EIROPAS SOCIĀLAIS FONDS
EIROPAS SAVIENĪBA
IEGULDĪJUMS TAVĀ NĀKOTNĒ
LATVIJAS UNIVERSITĀTE ANNO 1919

Terminology Integration in Statistical Machine Translation

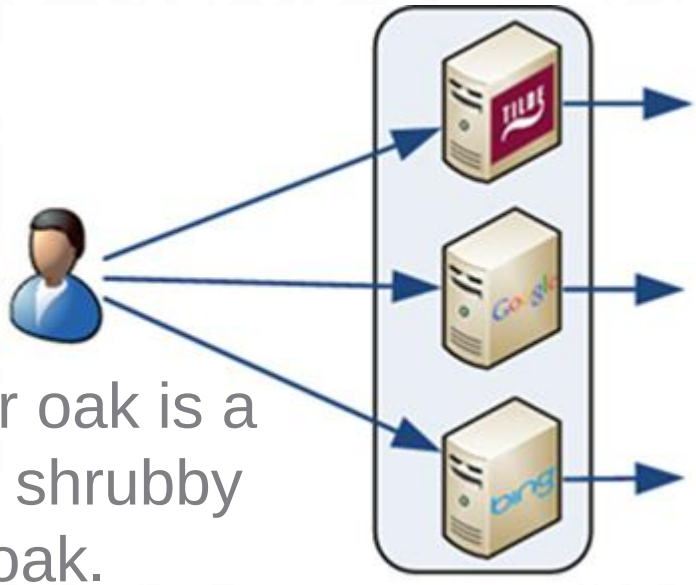
Mārcis Pinnis

Supervisor: Dr.sc.comp. Inguna Skadiņa

LATVIJAS UNIVERSITĀTE ANNO 1919
UNIVERSITY OF LATVIA

December 15, 2015

Translation of terminology using Statistical Machine Translation



A bear oak is a small shrubby oak.

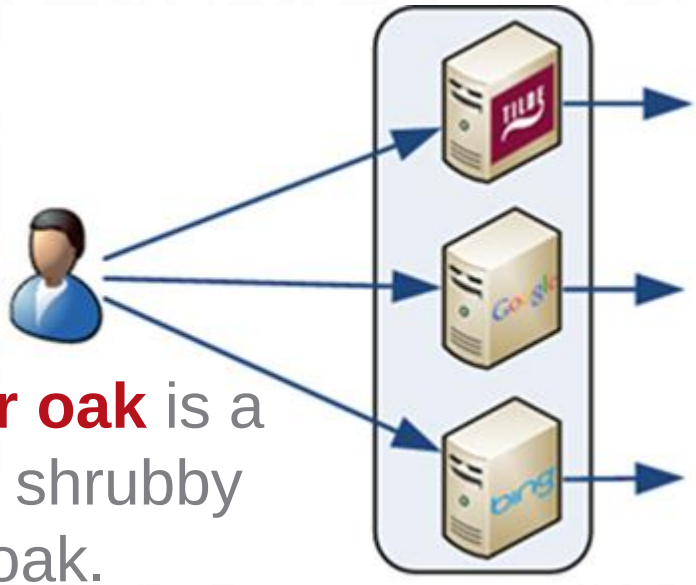
Lācis ozols ir nelielā kārklu ozols.

Lācis ozols ir mazs shrubby ozols.

Lācis ozols ir maza shrubby ozols.

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Translation of terminology using Statistical Machine Translation



A **bear oak** is a small shrubby oak.

Lācis ozols ir nelielā kārklu ozols.

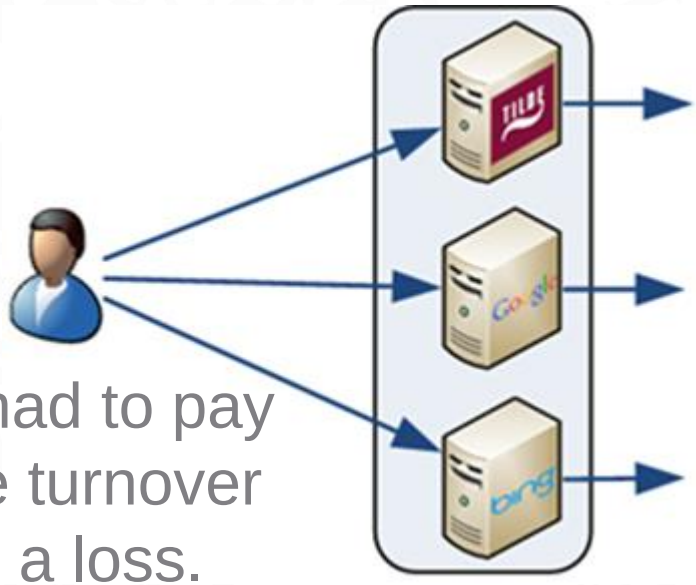
Lācis ozols ir mazs shrubby ozols.

Lācis ozols ir maza shrubby ozols.

bear oak = pundurozols

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Translation of terminology using Statistical Machine Translation



They had to pay for the turnover with a loss.

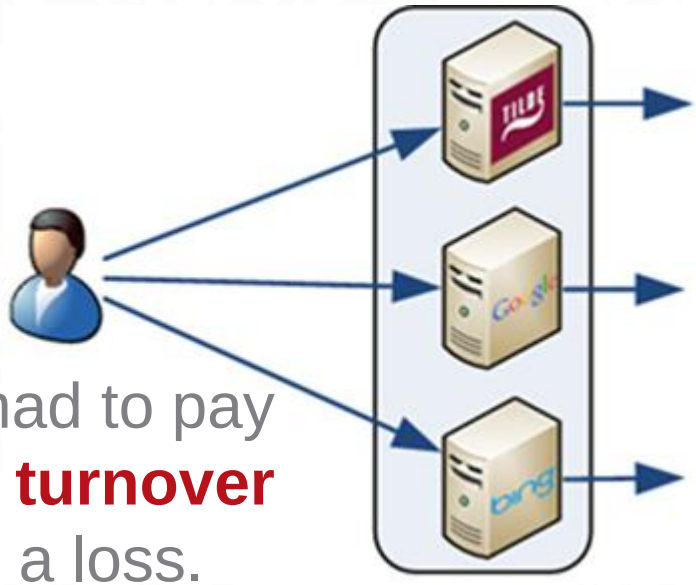
Viņi bija jāmaksā par apgrozījumu ar zaudējumiem.

Viņiem bija jāmaksā par apgrozījumu ar zaudējumiem.

Viņiem nācās maksāt apgrozījuma ar zaudējumiem.

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Translation of terminology using Statistical Machine Translation



They had to pay for the **turnover** with a loss.

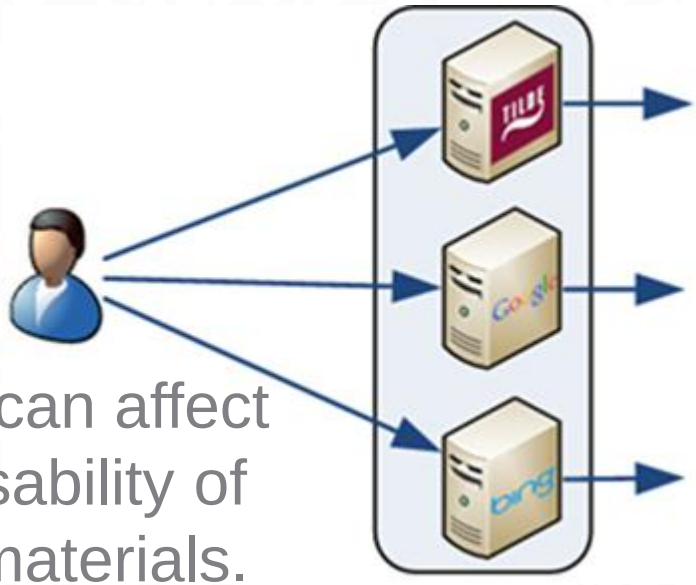
Viņi bija jāmaksā par apgrozījumu ar zaudējumiem.

Viņiem bija jāmaksā par apgrozījumu ar zaudējumiem.

Viņiem nācās maksāt apgrozījuma ar zaudējumiem.

turnover = bumbas pāreja pretiniekam
turnover = apgrozījums

Translation of terminology using Statistical Machine Translation



Humidity can affect the creasability of various materials.

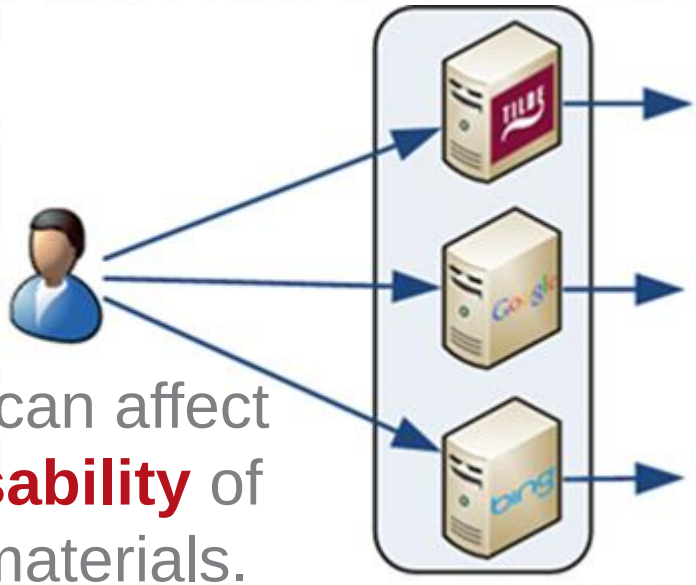
Mitrums var ietekmēt creasability no dažādiem materiāliem.

Mitrums var ietekmēt creasability dažādu materiālu.

Mitrums var ietekmēt dažādu materiālu creasability.

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Translation of terminology using Statistical Machine Translation



Humidity can affect the **creasability** of various materials.

Mitrums var ietekmēt creasability no dažādiem materiāliem.

Mitrums var ietekmēt creasability dažādu materiālu.

Mitrums var ietekmēt dažādu materiālu creasability.

creasability = lieces izturība

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Relevance of the Research Problem

- **SMT systems** (like all data-driven MT methods) **learn their models from large amounts of parallel data**
  - The data may not contain the necessary terms
  - The data may contain ambiguous terms
  - For morphologically rich languages, the data may not contain the terms in the necessary inflected forms
- This results in:
  - **Literal translation** of the words that comprise multi-word terms
  - Selection of a **translation from the wrong domain**
  - **Non-translation** of the terms
  - Translation with the **wrong inflected forms**

Object of Research

- **Methods and algorithms** for terminology integration in statistical machine translation

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Aim

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

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Objectives

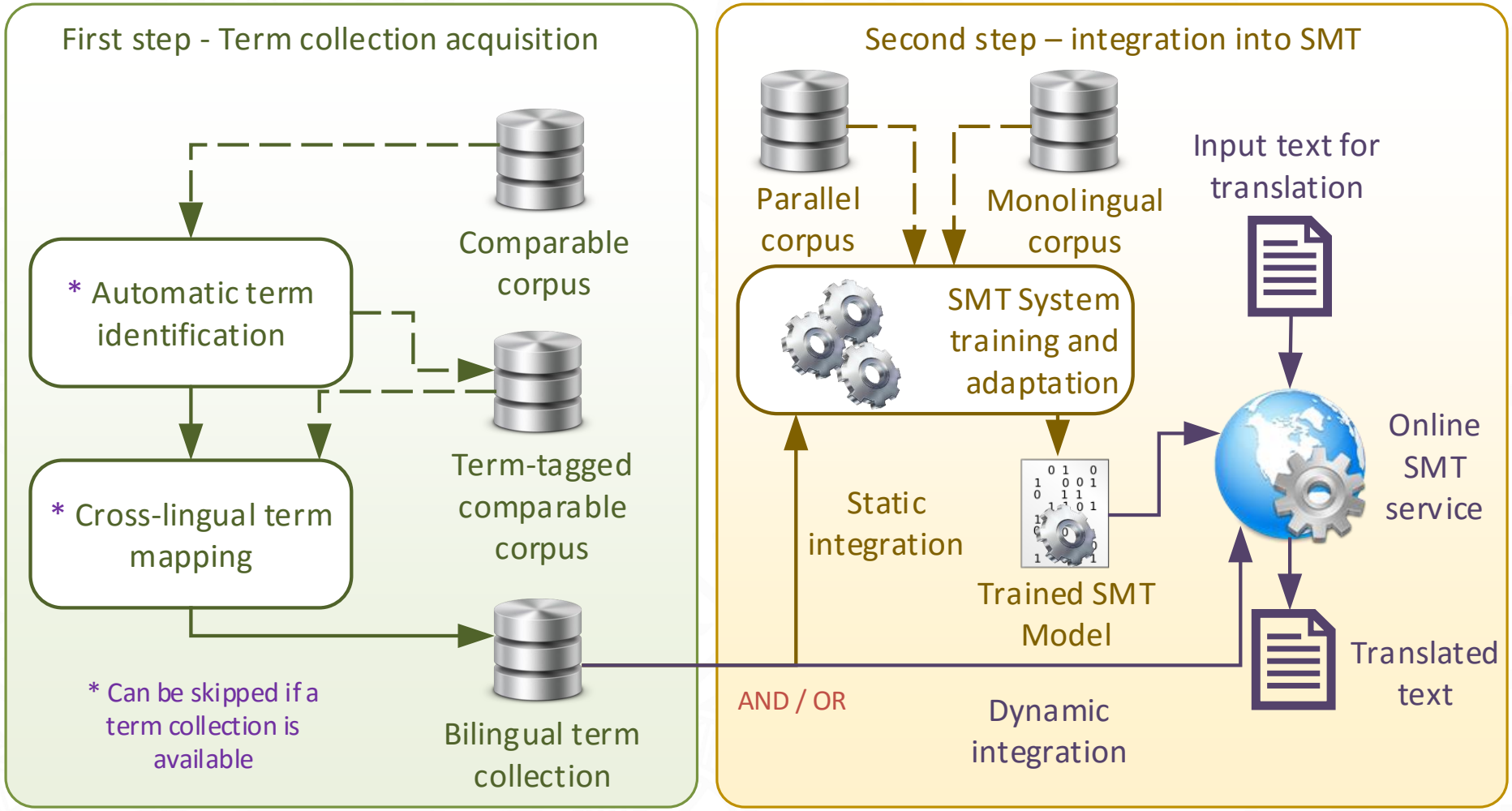
- To research methods and develop tools for:
  - **static terminology integration** in SMT systems
  - **dynamic terminology integration** in SMT systems
  - **term identification**
  - **cross-lingual term mapping**
  - **languages with complex morphologies** and **little** (or no) **parallel resources** in specific domains
- To evaluate the methods for the **English-Latvian language pair** and where applicable also other European languages

Research Hypotheses

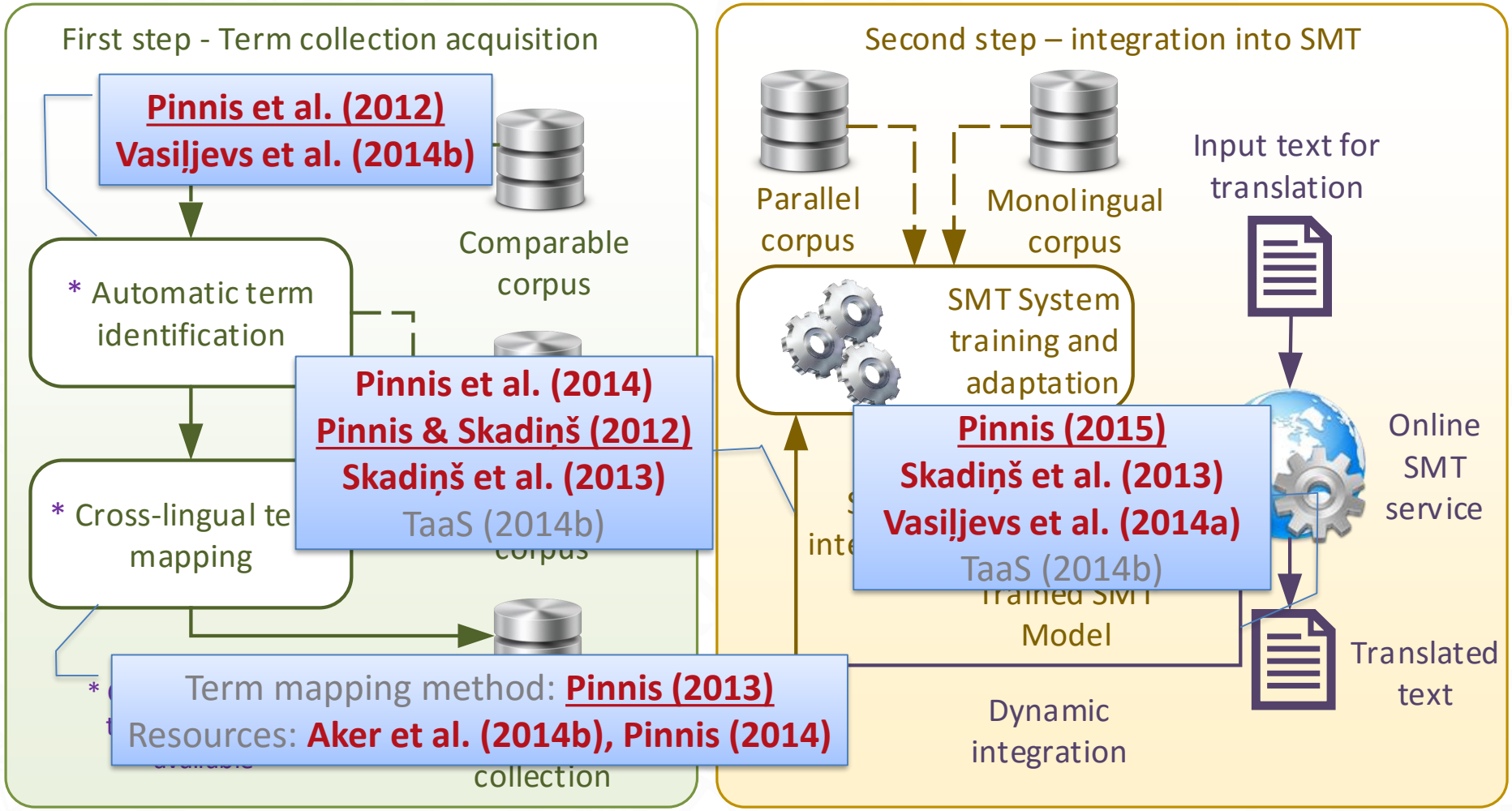
- Terminology translation quality as well as text translation quality in SMT systems can be improved by performing **static and dynamic terminology integration in SMT systems**
- In situations when authoritative term collections are not available, **automatic term identification in comparable corpora and cross-lingual term mapping** are effective methods to acquire bilingual term collections for the integration in SMT systems

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Workflow for Terminology Integration in SMT Systems



Workflow for Terminology Integration in SMT Systems



Term Identification (1)

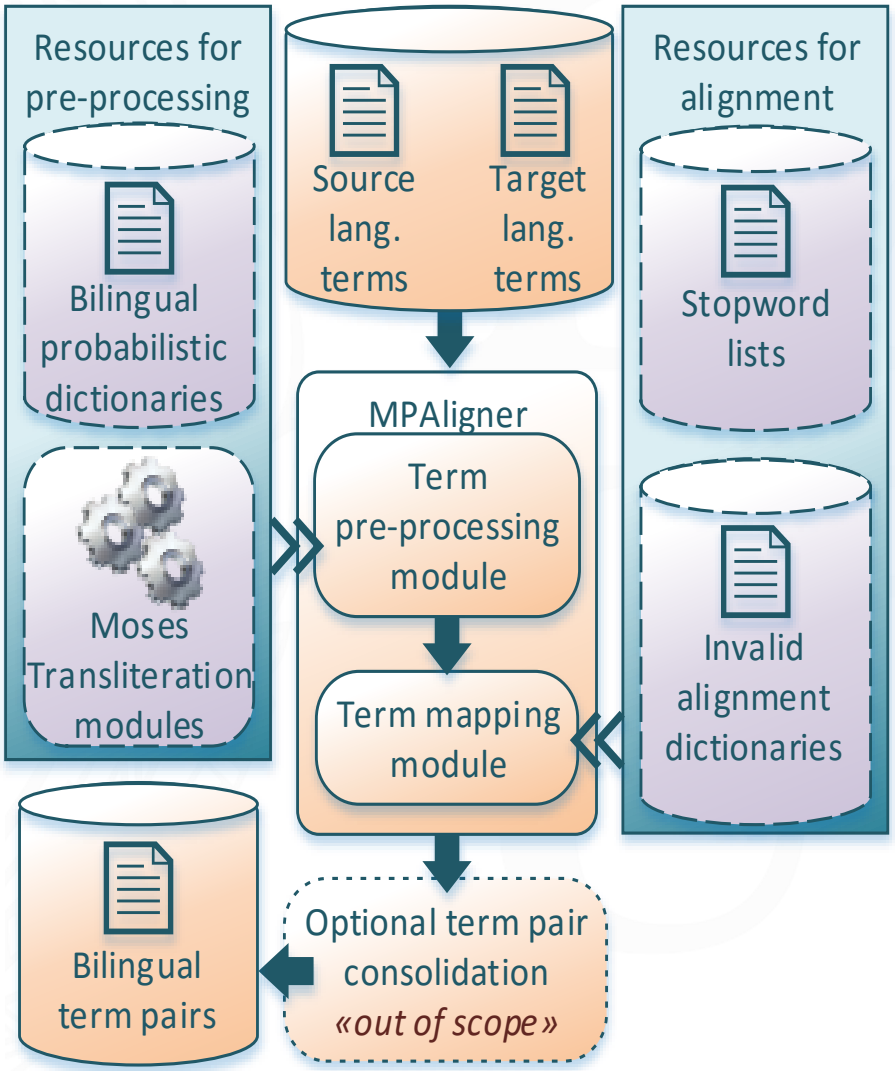
- Term identification methods:
  - **Analyse text** (a sentence, a paragraph, a document, or even a corpus)
  - **Identify** for each phrase (a single-word or multi-word unit) whether it can be **a term candidate** or not
  - Optionally, restrict the identified terms to a given term collection (for translation purposes)

Term Identification (2)

- Term identification for term collection creation
  - **Tilde's Wrapper System for CollTerm** (**TWSC**)
  - The innovative:
    - Combination of **linguistically, statistically, and reference corpus motivated term extraction** with **document level term tagging**
- Term identification for statistical machine translation
  - **Pattern-based Term Identification** (**Pattern**)
  - **Fast Term Identification** (**Fast**)
  - The innovative:
    - Allow the SMT systems to identify terms in different inflected forms (not covered by related research)
    - Suitable for morphologically rich languages

Cross-Lingual Term Mapping - *MPAligner*

- **Context independent**
- **Supports 25 languages**
- The innovative:
  - **Sub-word level term mapping**
  - Application of **character-based SMT transliteration systems**
  - Ability to map **compound terms with multi-word terms**
- Published in Pinnis (2013)



Term Mapping Example (1)

A candidate term pair

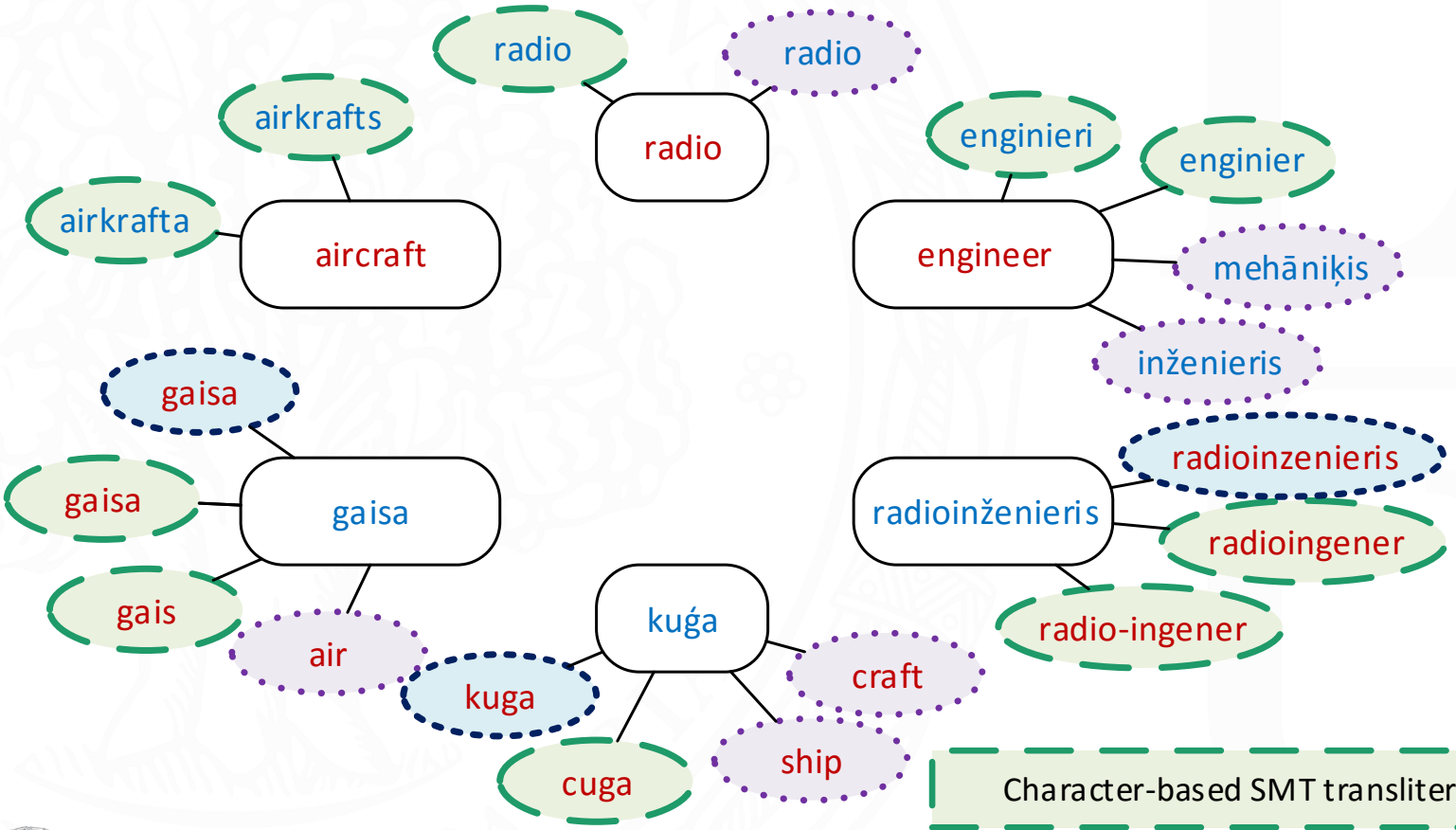
EN: aircraft radio engineer

LV: gaisa kuģa radioinženieris

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Term Mapping Example (2)

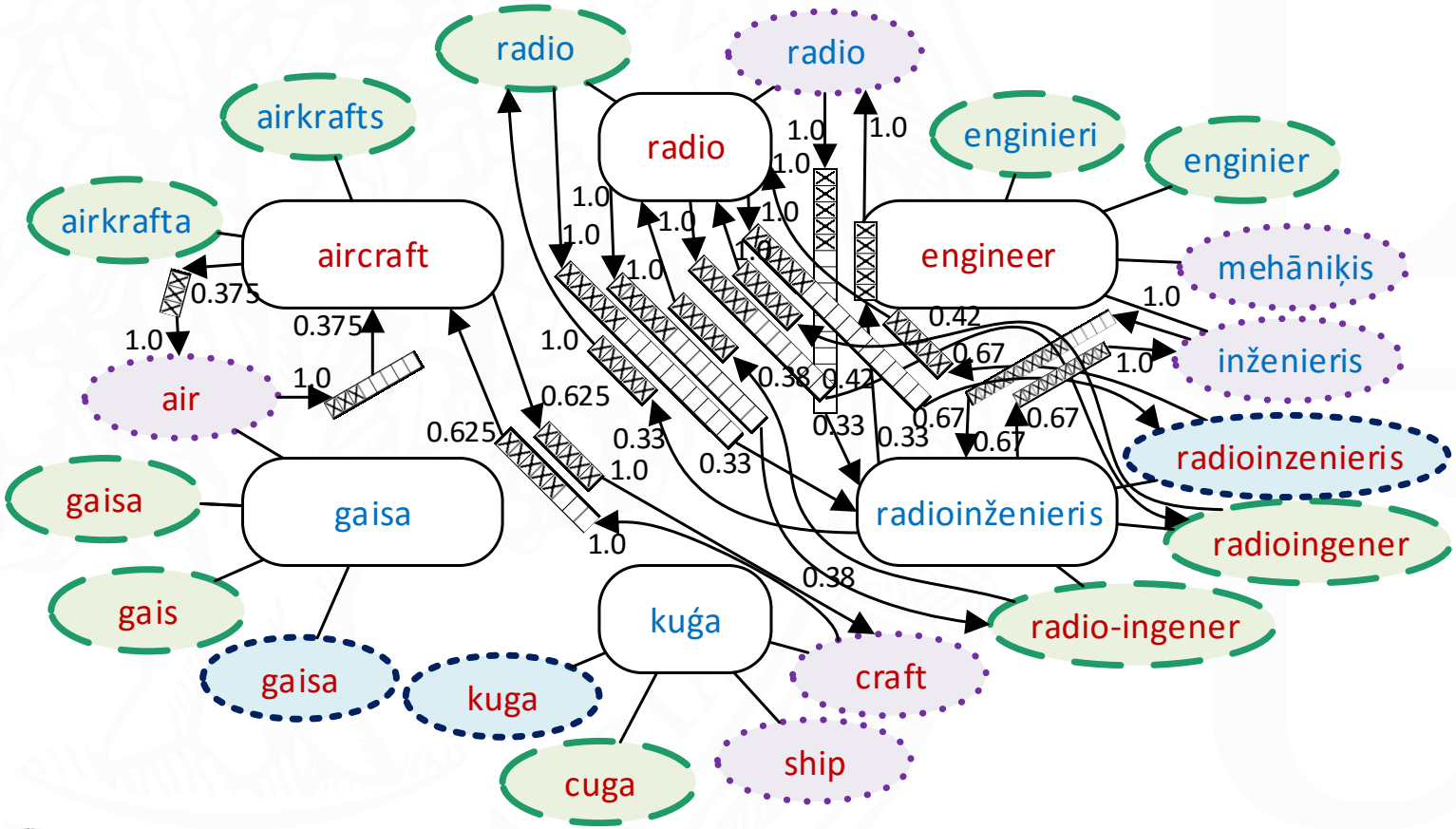
The terms are pre-processed



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Term Mapping Example (3)

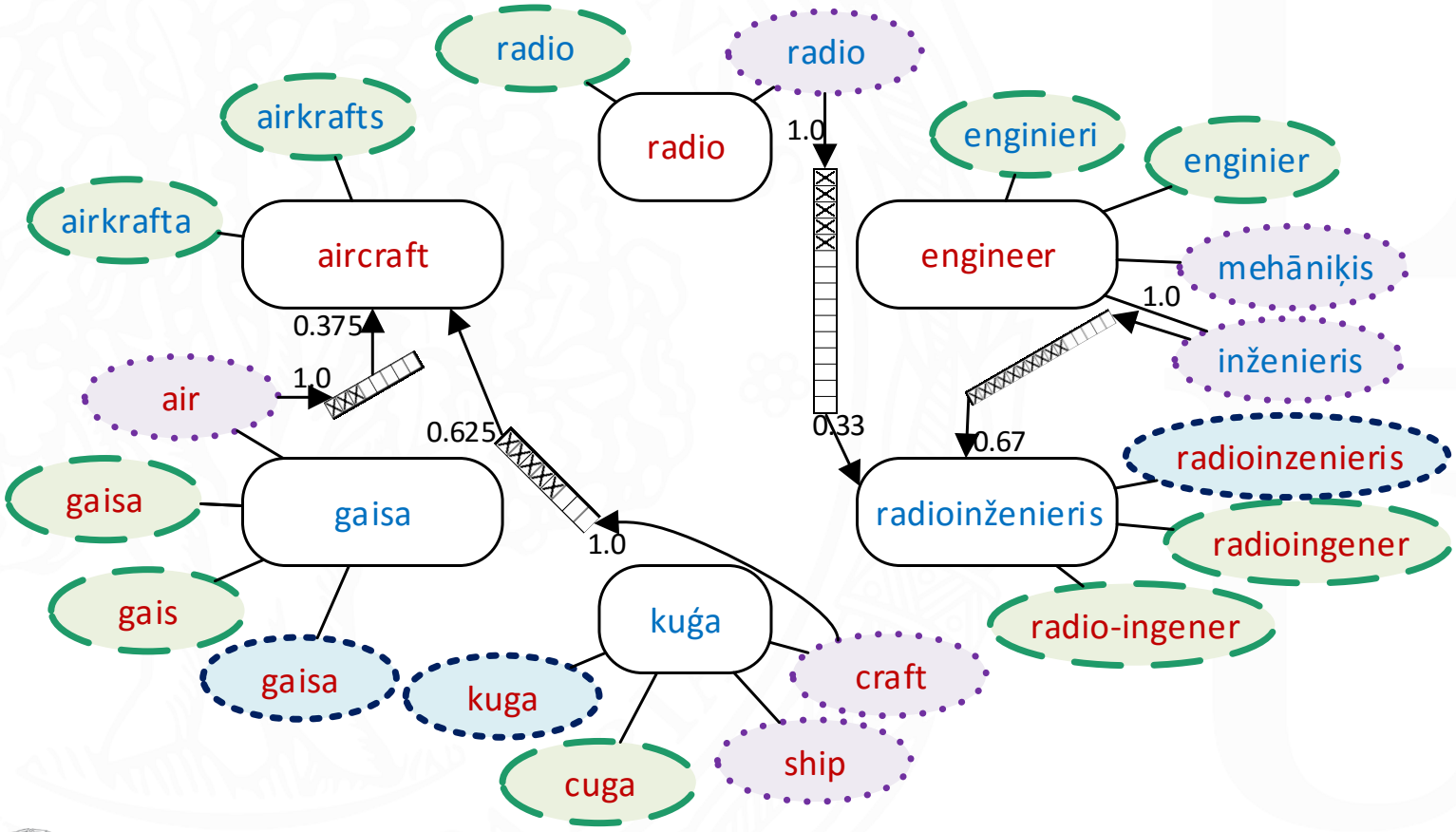
The content overlap is identified



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Term Mapping Example (4)

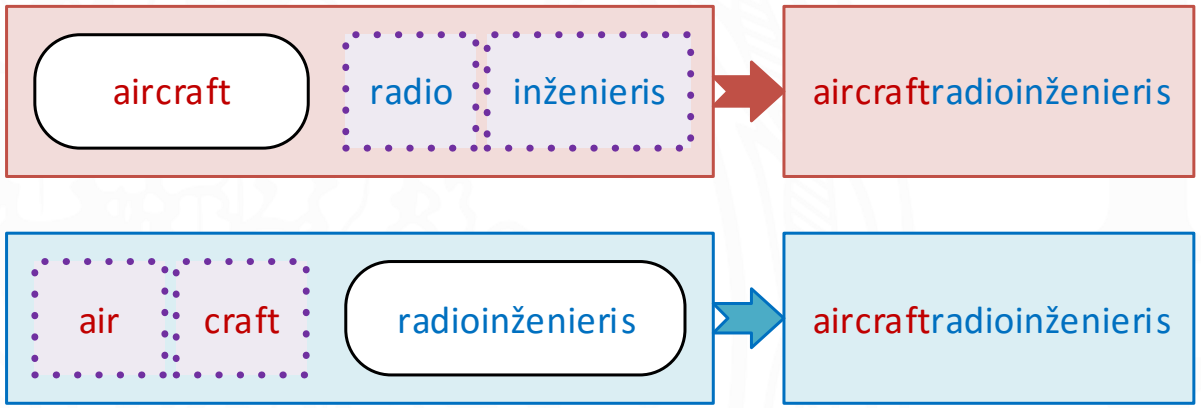
The content overlap is maximised



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Term Mapping Example (4)

Finally, the term pair candidate is scored



$$Similarity(s_1,s_2) = \frac{max(len(s_1),len(s_2))\text{-}LevenshteinDistance(s_1,s_2)}{max(len(s_1),len(s_2))}$$

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Evaluation

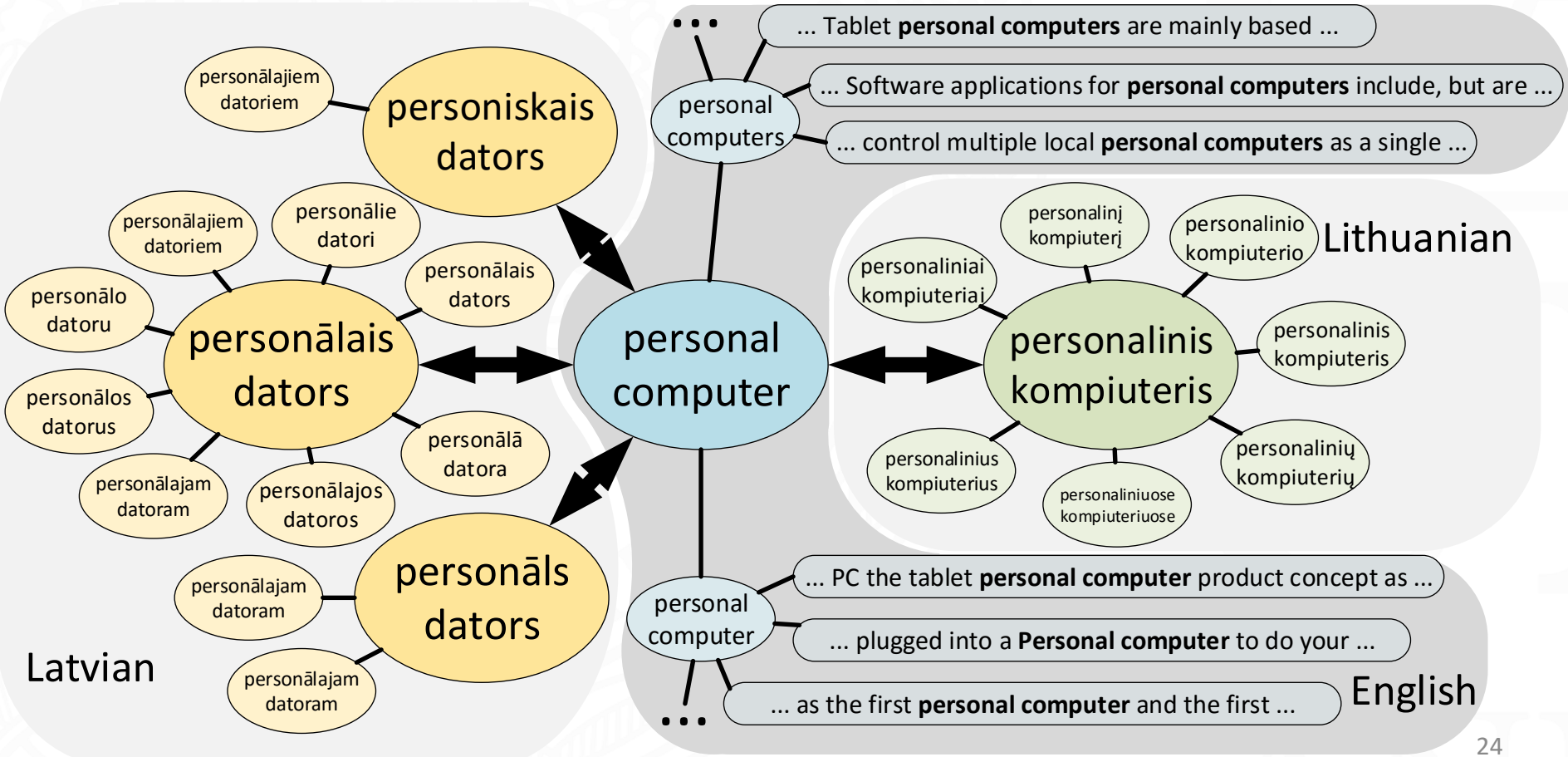
- **Automatic Evaluation** on term lists from the EuroVoc Thesaurus* for **23 languages**
  - **Precision** from 70.7% (en-fi) to **91.3%** (en-lv)
  - **Recall** from 31.8% (en-fi) to **72.2%** (en-mt)
    Recall for en-lv: 62.7%
  - **F1 score** from 43.8 (en-fi) to **80.2** (en-mt)
    F1 score for en-lv: 74.3
    (compared to 46.3 by Ştefănescu (2012))

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

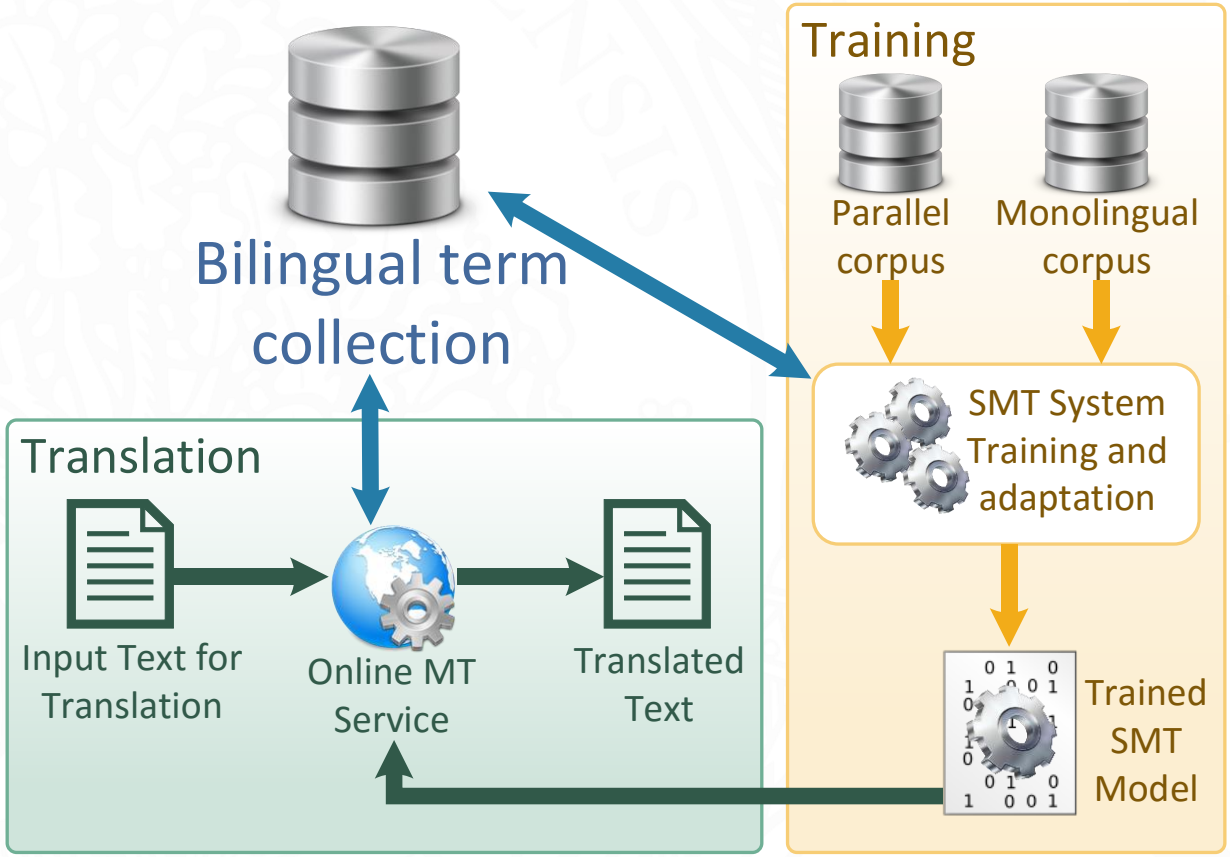
* The EuroVoc thesaurus is the EU's multilingual thesaurus: http://eurovoc.europa.eu/

An Example of Bilingual Terminology in the TaaS Statistical Data Base (created by MPAligner)

- **26 language pairs**, **over 20 million** unique inflected form pairs, **45 subject fields**
- **The largest resource** of automatically extracted bilingual terminology



Terminology Integration in SMT Systems



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Static Terminology Integration in SMT (1)

- Terminology as a corpus (related work)
- **Translation model adaptation**

```
English: jacks Latvian: domkrati

jack of earphones ||| austiņām ||| 0.5 0.009 1 0.325 1 2.71
jack ||| Jack ||| 1 1 0.333 0.111 1 2.718 ...
jack ||| domkrati ||| 1 1 0.333 0.111 2.718 2.718 ...
jack ||| domkratu ||| 1 0.5 0.333 0.222 2.718 2.718 ...
jack-knife ; ||| sasvērties ; ||| 1 0.295 1 0.866 1 2.718 .
```

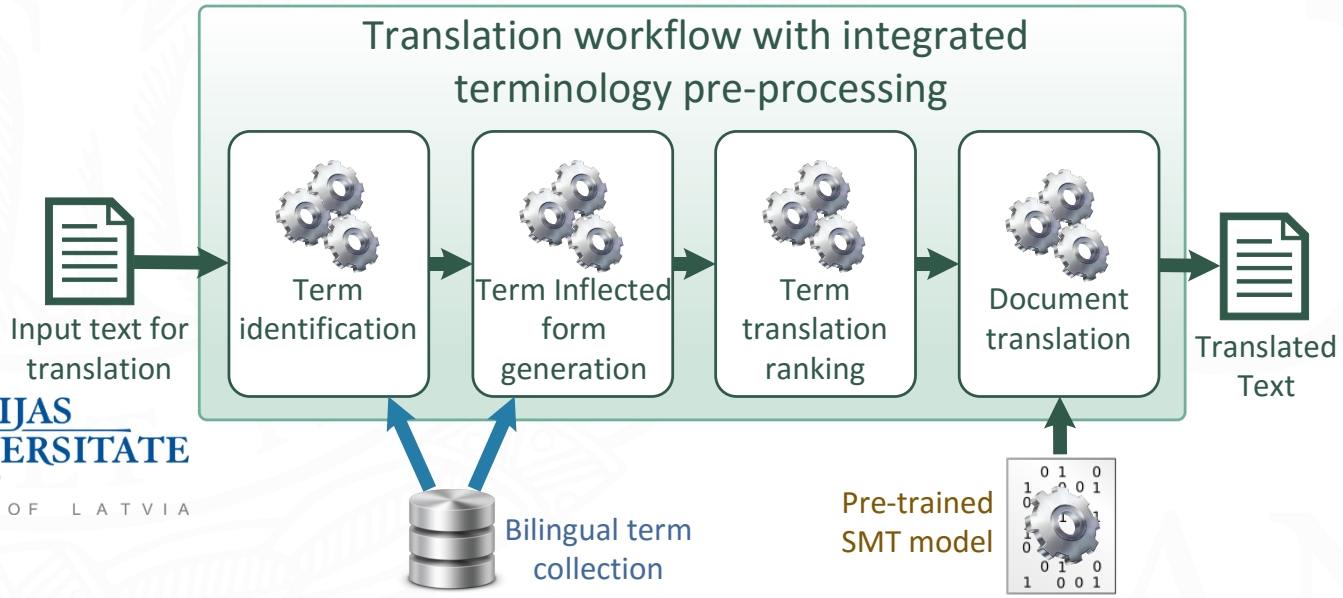
  - Novelty: **terms identified in inflected forms**, Contrary to related work (Arcan et al.,2014a), **adapts the whole phrase table**
  - **Published in Pinnis & Skadiņš (2012)**

Results* (BLEU)
Baseline: 12.68
15.51 (+2.83)
15.96 (+3.28)

** The results are cumulative for an English-Latvian SMT system trained on 1.9 million parallel sentences from the DGT-TM corpus.* ***For terminology integration, a non-filtered automatically extracted term collection was used****.*

Dynamic Terminology Integration in SMT (1)

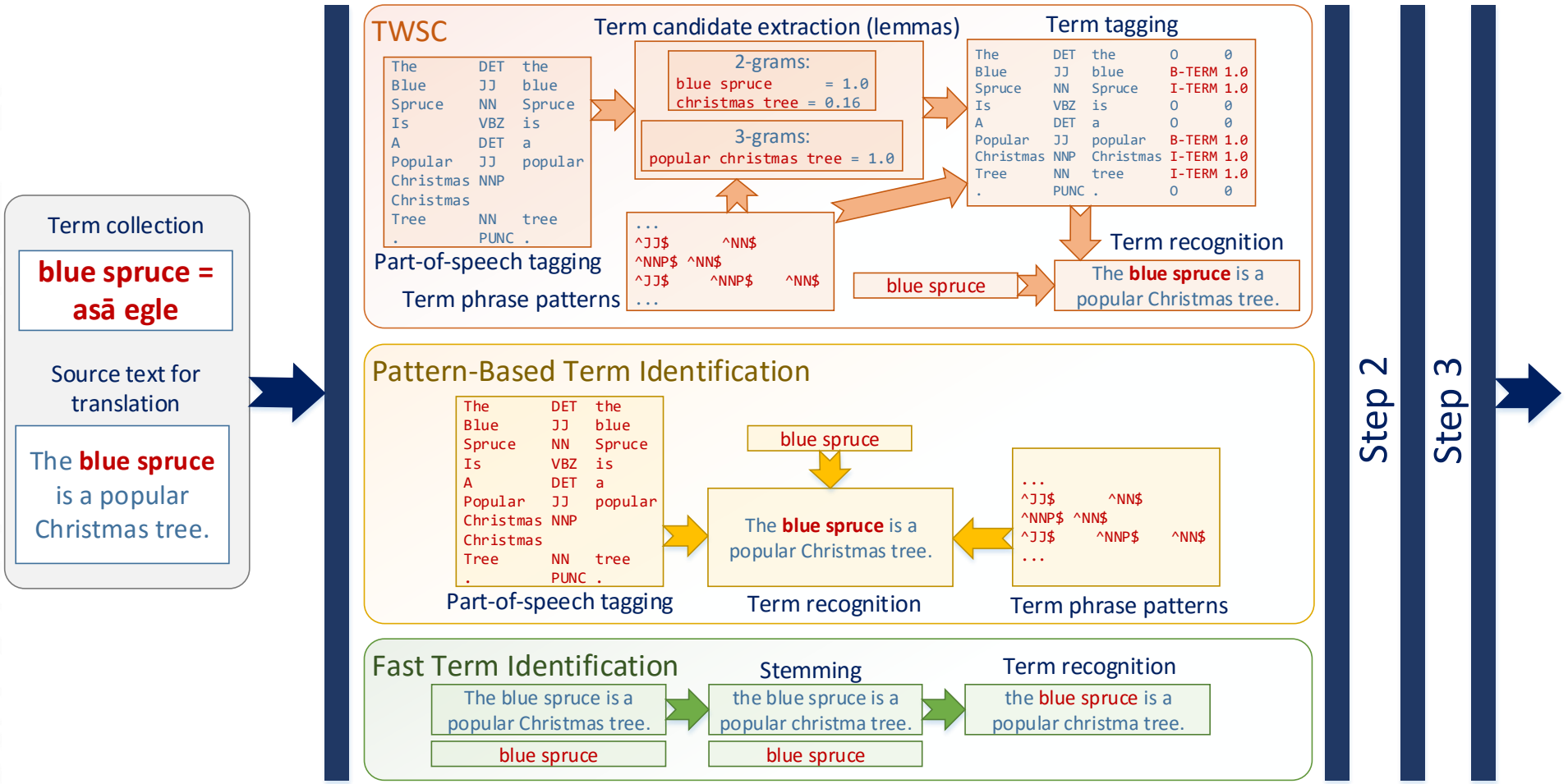
- **Does not require to re-train SMT systems**
- The novelty:
  - **The first** solution designed **for morphologically rich languages**
  - Multi-dimensional workflow with exchangeable components
- Transforms the SMT system into a **hybrid system**
  - Incorporates rule-based features based on **linguistic knowledge**
- The most significant result of the thesis



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

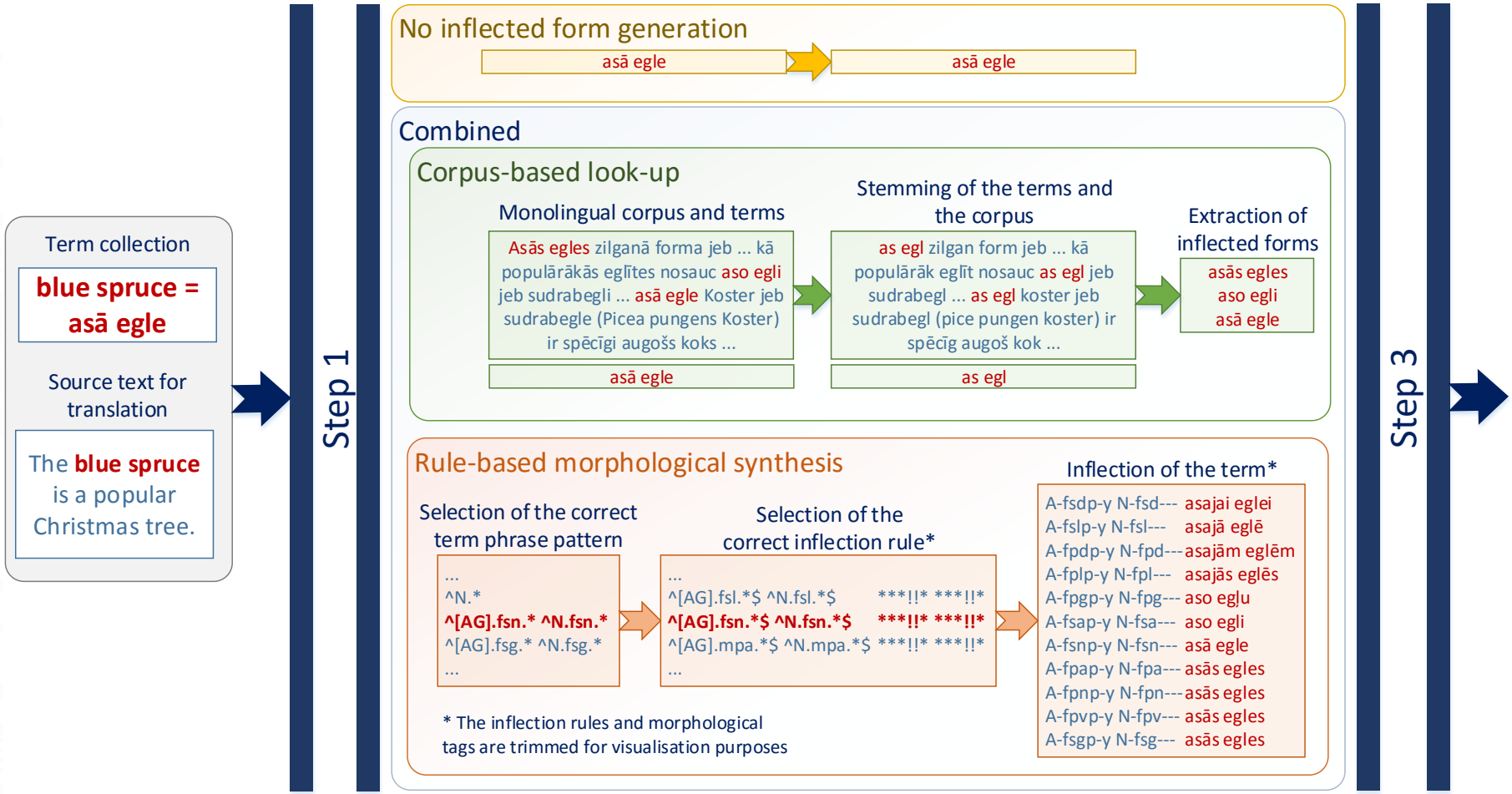
Dynamic Terminology Integration in SMT (2)

- **Step 1:** Term identification



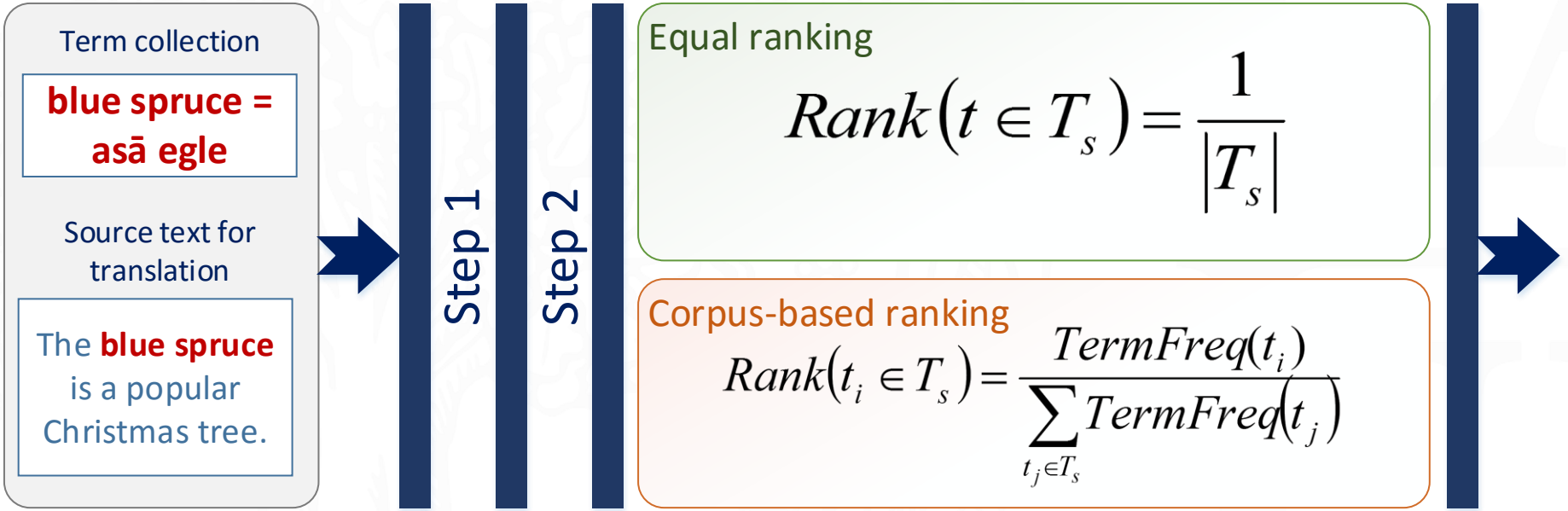
Dynamic Terminology Integration in SMT (3)

- **Step 2:** Inflected form generation for terms



Dynamic Terminology Integration in SMT (4)

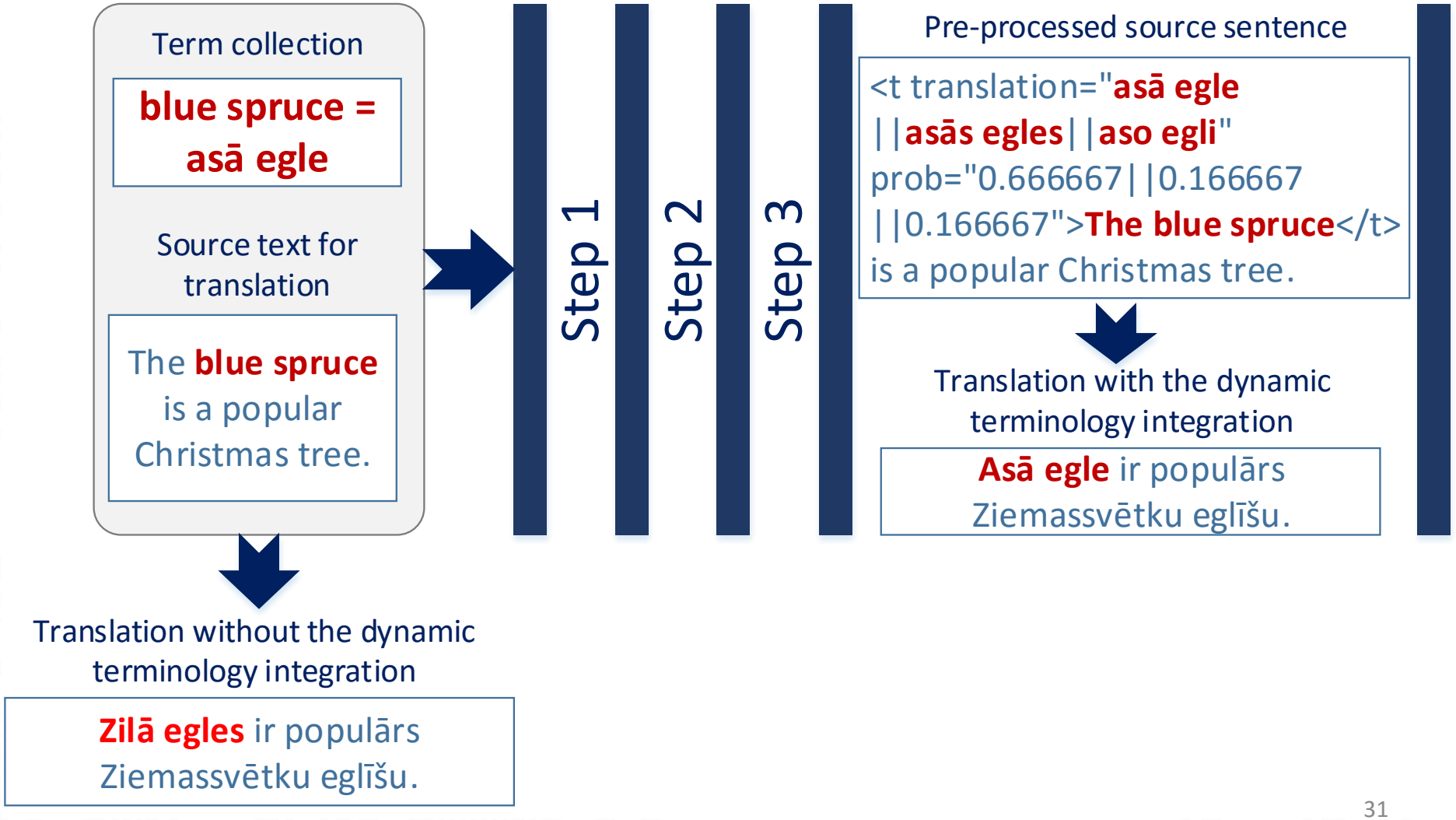
- **Step 3:** Term translation ranking



LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Dynamic Terminology Integration in SMT

- **Step 4:** Translation



Evaluation of Dynamic Terminology Integration in SMT

- Automatic evaluation
  - **4 language pairs**
  - **EN-LV** (↑**26.9%** or **+3.41 BLEU**), EN-LT (↑15.1%), EN-ET(↑6.4%), and EN-DE (↑13.7%)
- Manual evaluation on production systems in the IT domain
  - **7 language pairs**
  - Analysis of translations different from baseline (58%)
  - **Relative improvement of correct term translations**
    - from +1.6% (EN-ET) to **+52.6%** (EN-LT)
    - +32.3% for EN-LV / +30.01% on average
  - **Relative improvement of correct lexical choice** (includes wrong inflections)
    - from +26.4% (EN-DE) to **+65.2%** (EN-LT)
    - +52.4% for (EN-LV) / +45.36% on average
  - **Relative reduction of errors**
    - from +1.0% (EN-ET) to **+72.7%** (EN-RU)
    - +34.0% for EN-LV / +41.4% on average
- **The results clearly prove both hypotheses**

Scientific Novelty (1)

- The **linguistically**, **statistically**, and **reference corpus motivated term identification** method for semi-automatic creation of term collections (Pinnis et al., 2012)
- The **Fast Term Identification** method for term identification in SMT system training data designed **for morphologically rich languages** (Pinnis & Skadiņš, 2012; Pinnis, 2015)
- The **context independent cross-lingual term mapping** method that performs term mapping in **sub-word level** using **maximised character alignment maps** (Pinnis, 2013)

Scientific Novelty (2)

- The **probabilistic dictionary filtering** method using character-based SMT **transliteration systems** (Aker et al., 2014)
- The **character-based SMT transliteration system creation** method using transliteration dictionaries that have been automatically created with a **bootstrapping** method (Pinnis, 2014)
- The method for **static terminology integration in SMT systems** that transforms the SMT system phrase tables into **term-aware phrase tables** (Pinnis & Skadiņš, 2012)

Scientific Novelty (3)

- The **multi-dimensional method for dynamic terminology integration in SMT systems** using a source text pre-processing workflow (Pinnis, 2015)

Main Results

- **The toolkit for static and dynamic terminology integration in SMT systems**
- The tool for context-independent cross-lingual term mapping – ***MPAligner***, which has been used to create the largest resource of automatically extracted bilingual terminology that is integrated in the **TaaS Statistical Data Base**
- The tool for linguistically, statistically, and reference corpora motivated term identification - ***Tilde's Wrapper System* for *CollTerm* (*TWSC*)**

Approbation of the Work (1)

- **Main research projects:**
  - Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation (**FP7**, 2010-2012)
  - Terminology as a Service (**FP7**, 2012-2014)
  - 2.6. Multilingual Machine Translation (ICT Competence Centre, 2014-2015)

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Approbation of the Work (2)

- **The results** (tools & resources) **have been integrated and serve their purpose in**:
  - The **TaaS Platform** (term.tilde.com)
  - SMT platforms developed by Tilde:
    - **LetsMT** (platform for do-it-yourself SMT system development) (letsmt.eu)
    - **hugo.lv** (SMT system for the Latvian government)
    - **versti.eu** (SMT system for the Vilnius University)

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Publication of the Work

- **17 publications**
  - 10 form the basis of the thesis
  - 4 are indexed in Web of Science
  - 6 are indexed in Scopus
  - 15 are peer reviewed
- The work has been presented in **10 scientific conferences and 3 workshops**

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Publications (1)

- 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**.
- 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**.
- Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of RANLP 2013 (pp. 562–570). Hissar, Bulgaria. Indexed in **Scopus**.

Publications (2)

- 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**.
- Pinnis, M. (2015). Dynamic Terminology Integration Methods in Statistical Machine Translation. In Proceedings of EAMT 2015 (pp. 89–96). Antalya, Turkey.
- 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**.

Publications (3)

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

Publications (4)

- 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.
- 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**.
- 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.

Publications (5)

- 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**.
- 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.
- 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).

Publications (6)

- 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.
- 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**.

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

THANK YOU!

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu projektā «Atbalsts doktora studijām Latvijas Universitātē

ANSWERS TO REVIEWER QUESTIONS

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

The Status of (Phrase-based) Statistical Machine Translation?

- It achieves state-of-the-art results:
  - WMT 2015 – **1st place** among named entries **for all 5 translation directions into English**
  - WMT 2015 – **1st place** among named entries **for 2/5 translation directions from English**
  - WMT 2015 – **2nd place** among named entries **for 2/5 translation directions from English**
- CU-CHIMERA is a hybrid system between a deep syntactic system (CU-TECTO) and a **phrase-based SMT system**
- UEDIN-SYNTAX are **tree-to-string and string-to-tree SMT systems** (trained using Moses and Giza++) with similar models to phrase-based systems
- Neural Machine Translation (NMT) technologies were used by one submission – MONTREAL (Jean et al., 2015)
  - Shared the 1st place with UEDIN-SYNTAX for English-German
  - NMT technologies have a longer history (Forcada & Ñeco, 1997; Castaño & Casacuberta, 1997)
- Other submissions used neural networks as features for phrase-based SMT or to re-score translations in phrase-based SMT

How to tell apart domain-specific terms from general terms?

- Terms can be ranked with respect to specificity using **broad-domain corpus statistics** (e.g., inverse document frequencies)
- In the thesis I discuss it in sections (4.2.1.1 and 5.5.3)

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

- General terms (if they are not ambiguous) are learned well from data
- In dynamic terminology integration in SMT, general terms may be risky if they are ambiguous
- An interesting analysis for future work would be to perform analysis that answers to the following question: at what specificity (or ambiguity) level, terms are not learned well from data?

- *«Term extraction is based on a wild combination of word alignment information with transliteration, applied rules and etc. If the final aim is improving statistical translation without the necessity of extracting a terminology, would it make sense to update the default word alignment by additionally aligning all the detected terminology expressions and their translations?»*
- Question re-phrased (correct me if I am wrong): would it make sense to use MPAligner in combination with Fast Term Identification to improve word alignment in SMT when performing static terminology integration in SMT?
  - It may help to slightly improve the word alignment
  - It solves a problem that is a general SMT problem: how to improve word alignment
  - It is a broader task for multi-word expressions (i.e., not limited to terminology)
  - An interesting idea for future work

Is the development of the TWSC part of this thesis?

- **Yes**, it has been designed and developed by the author in the ACCURAT project and improved in the TaaS project
- For statistical analysis, TWSC uses CollTerm that has been developed by Nikola Ljubešić in the ACCURAT project (see section 2.2 in the thesis)
- TWSC uses also term phrase patterns - the Latvian and Lithuanian term phrase patterns were developed by Inguna Skadiņa and the author (see section 2.2.1 in the thesis).
- POS-taggers, of course, are of respective authors. In the thesis, for Latvian and Lithuanian, I used the tagger by Pinnis & Goba (2011).

Is section 5 on dynamic terminology integration the work of the author only?

- **Yes!**
  - The author's design
  - The author's implementation
  - The author's prepared evaluation tasks
  - The author's performed analysis of results
- **I shall say that this is the most significant result of my thesis!**

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

Was the brute-force automatic evaluation justified, or could one assume the independence of some steps and thus the invariance of the other steps to its results?

- The author believes that yes, it was justified!
  - An important question to answer was not whether «*non-filtered*», «*filtered*» or «*professional*» collections achieve higher results, but rather **what works best if you have each of those term collections**? Therefore, all three types had to be analysed.
  - The steps for term identification, inflected form acquisition, and ranking are dependent on each other in a sense that they operate with the data that the previous step produced and the outcome will differ (even within a single term) depending on which step was performed before
  - The «brute-force» approach allows clearly identifying, which are the best combinations
  - Also, the evaluation was performed with automated scripts that prepared the evaluation tasks, the evaluation data, and summarised the results

Why is the value of the additional term coverage feature in the phrase table the same ("1") for phrase pairs where no terms are included on either side and pairs where one phrase includes a term and the other one doesn't? Wouldn't it make sense to not penalize the former?

- Why is the value «1»?
  - There are two approaches: 1) either you give credit to the good, or 2) you penalise the bad
  - I selected the first approach
- However, it would make sense also to analyse the second approach
- **It is a very good suggestion!**
  - for future work


TILDE


TaaS


ACCURAT


KOMPETENCES
CENTRS

THANK YOU!


LATVIJAS
UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA