	*6.61*	*0.0904*	*0.9165*	
Baseline evaluated on DGT-TM data	*8.8771*	*48.12*	*0.3816*	*0.5202*	*8.6568*	*46.75*	*0.3222*	*0.5381*	
Professional terms and broad domain monolingual corpus for surface form analysis									
Fast + Corpus + Equal + Exclusive + POS	3.4579	7.75	0.1629	0.8834	3.2489	7.18	0.1097	0.8985	11.7%
Fast + Corpus + Equal + Inclusive + POS	3.3315	7.27	0.1520	0.8909	3.1513	6.74	0.0994	0.9032	4.8%
Fast + Corpus + Equal + Exclusive	3.4744	7.81	0.1637	0.8823	3.2626	7.17	0.1104	0.8975	12.5%
Fast + Corpus + Equal + Inclusive	3.3517	7.43	0.1527	0.8893	3.1701	6.91	0.1003	0.9014	7.1%
Fast + Corpus + Simple + Exclusive + POS	3.4883	7.92	0.1638	0.8810	3.2703	7.34	0.1103	0.8964	14.1%
Fast + Corpus + Simple + Inclusive + POS	3.3596	7.28	0.1529	0.8862	3.1746	6.76	0.1002	0.8989	4.9%
Fast + Corpus + Simple + Exclusive	**3.5090**	**7.99**	**0.1649**	**0.8798**	**3.2895**	**7.34**	**0.1114**	**0.8954**	15.1%
Fast + Corpus + Simple + Inclusive	3.3827	7.42	0.1537	0.8851	3.1954	6.90	0.1011	0.8976	6.9%
Pattern + Corpus + Equal + Exclusive + POS	3.3504	7.43	0.1567	0.8969	3.1579	6.96	0.1040	0.9095	7.1%
Pattern + Corpus + Equal + Inclusive + POS	3.2780	7.09	0.1494	0.8960	3.0902	6.55	0.0967	0.9089	2.2%
Pattern + Corpus + Simple + Exclusive + POS	3.3799	7.70	0.1579	0.8946	3.1786	7.22	0.1049	0.9080	11.0%
Pattern + Corpus + Simple + Inclusive + POS	3.3098	7.21	0.1505	0.8923	3.1206	6.67	0.0978	0.9050	3.9%
TWSC + Corpus + Equal + Exclusive + POS	3.2842	7.24	0.1494	0.9010	3.0989	6.79	0.0969	0.9113	4.3%
TWSC + Corpus + Equal + Inclusive + POS	3.2264	6.89	0.1457	0.9010	3.0337	6.38	0.0929	0.9116	-0.7%
TWSC + Corpus + Simple + Exclusive + POS	3.2904	7.31	0.1501	0.9018	3.1017	6.85	0.0976	0.9123	5.3%
TWSC + Corpus + Simple + Inclusive + POS	3.2425	7.01	0.1464	0.9004	3.0448	6.50	0.0928	0.9114	1.0%

The overall results suggest that for the English-Lithuanian pre-processing experiments the best results were achieved by the pre-processing scenario consisting of the *Fast Term Identification*, the *Monolingual Corpus-Based Look-up*, *Exclusive Decoding,* and *No POS-tagging Support*. The highest measured increase over the baseline system was 1.05 BLEU points (a relative improvement of 15.1%). Contrary to the results obtained for the English-Latvian experiments, for English-Lithuanian the *exclusive* decoding method allowed achieving higher results than the *inclusive* decoding. This indicates that either the term collection was not as ambiguous or the generated inflected forms of terms were in general better than the ones offered by the SMT system's translation model.

5.5.1.3. English-Estonian

The evaluation for English-Estonian was performed similarly to the English-Lithuanian evaluation. The results for English-Estonian are given in Table 38.

Table 38. English-Estonian automatic evaluation results for dynamic terminology integration in SMT systems

Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	3.1123	6.26	0.1399	0.9323	2.9347	5.90	0.0919	0.9419	
Baseline evaluated on DGT-TM data	8.8616	47.81	0.3722	-	8.5245	45.55	0.3115	-	
Professional terms and broad domain monolingual corpus for surface form analysis									
Fast + Corpus + Equal + POS + exclusive	3.2205	6.11	0.1428	0.9170	2.9702	5.67	0.0926	0.9296	-2.4%
Fast + Corpus + Equal + POS + inclusive	3.3324	6.47	0.1455	0.8950	3.0825	6.00	0.0947	0.9091	3.4%
Fast + Corpus + Equal + SKIP + exclusive	3.2469	6.27	0.1439	0.9149	2.9884	5.83	0.0935	0.9282	0.2%
Fast + Corpus + Equal + SKIP + inclusive	3.3554	**6.66**	**0.1465**	**0.8935**	**3.1032**	6.19	**0.0957**	**0.9079**	6.4%
Fast + Corpus + Simple + POS + exclusive	3.2224	6.10	0.1432	0.9176	2.9743	5.67	0.0927	0.9301	-2.6%
Fast + Corpus + Simple + POS + inclusive	3.3248	6.47	0.1453	0.8956	3.0796	6.01	0.0947	0.9093	3.4%
Fast + Corpus + Simple + SKIP + exclusive	3.2450	6.26	0.1441	0.9160	2.9887	5.82	0.0934	0.9292	0.0%
Fast + Corpus + Simple + SKIP + inclusive	**3.3478**	**6.66**	0.1463	0.8941	3.1003	**6.20**	0.0956	0.9080	6.4%
Pattern + Corpus + Equal + POS + exclusive	3.1799	6.08	0.1419	0.9262	2.9506	5.67	0.0923	0.9372	-2.9%
Pattern + Corpus + Equal + POS + inclusive	3.2573	6.40	0.1432	0.9082	3.0293	5.96	0.0930	0.9204	2.2%
Pattern + Corpus + Simple + POS + exclusive	3.1825	6.08	0.1423	0.9270	2.9561	5.67	0.0925	0.9380	-2.9%
Pattern + Corpus + Simple + POS + inclusive	3.2508	6.41	0.1430	0.9091	3.0260	5.97	0.0928	0.9209	2.4%
TWSC + Corpus + Equal + POS + exclusive	3.1758	6.34	0.1421	0.9249	2.9564	5.90	0.0921	0.9361	1.3%
TWSC + Corpus + Equal + POS + inclusive	3.1912	6.43	0.1418	0.9170	2.9830	5.98	0.0925	0.9281	2.7%
TWSC + Corpus + Simple + POS + exclusive	3.1738	6.33	0.1420	0.9253	2.9558	5.90	0.0919	0.9364	1.1%
TWSC + Corpus + Simple + POS + inclusive	3.1908	6.43	0.1416	0.9176	2.9841	5.99	0.0924	0.9285	2.7%

The results are quite different from the results obtained for English-Latvian and English-Lithuanian. The decoding with "*inclusive*" treatment of translation equivalents has shown to work better than the "*exclusive*" option. As the results do not correlate with the previous results, an in-depth analysis of what could the cause be for such results was performed. After analysing the term collection used for pre-processing it was identified that the Monolingual Corpus-Based Look-up method for inflected form generation for terms failed to produce translation equivalents for most of the terms. This was due to issues in the stemmer implemented for Estonian. As a result, most of the Estonian terms had only one translation equivalent and that was the term in its canonical form. The results show that also for Estonian it is not sufficient to provide for terms only their canonical forms as the translation equivalents. The translation equivalents have to be provided also in different inflected forms.

5.5.1.4. English-German

Similarly to the other language pairs, dynamic terminology integration experiments were performed for English-German. The results are given in Table 39.

Table 39. English-German automatic evaluation results for dynamic terminology integration in SMT systems

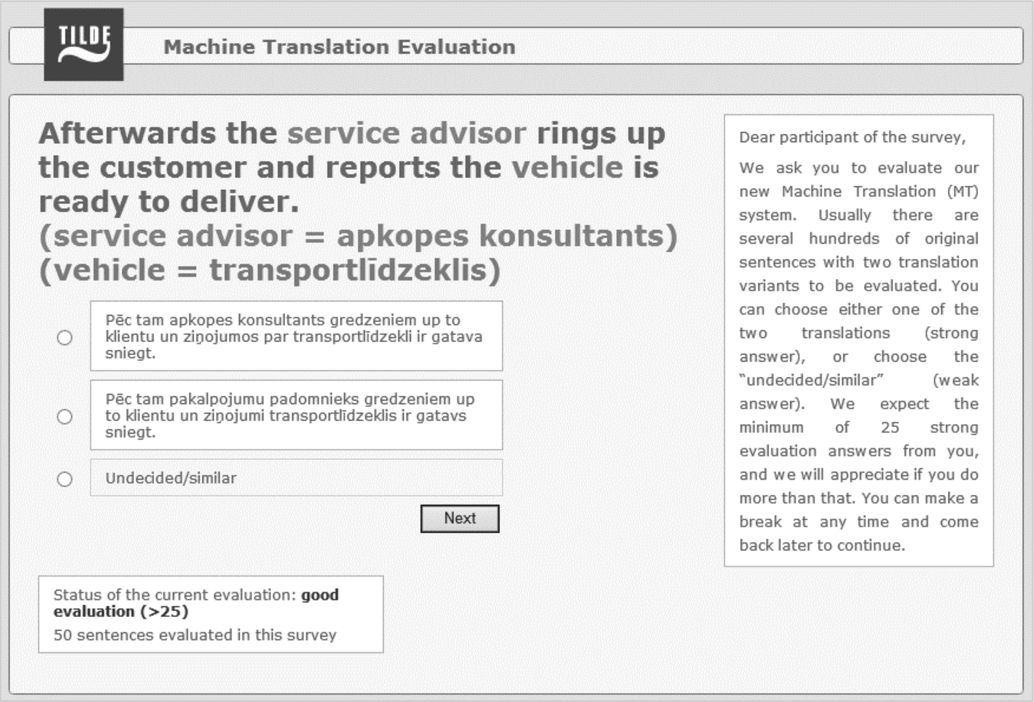
Pre-processing scenario	NIST	BLEU	METEOR	TER	NIST (C)	BLEU (C)	METEOR (C)	TER (C)	Change over baseline
Baseline	3.646	8.27	0.1639	0.8812	3.5030	7.94	0.1150	0.8898	
Baseline evaluated on DGT-TM data	9.4224	54.03	0.4055	0.4718	9.3024	53.10	0.3568	0.4821	
Professional terms and broad domain monolingual corpus for surface form analysis									
Fast + Corpus + Equal + POS + exclusive	4.2881	8.95	0.1891	0.8263	4.1491	8.56	0.1362	0.8357	8.2%
Fast + Corpus + Equal + POS + inclusive	4.1152	8.87	0.1811	0.8323	3.9714	8.56	0.1299	0.8427	7.3%
Fast + Corpus + Equal + SKIP + exclusive	4.3338	9.07	0.1907	**0.8214**	**4.1977**	8.68	**0.1381**	**0.8306**	9.7%
Fast + Corpus + Equal + SKIP + inclusive	4.1655	9.04	0.1829	0.8280	4.0260	8.73	0.1316	0.8374	9.3%
Fast + Corpus + Simple + POS + exclusive	4.2821	9.01	0.1889	0.8284	4.1387	8.62	0.1355	0.8382	8.9%
Fast + Corpus + Simple + POS + inclusive	4.1305	9.16	0.1817	0.8309	3.9857	8.82	0.1301	0.8413	10.8%
Fast + Corpus + Simple + SKIP + exclusive	**4.3384**	9.17	**0.1911**	0.8228	4.1944	8.76	0.1378	0.8319	10.9%
Fast + Corpus + Simple + SKIP + inclusive	4.1958	**9.40**	0.1842	0.8258	4.0572	**9.06**	0.1326	0.8351	13.7%
Pattern + Corpus + Equal + POS + exclusive	4.1996	9.07	0.1854	0.8379	4.0670	8.72	0.1335	0.8467	9.7%
Pattern + Corpus + Equal + POS + inclusive	4.0399	8.96	0.1780	0.8433	3.8996	8.63	0.1273	0.8534	8.3%
Pattern + Corpus + Simple + POS + exclusive	4.1901	9.15	0.1851	0.8397	4.0539	8.77	0.1328	0.8489	10.6%
Pattern + Corpus + Simple + POS + inclusive	4.0468	9.20	0.1781	0.8428	3.9067	8.84	0.1273	0.8528	11.2%
TWSC + Corpus + Equal + POS + exclusive	3.9200	8.73	0.1748	0.8635	3.7652	8.39	0.1227	0.8724	5.6%
TWSC + Corpus + Equal + POS + inclusive	3.8968	8.85	0.1729	0.8604	3.7501	8.53	0.1210	0.8694	7.0%
TWSC + Corpus + Simple + POS + exclusive	3.9293	8.72	0.1753	0.8639	3.7711	8.36	0.1227	0.8736	5.4%
TWSC + Corpus + Simple + POS + inclusive	3.8988	8.91	0.1729	0.8604	3.7514	8.56	0.1209	0.8693	7.7%

Also for English-German the evaluation results in comparison to the baseline system show a stable translation quality increase when performing source text pre-processing (although the baseline system's performance on the in-domain data is also relatively low). Different from the other language pairs, it is not evident whether inclusive or exclusive decoding is more beneficial as the results fluctuate. However, it is evident that for all language pairs one particular evaluation scenario (and pre-processing process chain) allows achieving the highest results: "*Fast + Corpus + Simple*".

5.5.2. Manual Evaluation Using Out-of-domain Systems

The goal of the manual evaluation is to allow human evaluators (instead of automatic means) to provide a natural view on whether the terminology integration has achieved its goal (that is, whether the SMT quality and terminology translation quality has improved) or not. In order to do so, for three language pairs (English-Latvian, English-Lithuanian, and English-Estonian) the baseline scenario (SMT without integrated terminology support) was compared with an improved scenario (with integrated terminology support). Taking into account the automatic evaluation results, for the improved scenario the author selected the source text pre-processing configuration consisting of the "*Fast + Corpus + Simple*" methods. For the baseline scenario, the same broad domain SMT systems (trained on only DGT-TM corpora) from the automatic evaluation experiments were used. For terminology integration, the same *professional* term collections in the automotive domain were used as in the automatic evaluation.

For comparative evaluation the Tilde’s web based evaluation environment (Skadiņš et al. 2010) was used. The system’s interface is shown in Figure 34.



Afterwards the service advisor rings up the customer and reports the vehicle is ready to deliver.
(service advisor = apkopes konsultants)
(vehicle = transportlīdzeklis)

○ Pēc tam apkopes konsultants gredzeniem up to klientu un ziņojumos par transportlīdzekli ir gatava sniegt.

○ Pēc tam pakalpojumu padomnieks gredzeniem up to klientu un ziņojumi transportlīdzeklis ir gatavs sniegt.

○ Undecided/similar

Status of the current evaluation: **good evaluation (>25)**
50 sentences evaluated in this survey

Dear participant of the survey,

We ask you to evaluate our new Machine Translation (MT) system. Usually there are several hundreds of original sentences with two translation variants to be evaluated. You can choose either one of the two translations (strong answer), or choose the “undecided/similar” (weak answer). We expect the minimum of 25 strong evaluation answers from you, and we will appreciate if you do more than that. You can make a break at any time and come back later to continue.

Figure 34. Tilde’s web based evaluation environment for the system comparison task

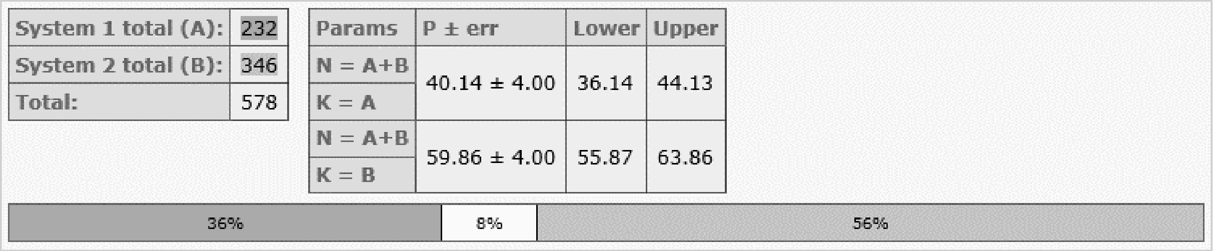
The figure shows that the evaluators (professional translators) were given a source segment where terms were marked in different colours. The term entries from the term collection were included below the source segment. This is an adaptation of the evaluation platform and its purpose is to inform evaluators about in-domain terminology in the source text. Then, below the source segment, two translations from the two different scenarios (the baseline scenario and the improved scenario) were shown. The system translation hypotheses were presented to the evaluators in a randomised order so that evaluators would not be able to identify the two systems.

Translators who took part in the manual evaluation efforts were asked to select the translation that they think is better taking into account that the aim was to achieve consistent and correct terminology translation with improved overall translation quality (or at least not decreased overall translation quality). If the translators could not decide, which translation hypothesis is better, they were asked to select the third option “*Undecided/similar*”. Evaluators were asked to evaluate at least 25 sentences. The statistics of the evaluated sentences and translators are given in Table 40.

Table 40. Evaluator and rating statistics

Language pair	Number of evaluators	Total number of ratings
English-Latvian	7	578
English-Lithuanian	6	534
English-Estonian	8	617

The evaluation methodology is based on the comparative evaluation methodology introduced by Skadiņš et al. (2010). The summary of the manual comparative evaluation for English-Latvian is presented in Figure 35, for English-Lithuanian – in Figure 36, for English-Estonian – in Figure 37. In the results, *System 1* is the baseline scenario and *System 2* is the improved scenario using the *professional* term collection.



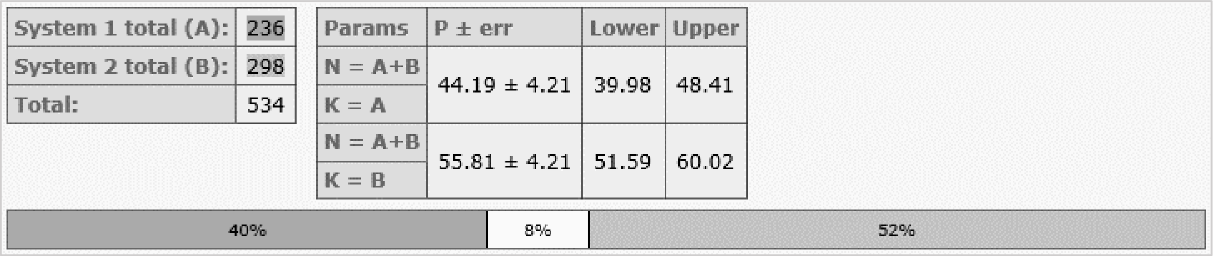
System 1 total (A):	232
System 2 total (B):	346
Total:	578

Params	P ± err	Lower	Upper

N = A+B
K = A
	40.14 ± 4.00	36.14	44.13

N = A+B
K = B
	59.86 ± 4.00	55.87	63.86

Figure 35. English-Latvian system comparison by total points



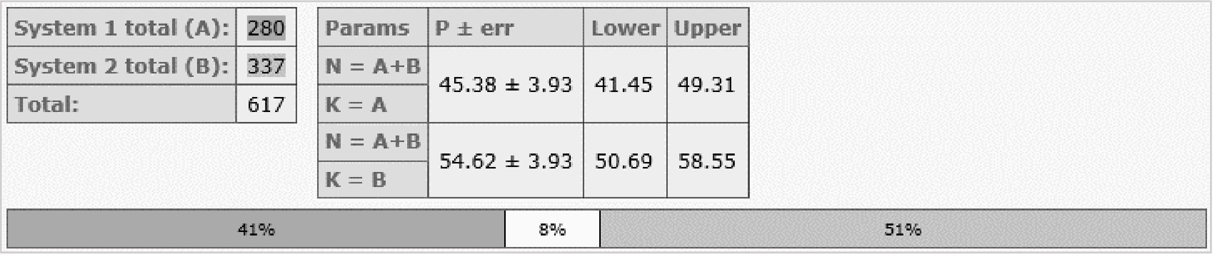
System 1 total (A):	236
System 2 total (B):	298
Total:	534

Params	P ± err	Lower	Upper

N = A+B
K = A
	44.19 ± 4.21	39.98	48.41

N = A+B
K = B
	55.81 ± 4.21	51.59	60.02

Figure 36. English-Lithuanian system comparison by total points



System 1 total (A):	280
System 2 total (B):	337
Total:	617

Params	P ± err	Lower	Upper

N = A+B
K = A
	45.38 ± 3.93	41.45	49.31

N = A+B
K = B
	54.62 ± 3.93	50.69	58.55

Figure 37. English-Estonian system comparison by total points

The results show that for all three language pairs it is weakly sufficient[26] to state that the translations of the improved scenario were preferred more than the translations of the baseline scenario. The translations of the improved scenario were preferred in 59.86±4.00% cases for

[26] According to the methodology by Skadiņš et al. (2010) it is **weakly sufficient** to say that *System 1* is preferred more than the *System 2* if the proportion of the preferences by total points for *System 1* minus the 95% confidence interval of the proportions of preferences by total points is greater than 50%. In this analysis the "*Undecided/similar*" ratings are added to both system preferences (thereby minimising the quality difference and penalising both systems).

English-Latvian, 55.81±4.21% cases for English-Lithuanian, and 54.62±3.93% cases for English-Estonian. Whereas the translations of the baseline scenario were preferred just in 40.14±4.00% cases for English-Latvian, 44.19±4.21% cases for English-Lithuanian, and 45.38±3.93% cases for English-Estonian. It has to be noted that indecisive answers have been counted for both scenarios.

Further analysis was performed by identifying sentences with sufficient confidence[27] (sentences rated as "*Undecided/similar*" were ignored). The analysis revealed that the translations of the improved scenario were preferred for 81.25±19.13% of sentences for English-Latvian, for 85.71±25.92% – for English-Lithuanian, and for 66.67±37.72% – for English-Estonian. For English-Latvian and English-Lithuanian we can conclude that the results are sufficient[28] to say that the dynamic terminology integration method allows creating translations of higher quality. For English-Estonian only six sentences were with sufficient confidence. Therefore, the confidence interval is too large and the results are just weakly sufficient to prove that the dynamic terminology integration method increases translation quality. However, the author believes that if the evaluation for English-Estonian would have been extended, a sufficient number of sentences with sufficient confidence would have been identified.

5.5.3. Manual Evaluation Using In-domain Systems

Although the previous evaluation results showed that the dynamic terminology integration method allows achieving significantly better results compared to the baseline scenario, the SMT systems in the baseline scenario achieved relatively low results and the term collections were relatively small (although focussed to a narrow domain). Therefore, an additional manual evaluation experiment was performed for seven language pairs using in-domain SMT systems (contrary to out-of-domain systems in the previous experiments) in the information technology domain that are used also by translators in their professional duties. For terminology integration, the author used the *Microsoft Terminology Collection*[29].

The term collection contains many ambiguous terms that can be confused with general language words and phrases (e.g., "*AND*", "*about*", "*name*", "*form*", "*order*", etc.). The combination of "*Fast + Corpus + Simple*" methods for source text pre-processing (contrary to

[27] Sentences with sufficient confidence according to the methodology by Skadiņš et al. (2010) are sentences for which the translations of one of the systems have been preferred by at least six evaluators more than the other system's translation.

[28] According to the methodology by Skadiņš et al. (2010) it is **sufficient** to say that *System 1* is preferred more than the *System 2* if the proportion of the preferences of statistically justified sentences for *System 1* minus the 95% confidence interval of the proportions of preferences of statistically justified sentences is greater than 50%. In this analysis the "*Undecided/similar*" ratings are ignored.

[29] The Microsoft Terminology Collection can be freely downloaded from: http://www.microsoft.com/Language.

the more linguistically motivated methods and methods that perform SMT system model adaptation) is sensitive to the level of ambiguity of the included terms. Therefore, it is important to filter out the ambiguous terms. For term filtering, the author used the term pair specificity estimation method (3) that was introduced in section 4.2.1.1. The statistics of the term collection before and after filtering are shown in Table 41.

Table 41. Term collection statistics before and after filtering (languages are given in the ISO 639-1 format)

Language pair	Terms (initial)	Terms (filtered)
en-es	23,094	18,871
en-fr	24,160	19,665
en-et	12,648	10,175
en-lt	12,726	10,352
en-lv	12,926	10,497
en-ru	22,669	18,416
en-de	24,997	20,308

For the evaluation, pre-trained SMT systems from the LetsMT platform were used. The SMT system performance on in-domain evaluation sets (however, different from the evaluation data used in the manual evaluation experiment) is given in Table 42.

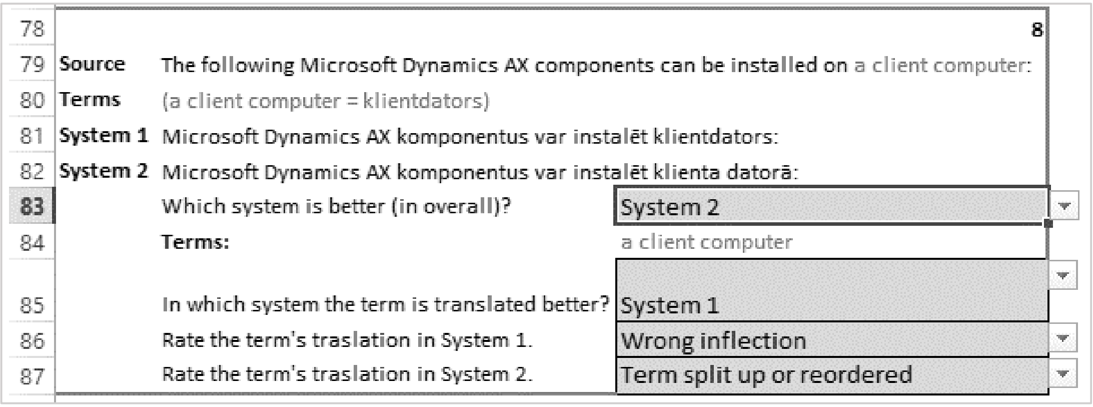
Table 42. SMT system performance on held-out evaluation sets; the systems were created by Valters Šics in the LetsMT platform (languages are given in the ISO 639-1 format)

Language pair	BLEU
en-es	74.61
en-fr	68.76
en-et	55.23
en-lt	60.42
en-lv	66.98
en-ru	60.79
en-de	61.35

The manual evaluation was performed by comparing the SMT system performance without (the baseline scenario) and with (the improved scenario) integrated terminology. The evaluation data for each language pair consists of 100 in-domain sentences for which the outputs of the SMT systems in the two scenarios differed (different translations were produced in average for 56% of sentences). For each language pair two professional translators were involved in the evaluation. The translators were asked to perform three ratings using an Excel spreadsheet (an example of the evaluation task is given in Figure 38):

- For each sentence, translators had to decide which scenario produced a better translation. If both scenarios produced translations of equal quality, the translators had to decide whether both scenarios produced acceptable or not acceptable translations.
- Similarly to the sentence level, for each term that was identified in the source text using the *Fast Term Identification* method, translators had to decide which scenario produced a better translation.

- The first two are quantitative analysis measures, therefore as a third rating translators were asked to rate the term translation quality in both scenarios separately. The translators had to decide whether:
  - the term is translated correctly,
  - a wrong inflectional form is used,
  - the term is left untranslated,
  - the term is split up or its words are in a wrong order,
  - a wrong lexical choice is made,
  - the marked phrase is actually not a term and has been wrongly identified as a term,
  - the term is not translated correctly, but there is a different issue.



8

Source The following Microsoft Dynamics AX components can be installed on a client computer:

Terms (a client computer = klientdators)

System 1 Microsoft Dynamics AX komponentus var instalēt klientdators:

System 2 Microsoft Dynamics AX komponentus var instalēt klienta datorā:

Which system is better (in overall)? System 2

Terms: a client computer

In which system the term is translated better? System 1

Rate the term's traslation in System 1. Wrong inflection

Rate the term's traslation in System 2. Term split up or reordered

Figure 38. An example of the evaluation task for English Latvian showing a sentence containing one term from the filtered term collection

The sentence level evaluation summary in Table 43 shows that the translations of the improved scenario were preferred more than the baseline scenario for six language pairs (results are sufficient to say that the improved scenario produces better quality translations for English-Latvian, English-German, and English-French; see Figure 39). It is evident that the task of comparing sentence level quality is very challenging for evaluators, because the agreement scores in terms of the *Free-marginal Kappa* (Randolph, 2005) are mainly in the levels of fair to moderate.

Table 43. Evaluation summary for sentence level ratings where evaluators were in agreement (languages are given in the ISO 639-1 format)

Language pair	Baseline	Improved	Both	None	Total	Free Kappa
en-es	11	8	15	19	53 / 99	0.38
en-fr	8	21	35	18	82 / 221	0.16
en-et	8	16	3	36	63 / 101	0.50
en-lt	6	8	23	16	53 / 100	0.37
en-lv	1	9	9	57	76 / 100	0.68
en-ru	9	17	7	27	60 / 100	0.47
en-de	5	15	29	9	58 / 99	0.45

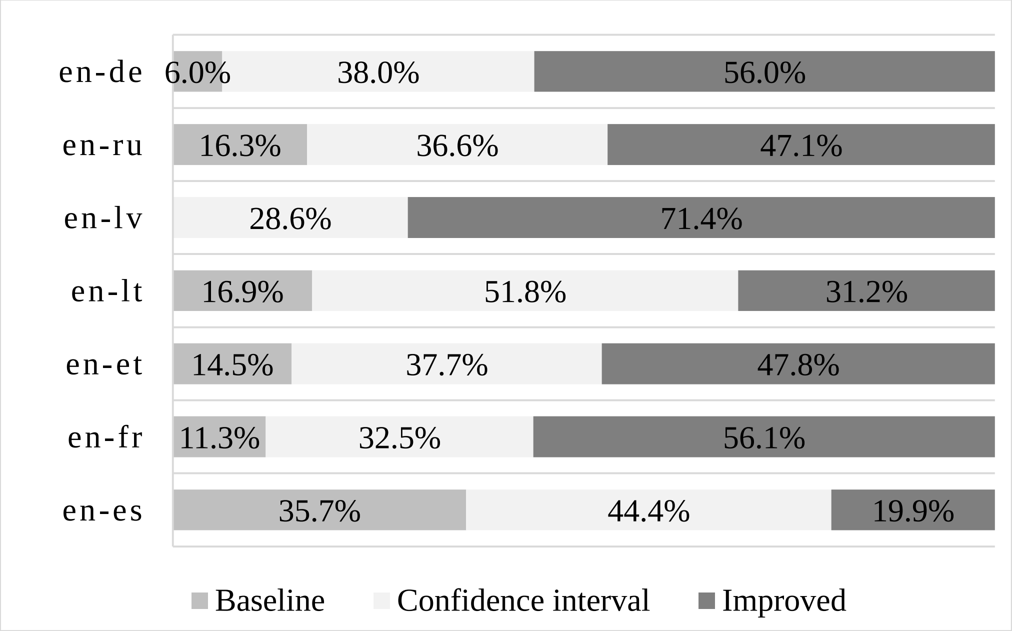


Figure 39. Confidence intervals for sentence level summary of ratings for sentences where both evaluators were in agreement (languages are given in the ISO 639-1 format)

The term level evaluation summary is given in Table 44. It is evident that translation quality has improved over the baseline scenario for all language pairs that were evaluated. The results are sufficient (see Figure 40) for all seven language pairs to state that the dynamic terminology integration method allows producing better quality translations for terms than the baseline scenario. Even more, the agreement scores for evaluators show that the task of comparing in which system terms were translated better was fairly easy and in general well understood.

Table 44. Evaluation summary for term level ratings where evaluators were in agreement (languages are given in the ISO 639-1 format)

Language pair	Baseline	Improved	Both	None	Total	Free Kappa
en-es	4	34	77	0	115 / 157	0.64
en-fr	4	71	141	4	220 / 380	0.44
en-et	21	51	53	0	125 / 162	0.70
en-lt	1	40	54	3	98 / 158	0.49
en-lv	6	46	67	4	123 / 151	0.75
en-ru	1	49	93	0	143 / 166	0.82
en-de	2	30	87	0	119 / 153	0.70

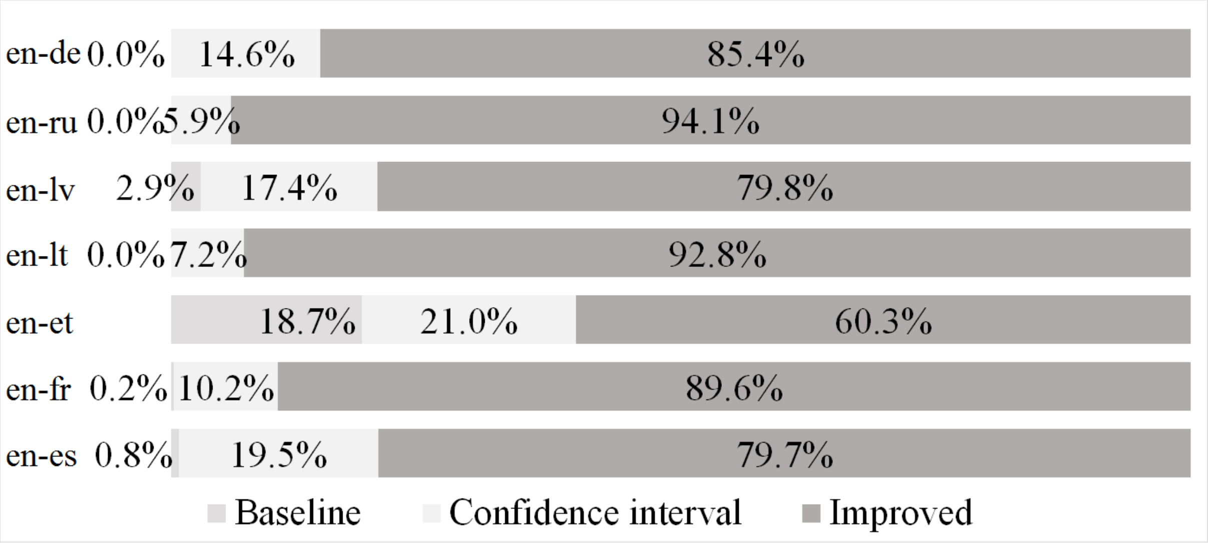


Figure 40. Confidence intervals for term level summary of ratings for terms where both evaluators were in agreement (languages are given in the ISO 639-1 format)

The summary of the term translation quality evaluation for the individual scenarios is given in Table 45. The results show that the proportion of correct term translations has improved for all language pairs from +1.6% for English-Estonian to +52.6% for English-Lithuanian. The minimal improvement for English-Estonian is mainly due to selection of wrong inflected forms (which is a lesser quality issue, but an issue nonetheless) rather than wrong term lexical choices (which is a greater quality issue). The author believes that the relatively low performance for English-Estonian is caused by the under-performance of the word stemming component for Estonian that is used for inflectional form acquisition for terms. It is evident that in terms of using the correct lexical choice, the quality has improved from +26.4% for English-German to +65.2% for English-Lithuanian. This means that the method allows ensuring terminology translation consistency better than in the baseline scenario. If we analyse further the reduction of term translation mistakes, the English-Russian system achieved the best results with an error reduction of 72.7%.

The results show that for morphologically richer and less resourced languages (e.g., Latvian, Lithuanian) the proportion of correct term translations in the baseline scenario is lower than for morphologically less rich and well-resourced languages (Spanish, German). This is because of two reasons: 1) the amount of training data for SMT system development differs, and 2) for morphologically rich languages it is more challenging for the SMT system to select the correct inflected form of terms (which is shown by the higher proportion of wrong inflected forms selected in translations). For morphologically rich languages the improvement of correct selection of term lexical forms is approximately 30%, which is higher by 10% than for morphologically simpler languages. However, the results show that the improvement after dynamic terminology integration is over 14% for all language pairs (except for English-Estonian due to the reasons explained earlier).

Table 45. Evaluation summary for term translation quality
(languages are given in the ISO 639-1 format; "B" is the baseline scenario and "I" is the improved scenario)

Percentage of terms	**en-es**		**en-fr**		**en-et**		**en-lt**		**en-lv**		**en-ru**		**en-de**	
	B	**I**	**B**	**I**	**B**	**I**	**B**	**I**	**B**	**I**	**B**	**I**	**B**	**I**
Term correct	71.3	85.4	55.9	75	39.8	40.4	42.1	64.2	51.3	67.9	60.2	89.2	70.3	85.6
Wrong inflection	1.9	11.5	1.6	5.8	19.4	50.3	7.9	18.4	11.3	27.5	6	8.7	1.6	5.2
Not translated	8.6	0.6	19.2	14.3	9.9	1.2	0.6	0	4.6	0	16.3	0.9	7.8	0.3
Term split up or reordered	2.2	0.3	6.7	0.9	2.2	0.3	1.9	2.2	2.6	0	6.6	0.6	0.7	1
Wrong lexical choice	7.3	1.3	13.3	1.6	18.8	4.6	30.4	2.8	20.9	0	10.8	0.6	5.9	5.6
Not a term	6.4	0.6	1.7	1.7	0.6	0.6	10.8	10.8	1.7	1.7	0	0	0.7	1
Other	2.2	0.3	1.6	0.7	9.3	2.5	6.3	1.6	7.6	3	0	0	13.1	1.3
Rel. impr. of correct term translations	19.60%		34.10%		1.60%		52.60%		32.30%		48.00%		21.90%	
Rel. impr. of correct lexical choice	32.20%		40.50%		53.10%		65.20%		52.40%		47.70%		26.40%	
Rel. red. of errors	48.90%		43.30%		1.00%		38.30%		34.00%		72.70%		51.60%	

5.6. Summary of Dynamic Integration of Terminology in SMT Systems

In this section, the author presented a novel workflow for dynamic terminology integration in SMT systems using source text pre-processing methods. The workflow consists of four steps: 1) term identification in the source text, 2) inflected form generation for terms, 3) term translation equivalent ranking for translation, and 4) the translation of the pre-processed source text with an SMT system. The methods have been evaluated in three evaluation experiments: 1) automatic evaluation for four language pairs using standard SMT evaluation metrics, 2) manual comparative system evaluation for three language pairs using broad domain SMT systems, and 3) manual comparative system evaluation and term translation quality evaluation for seven language pairs using production level SMT systems in the information technology domain.

For term identification three methods were analysed: 1) the linguistically and statistically motivated term identification using *TWSC*, 2) the *Pattern-Based Term Identification*, and 3) the *Fast Term Identification*. The results show that the *Fast Term Identification* method outperforms the linguistically motivated term identification methods ("*Pattern*" and "*TWSC*"). This may be explained with the fact that recall of the *Fast Term Identification* method is higher than the recall of the *Pattern-Based Term Identification* method and much higher than that of the *TWSC-based Term Identification* method. This is a very positive result, because it is possible to achieve the highest performance (in terms of speed) and still maintain the best translation quality. However, this also means that the linguistically motivated methods (1 and 2), which rely on linguistic resources that are exhaustive, may need to be improved. For instance, the term patterns may not contain all patterns necessary for a given term collection or the morpho-syntactic tagger may not know how to tag an unknown word of a term phrase. However, this is a possible area for future improvements.

For inflected form generation four methods were analysed: 1) no inflected form generation, 2) *Rule-based Morphological Synthesis* of inflected forms for terms, 3) *Monolingual Corpus Look-up* of inflected forms, and 4) the combination (through union) of the second and third methods. The evaluation results have shown that the highest results can be achieved with the *Monolingual Corpus Look-up* method. The *Rule-based Morphological Synthesis* method and the combined method performed worse than the remaining methods, because of high ambiguity introduced to the SMT system's decoder.

The experiments have shown that by generating different inflected forms of terms and preferring higher scores for more frequent inflected forms (with the frequency-based ranking method) it is possible to achieve a higher SMT quality than with equal ranking.

The results also showed that using just the translation equivalents from a term collection (without further inflected form generation for terms) allows achieving the highest results when an automatically created bilingual term collection is used. Although the experiment results show stable translation quality improvements, this scenario does not allow achieving the highest results when using a *professional* term collection. This proves the hypothesis that for morphologically rich languages modelling of correct inflected forms is very important in order to achieve as good results as possible.

CONCLUSIONS

The author in this thesis presented novel methods for terminology integration in statistical machine translation in both SMT systems during training (through static integration) and during translation (through dynamic integration). The work focussed not only on the SMT integration techniques, but also on methods for acquisition of linguistic resources (including the bilingual term collections) necessary for different tasks involved in the workflows for terminology integration in SMT systems. For instance, monolingual term identification, term normalisation for acquisition of canonical forms of terms from terms in different inflected forms, and cross-lingual term mapping for semi-automated creation of bilingual term collections. To increase performance of the cross-lingual term mapping methods, the author presented novel methods for probabilistic dictionary filtering and character-based SMT transliteration system development using probabilistic dictionaries. For static and dynamic terminology integration in SMT systems, the author designed and implemented methods that allow performing bilingual term identification in SMT training data and the source text (for dynamic integration), inflected form generation for terms using rule-based morphological synthesis or monolingual corpus look-up methods, etc.

The terminology integration methods were specifically designed for the *Moses* SMT system and the *LetsMT* platform (that uses the *Moses* SMT system), however the methods designed are fairly general and can be applied for any phrase-based SMT system that operates similarly to the *Moses* system.

The methods presented in the thesis have been evaluated using both automated evaluation methods as well as manual evaluation methods. The monolingual term identification method using *TWSC* for term collection creation purposes and cross-lingual term mapping method using *MPAligner* have shown to achieve state-of-the-art performance, which has been also validated by third party (independent) evaluation efforts. The term mapping quality of *MPAligner* (including also the evaluation of methods for linguistic resource creation for *MPAligner*) and the monolingual term identification method using *TWSC* have been evaluated for all official languages of the European Union by the author and also by third party researchers, thus showing the language independence of the methods designed by the author.

For static terminology integration in SMT systems, the evaluation shows that the designed methods for English-Latvian on the automotive domain evaluation data allowed to achieve a cumulative SMT quality improvement of up to 28.1% (or 3.56 absolute BLEU points) over an initial baseline system. The translation model adaptation method that introduces a bilingual terminology identifying feature in the SMT system's translation model has shown to be stable in increasing SMT quality.

However, the most significant achievement of the author's work is the dynamic terminology integration method in SMT systems using the source text pre-processing workflow. In almost all experiments, different combinations of the pre-processing workflow showed SMT quality improvements. The dynamic terminology integration method was also evaluated in three different evaluation experiments. Automatic evaluation in the automotive domain was performed for four different language pairs (English-Latvian, English-Lithuanian, English-Estonian, and English-German). It showed SMT quality improvements for all language pairs ranging from 6.4% (or 0.40 absolute BLEU points) for English-Estonian up to 26.9% (or 3.41 absolute BLEU points) for English-Latvian in terms of BLEU points over the results of the baseline systems. Manual comparative system evaluation for three language pairs (English-Latvian, English-Lithuanian, and English-Estonian) in the automotive domain further validated that the dynamic terminology integration methods allow improving SMT system quality using bilingual term collections. Furthermore, manual comparative evaluation in the information technology domain using production level SMT systems for seven language pairs showed that the proportion of correct term translations has improved for all language pairs from +1.6% for English-Estonian to +52.6% for English-Lithuanian. The methods allow reducing the proportion of mistranslated terms from +1.0% for English-Estonian to an impressive +72.7% for English-Russian.

The positive evaluation results of both the static terminology integration experiments and the dynamic terminology integration experiments allow the author to conclude that the research hypothesis that terminology translation quality as well as text translation quality in SMT systems can be improved by performing static and dynamic terminology integration in SMT systems has been successfully proven. The results of the static and dynamic terminology integration experiments also prove the second hypothesis that in situations when authoritative term collections are not available, automatic term identification in comparable corpora and cross-lingual term mapping are effective methods to acquire bilingual term collections for the integration in SMT systems. The goal of the thesis has been reached and all objectives have been completed.

The thesis also drafted possible areas of future improvements for the methods designed by the author. The possible areas are as follows:

- *TWSC* requires a property file that specifies different thresholds. These thresholds are sensitive to document length. In future work algorithms could be improved so that the performance of *TWSC* is less affected by document length variations.

- *TWSC* and the *Pattern-Based Term Identification* methods both rely on term patterns. The patterns are exhaustive resources and may not be defined for all terms given in term collections. A method for dynamic acquisition of term patterns for terms could be investigated.
- Currently, each term pattern is required to have exactly one normalisation rule. However, the normalisation rules could be dynamically predicted. For instance, for multi-word terms in Latvian, nouns before the head noun are usually kept in their respective inflected forms, however adjectives are inflected corresponding to the head noun they are attached to. This behaviour could be captured using dynamic rules that are not fixed to a single term pattern, thus eliminating the need for the definition of one-to-one rules.
- The term mapping tool *MPAligner* has many parameters that may need to be adjusted for different language pairs. Machine learning methods could be investigated to fine-tune the system's parameters in order to achieve higher recall and precision.
- The evaluation of the probabilistic dictionary filtering methods revealed that better recall dictionaries could be acquired by combining the different filtering methods.
- Currently, the context independent term mapping method does not use any reference corpus based statistics, however, using statistics that are calculated on very large (e.g., hundred million words or more) monolingual corpora could be beneficial by minimising misalignments similarly as it is implemented in the probabilistic dictionary filtering method.
- Regarding transliteration systems, the evaluation has indicated that their application in SMT may not be productive due to a limited precision (around 50%) for the top one transliteration equivalents. However, if language specific knowledge (in terms of morphological analysis) would be introduced, the SMT-based transliteration systems could be trained to transliterate words into specific inflected forms. This would allow providing SMT systems with linguistically motivated transliteration equivalents and could potentially provide a method for translation of out-of-vocabulary words.
- Regarding static terminology integration, to ensure that the tuning process considers the term identifying feature (during translation model adaptation) productive, the SMT system training workflow could be modified to validate whether the tuning data is rich with in-domain terminology or not. If the tuning data does not contain in-domain terminology, a new process could be integrated that selects additional (or even new) sentence pairs from the parallel corpus so that the selected tuning data would be rich in in-domain terminology.

- Regarding dynamic terminology integration, the experiments relied only on the target language data in order to rank the inflected forms of terms. By doing so the methods do not take into account the linguistic information transfer from the source language to the target language, which is important to guess the necessary inflected forms of terms in the translated text. Machine learning methods could be investigated that allow ranking the inflected forms of terms according to the source text.

REFERENCES

ACCURAT. (2011). *D2.3 Report on Information Extraction from Comparable Corpora* (p. 55). ACCURAT project: Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation.

Aker, A., Paramita, M., & Gaizauskas, R. (2013). Extracting Bilingual Terminologies from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 402–411). Sofia, Bulgaria: Association for Computational Linguistics.

Aker, A., Paramita, M. L., Barker, E., & Gaizauskas, R. (2014a). Bootstrapping Term Extractors for Multiple Languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)* (pp. 483–489). Reykjavik, Iceland: European Language Resources Association (ELRA).

Aker, A., Pinnis, M., Paramita, M. L., & Gaizauskas, R. (2014b). Bilingual Dictionaries for All EU Languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)* (pp. 2839–2845). Reykjavik, Iceland: European Language Resources Association (ELRA).

Arbabi, M., Fischthal, S. M., Cheng, V. C., & Bart, E. (1994). Algorithms for Arabic Name Transliteration. *IBM Journal of Research and Development*, *38*(2), 183–194.

Arcan, M., Giuliano, C., Turchi, M., & Buitelaar, P. (2014a). Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*.

Arcan, M., Turchi, M., Tonelli, S., & Buitelaar, P. (2014b). Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)* (pp. 54–68).

Babych, B., & Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*, 1313–1316.

Bertoldi, N. (2014). Dynamic Models in Moses for Online Adaptation. *The Prague Bulletin of Mathematical Linguistics*, (101), 7–28.

Bertoldi, N., & Federico, M. (2009). Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 182–189). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bertoldi, N., Haddow, B., & Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. *The Prague Bulletin of Mathematical Linguistics*, *91*(1), 7–16.

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 674–679).

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the Biennial GSCL Conference* (pp. 31–40).

Bourigault, D. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3* (pp. 977–981). Association for Computational Linguistics.

Brown, P. E., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2).

Callison-Burch, C., Bannard, C., & Schroeder, J. (2005). Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 255–262). Association for Computational Linguistics.

Carl, M., & Langlais, P. (2002). An Intelligent Terminology Database as a Pre-processor for Statistical Machine Translation. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14* (pp. 1–7).

Chen, Y., & Eisele, A. (2010). Integrating a Rule-based with a Hierarchical Translation System. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 1746–1752). Valletta, Malta.

Chinchor, N. (1997). MUC-7 Named Entity Task Definition. In *Proceedings of the 7th Conference on Message Understanding*.

Church, K., Gale, W., Hanks, P., & Kindle, D. (1991). Using Statistics in Lexical Analysis. In *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon* (pp. 115–164).

Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, *16*(1), 22–29.

Dagan, I., & Church, K. (1994). Termight : Identifying and Translating Technical Terminology. In *Proceedings of the fourth conference on Applied Natural Language Processing* (pp. 34–40). Association for Computational Linguistics.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of the Workshop the Balancing Act: Combining Symbolic and Statistical Approaches to Language (Language, Speech, and Communication)* (pp. 29–36). Las Cruces, New Mexico, USA: Association for Computational Linguistics.

De Groc, C. (2011). Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2011)* (Vol. 1, pp. 497–498).

Deksne, D. (2013). Finite State Morphology Tool for Latvian. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing* (pp. 49–53).

Delač, D., Krleža, Z., Šnajder, J., Bašić, B. D., & Šarić, F. (2009). TermeX: A Tool for Collocation Extraction. In *Computational Linguistics and Intelligent Text Processing* (pp. 149–157). Springer.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 85–91).

Dias, G., Guilloré, S., Bassano, J., Gabriel, J., & Lopes, P. (2000). Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? In *Recherche d'Informations Assistée par Ordinateur (RIAO 2000)*. Paris, France.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association between Species. *Ecology*, *26*(3), 297–302.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the second international conference on*

Human Language Technology Research (pp. 138–145). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)* (pp. 644–648). Atlanta, USA.

Easterbrook, S., Singer, J., Storey, M.-A., & Damian, D. (2008). Selecting empirical methods for software engineering research. In *Guide to advanced empirical software engineering* (pp. 285–311). Springer.

Eisele, A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)* (pp. 2868–2872). Valletta, Malta.

Esplá-Gomis, M. (2009). Bitextor, a Free/Open-Source Software to Harvest Translation Memories from Multilingual Websites. In *Proceedings of MT Summit XII: the Twelfth Machine Translation Summit*. Ottawa, Canada: Association for Machine Translation in the Americas.

Federmann, C., Gromann, D., Declerck, T., Hunsicker, S., Krieger, H., & Budin, G. (2012). Multilingual Terminology Acquisition for Ontology-Based Information Extraction. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)* (pp. 166–175). Madrid, Spain.

Finch, A., & Sumita, E. (2008). Phrase-Based Machine Transliteration. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)* (pp. 13–18). Hyderabad, India: Asian Federation of Natural Language Processing.

Flournoy, R., & Duran, C. (2009). Machine translation and document localization at Adobe: From pilot to production. M*T Summit XII: Proceedings of the Twelfth Machine Translation Summit*, 425–428.

Foo, J. (2012). *Computational Terminology : Exploring Bilingual and Monolingual Term Extraction*. Linköping University.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, *3*(2), 115–130.

Frantzi, K. T., & Ananiadou, S. (1997). Automatic Term Recognition Using Contextual Cues. In *In Proceedings of 3rd DELOS Workshop*.

Fung, P., & Yee, L. Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1* (pp. 414–420). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gaussier, E., Hull, D. A., Salah, A., & Ait-Mokhtar, S. (2000). Term Alignment in Use: Machine-Aided Human Translation. *Véronis, Jean: Parallel Text Processing. Alignment and Use of Translation Corpora. Dordrecht*, 253–274.

Grigonyte, G., Rimkute, E., Utka, A., & Boizou, L. (2011). Experiments on Lithuanian Term Extraction. In *Proceedings of the NODALIDA 2011 Conference* (pp. 82–89).

Hálek, O., Rosa, R., Tamchyna, A., & Bojar, O. (2011). Named Entities from Wikipedia for Machine Translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies (ITAT 2011)* (pp. 23–30).

Hildebrand, A. S., Eck, M., Vogel, S., & Waibel, A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of EAMT* (Vol. 2005, pp. 133–142).

Hong, M., Fissaha, S., & Haller, J. (2001). Hybrid Filtering for Extraction of Term Candidates from German Technical Texts. In *Terminologie et intelligence artificielle (TIA 2001)* (pp. 223–232).

Jacquemin, C., Klavans, J. L., & Tzoukermann, E. (1997). Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 24–31). Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2nd ed.). Pearson Education International.

Justeson, J. S., & Katz, S. M. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering, 1*(01), 9–27.

Kirschenbaum, A., & Wintner, S. (2010). A General Method for Creating a Bilingual Transliteration Dictionary. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC-2010)* (pp. 273–276).

Knight, K., & Graehl, J. (1997). Machine Transliteration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 128–135).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48–54). Association for Computational Linguistics.

Koehn, P., & Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 224–227). Prague, Czech Republic.

Koster, J. (1975). Dutch as an SOV Language. *Linguistic Analysis*, 1(2), 111–136.

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

Lardilleux, A., Yvon, F., & Lepage, Y. (2012). Hierarchical Sub-Sentential Alignment with Anymalign. In *Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012)* (pp. 279–286).

Laroche, A., & Langlais, P. (2010). Revisiting Context-Based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 617–625). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lefever, E., Macken, L., & Hoste, V. (2009). Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus. In *Proceedings of the 12th Conference of* *the European Chapter of the Association for Computational Linguistics* (pp. 496–504). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lehmann, W. P. (1978). English: A Characteristic SVO Language. In *Syntactic typology: Studies in the phenomenology of language* (pp. 169–222). University of Texas Press.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lewis, W. D., Wendt, C., & Bullock, D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 2878–2883).

Lokmane, I. (2010). Vārdu secības funkcijas latviešu valodā. In *Proceedings "Latvistika un somugristika Latvijas Universitātē"* (pp. 59–68). Riga, Latvia: University of Latvia.

Mastropavlos, N., & Papavassiliou, V. (2011). Automatic Acquisition of Bilingual Language Resources. In *Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece*.

Mateo, R. M. (2014). A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea: A Journal of English and American Studies*, *49*(2014), 73–94.

Merkel, M., & Foo, J. (2007). Terminology Extraction and Term Ranking for Standardizing Term. In *Proceedings of NODALIDA 2007* (pp. 349–354).

Morin, E., & Daille, B. (2010). Compositionality and Lexical Alignment of Multi-Word Terms. *Language Resources and Evaluation*, *44*(1-2), 79–95.

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2010). Brains, not Brawn: The Use of "Smart" Comparable Corpora in Bilingual Terminology Mining. *ACM Transactions on Speech and Language Processing (TSLP)*, *7*(1), 1.

Munteanu, D., & Marcu, D. (2006). Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (pp. 81–88). Sydney, Australia: Association for Computational Linguistics.

Nikoulina, V., Sandor, A., & Dymetman, M. (2012). Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation. In *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (ML4HMT-12)* (pp. 1–16). Mumbai, India.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29*(1), 19–51.

Okuma, H., Yamamoto, H., & Sumita, E. (2008). Introducing a Translation Dictionary into Phrase-Based SMT. *IEICE Transactions on Information and Systems, 91*(7), 2051–2057.

Pantel, P., & Lin, D. (2001). A Statistical Corpus-Based Term Extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence - Advances in Artificial Intelligence (AI 2001)* (pp. 36–46). Ottawa, Canada: Springer-Verlag Berlin Heidelberg.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology Extraction: an Analysis of Linguistic and Statistical Approaches. In *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference* (pp. 255–279). Springer.

Pecina, P. (2005). An Extensive Empirical Study of Collocation Extraction Methods. In *Proceedings of the ACL Student Research Workshop* (pp. 13–18). Association for Computational Linguistics.

Pecina, P., & Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 651–658). Association for Computational Linguistics.

Petrović, S., Šnajder, J., & Bašić, B. D. (2010). Extending Lexical Association Measures for Collocation Extraction. *Computer Speech & Language, 24*(2), 383–394.

Pinnis, M. (2012). Latvian and Lithuanian Named Entity Recognition with TildeNER. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1258–1265). Istanbul, Turkey: European Language Resources Association (ELRA).

Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)* (pp. 562–570). Hissar, Bulgaria.

Pinnis, M. (2014). Bootstrapping of a Multilingual Transliteration Dictionary for European Languages. In *Human Language Technologies – The Baltic Perspective - Proceedings of* *the Sixth International Conference Baltic HLT 2014* (pp. 132–140). Kaunas, Lithuania: IOS Press.

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).* Antalya, Turkey: European Association for Machine Translation.

Pinnis, M., & Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In C. Mahlow & M. Piotrowski (Eds.), *Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology* (pp. 14–22). Zurich, Switzerland: Springer Berlin Heidelberg.

Pinnis, M., Gornostay, T., Skadiņš, R., & Vasiļjevs, A. (2013). Online Platform for Extracting, Managing, and Utilising Multilingual Terminology. In *Proceedings of the Third Biennial Conference on Electronic Lexicography, eLex 2013* (pp. 122–131). Tallinn, Estonia: Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia) / Eesti Keele Instituut (Tallinn, Estonia).

Pinnis, M., Ljubešić, N., Ştefãnescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)* (pp. 193–208). Madrid.

Pinnis, M., & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In A. Tavast, K. Muischnek, & M. Koit (Eds.), *Human Language Technologies – The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012* (Vol. 247, pp. 176–184). Tartu, Estonia, Estonia: IOS Press.

Pouliquen, B., Steinberger, R., Ignat, C., Temnikova, I., Widiger, A., Zaghouani, W., & Zizka, J. (2005). Multilingual Person Name Recognition and Transliteration. *Journal CORELA - Cognition, Representation, Langage. Numéros Spéciaux, Le Traitement Lexicographique Des Noms Propres*.

Randolph, J. J. (2005). Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Joensuu Learning and Instruction Symposium*.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 519–526). Stroudsburg, PA, USA: Association for Computational Linguistics.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.

Schmidtke, D. (2008). Microsoft office localization: use of language and translation technology. URL Http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf.

Shao, L., & Ng, H. T. (2004). Mining New Word Translations from Comparable Corpora. In *Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, 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 N. C. C. Calzolari, 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)* (pp. 438–445). Istanbul, Turkey: European Language Resources Association (ELRA).

Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. In *Human Language Technologies: The Baltic Perspective: Proceedings of the Fourth International Conference, Baltic HLT 2010* (Vol. 219, pp. 125–132). Riga, Latvia: IOS Press.

Skadiņš, R., Pinnis, M., Gornostay, T., & Vasiļjevs, A. (2013). Application of Online Terminology Services in Statistical Machine Translation. In *Proceedings of the XIV Machine Translation Summit* (pp. 281–286). Nice, France: The European Association for Machine Translation.

Skadiņš, R., Skadiņa, I., Pinnis, M., Vasiļjevs, A., & Hudík, T. (2014). Application of Machine Translation in Localization into Low-resourced Languages. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 2014)* (pp. 209–216).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas* (pp. 223–231). Cambridge, MA, USA.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation, 28,* 11–21.

Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In *The 5th Workshop on Building and Using Comparable Corpora* (pp. 98–103). Turkey, Istanbul.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlter, P. (2012). DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 454–459).

Steinberger, R., & Pouliquen, B. (2011). JRC-Names: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP'2011)* (pp. 104–110). Hissar, Bulgaria.

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EuroVoc. *Computational Linguistics and Intelligent Text Processing*, 115–424.

TaaS. (2014a). *Final Publishable Summary* (p. 46). TaaS Project: Terminology as a Service.

TaaS. (2014b). Public Deliverable D4.4 Integration with SMT Systems. TaaS Project: Terminology as a Service.

Thurmair, G. (2004). Comparing Rule-Based and Statistical MT Output. In *Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora* (pp. 5–9). Lisbon, Portugal.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the sixth conference on Natural Language Learning* (pp. 142–147). Taipei, Taiwan: Association for Computational Linguistics.

TTC. (2013). *Public Deliverable D7.3: Evaluation of the Impact of TTC on Statistical MT* (p. 38). TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora.

Vancāne, I., Krugļevskis, V (2003). *Vārdkopterminu struktūra un datorizēta meklēšana tekstos*. In Linguistica LETTICA. Latvian Language Institute, Rīga, Latvia.

Vasiļjevs, A., Rirdance, S., & Liedskalnins, A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)* (pp. 213–220). Hong Kong.

Vasiļjevs, A., Kalniņš, R., Pinnis, M., & Skadiņš, R. (2014a). Machine Translation for e-Government - the Baltic Case. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA 2014), vol. 2: MT Users* (pp. 181–193). Vancouver, BC Canada.

Vasiļjevs, A., Pinnis, M., & Gornostay, T. (2014b). Service Model for Semi-Automatic Generation of Multilingual Terminology Resources. In *Proceedings of the 11th Conference on Terminology and Knowledge Engineering (TKE 2014)* (pp. 67–76). Berlin, Germany.

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

Vogel, S. (2003). SMT Decoder Dissected: Word Reordering. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering* (pp. 561–566). IEEE.

Vu, T., Aw, A. T., & Zhang, M. (2008). Term Extraction through Unithood and Termhood Unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)* (pp. 631–636). Hyderabad, India: Asian Federation of Natural Language Processing.

Wentland, W., Knopp, J., Silberer, C., & Hartung, M. (2008). Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Wermter, J. (2008). *Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods*. Friedrich Schiller University Jena.

Wermter, J., & Hahn, U. (2006). You Can’t Beat Frequency (Unless You Use Linguistic Knowledge) – A Qualitative Evaluation of Association Measures for Collocation and Term Extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 785–792). Association for Computational Linguistics.

Wohlin, C., Höst, M., & Henningsson, K. (2003). Empirical research methods in software engineering. In *Empirical methods and studies in software engineering* (pp. 7–23). Springer.

Wolf, P., Bernardi, U., Federmann, C., & Hunsicker, S. (2011). From Statistical Term Extraction to Hybrid Machine Translation. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation* (pp. 225–232). Leuven.

Wong, W., Liu, W., & Bennamoun, M. (2008). Determination of Unithood and Termhood for Term Recognition. In *Handbook of Research on Text and Web Mining Technologies*. IGI Global.

APPENDIX

Appendix 1: Publications in Peer Reviewed Conference Proceedings. The appendix contains a list of publications by the author (including co-authored publications) from peer-reviewed international conference proceedings. Only publications relevant to the topics discussed in the thesis are included in the list.