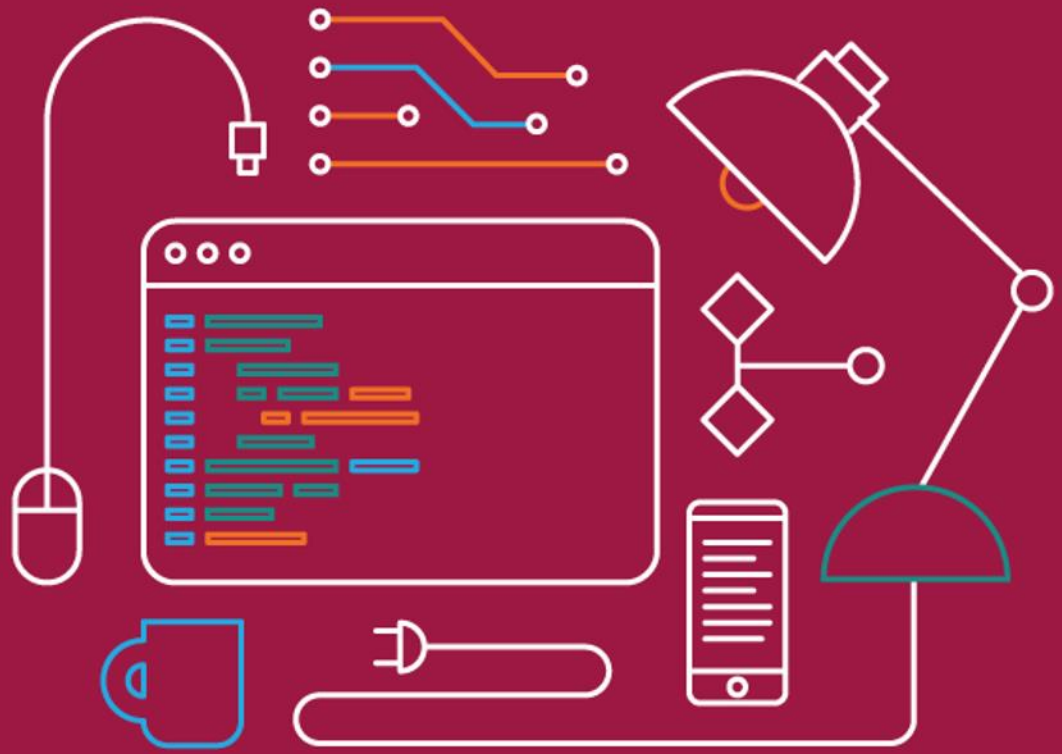


Neural MT and other Language technologies at Tilde

Dr. Raivis SKADIŅŠ

Tilde, Director of Research and Development

Fifth GF Summer School 2017, Riga, August 18, 2017

In my talk

- About Tilde and what we do
- Grammar Checking
- Neural Machine Translation

Tilde

Offices in Riga, Vilnius & Tallinn

135 employees

7 PhDs
150+ research publications

Founded in 1991, Riga

Almost everybody in the Baltic countries uses some Tilde software or product localized by Tilde

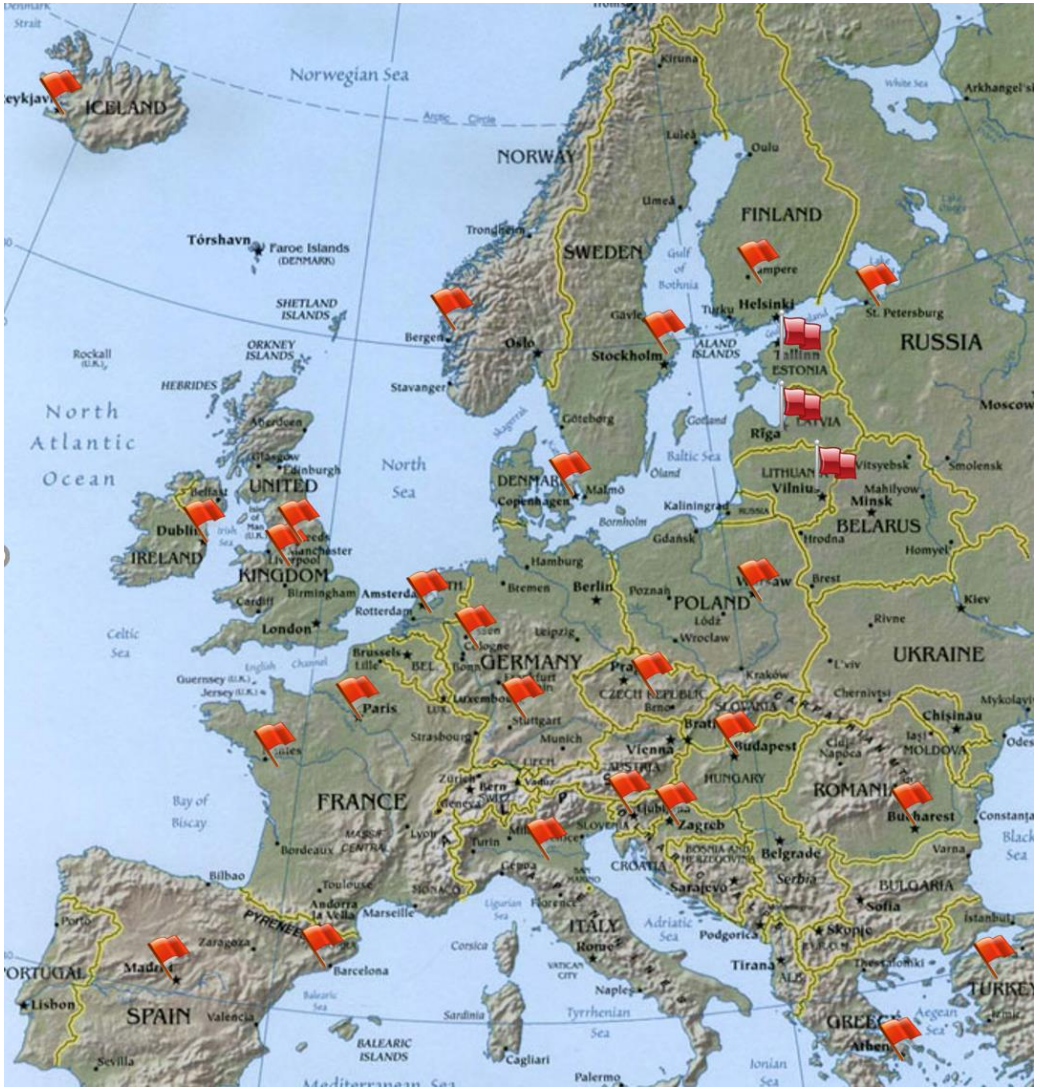
European Commission, Microsoft, IBM, Oracle and other global clients

What we do

- All kinds of language technologies
  - spelling checkers
  - electronic dictionaries
  - terminology
  - encyclopedias
  - grammar checkers
  - machine translation
  - speech recognition and synthesis
  - virtual assistants and chatbots

What we do

- Wide range of clients
  - home and office users
  - localization companies
  - enterprise clients
  - governments
  - EU infrastructure projects
- Research projects



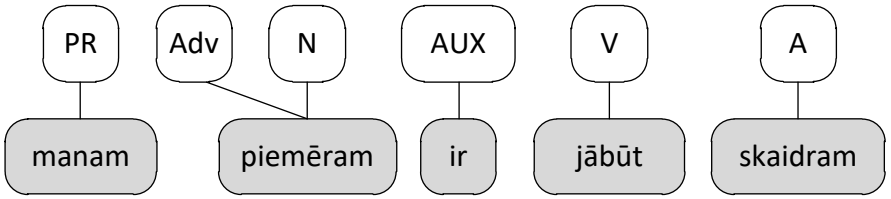
Grammar Checking

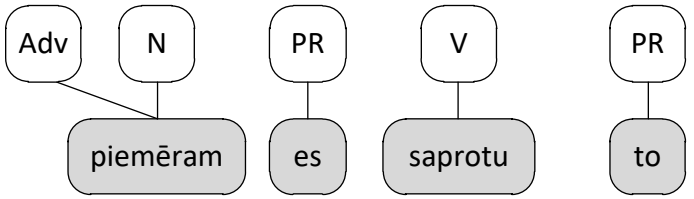
How we do it

- If you can parse the sentence, then it is correct
- **But**, if you cannot parse it
  - It is wrong
  - Your grammar is incomplete

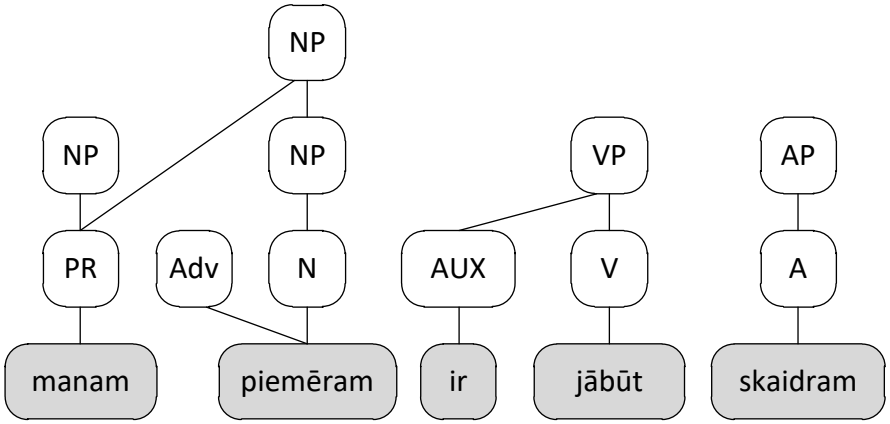
- Is it really so simple?
- Will any parser do?
- How to find the error? How to fix it?

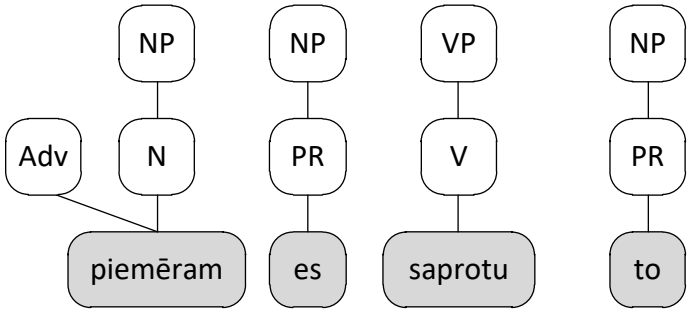
Example


PR
Adv
N
AUX
V
A
manam
piemēram
ir
jābūt
skaidram

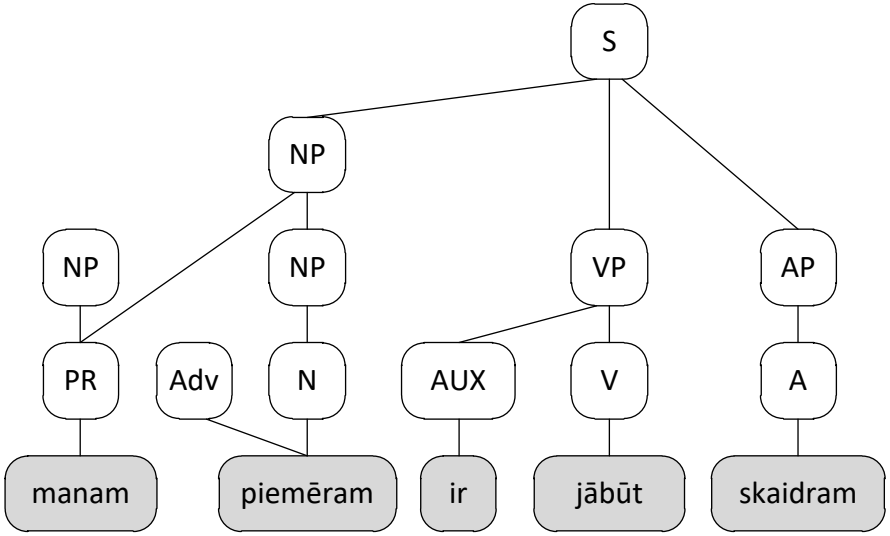

Adv
N
PR
V
PR
piemēram
es
saprotu
to

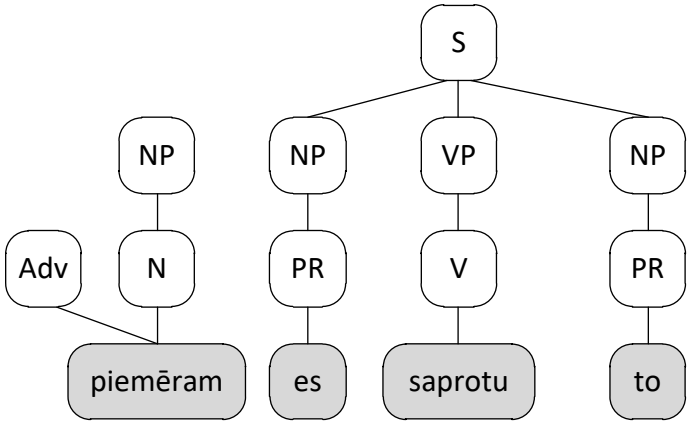
Example


NP
NP
NP
PR
Adv
N
manam
piemēram
VP
AP
AUX
V
A
ir
jābūt
skaidram


NP
NP
VP
NP
Adv
N
PR
V
PR
piemēram
es
saprotu
to

Example


S
NP
NP
NP
VP
AP
PR
Adv
N
AUX
V
A
manam
piemēram
ir
jābūt
skaidram


S
NP
NP
VP
NP
Adv
N
PR
V
PR
piemēram
es
saprotu
to

Some examples of rules

```
NP -> attr:AP main:NP
   Agree(attr:AP, main:NP, Case, Number, Gender)

S -> subj:NP main:VP obj:NP
   Agree(subj:NP, main:VP, Person)
   subj:NP.Case == Nom
   obj:NP.Case == Acc
```

- And there are hundreds of them; (Deksne et al., 2014)

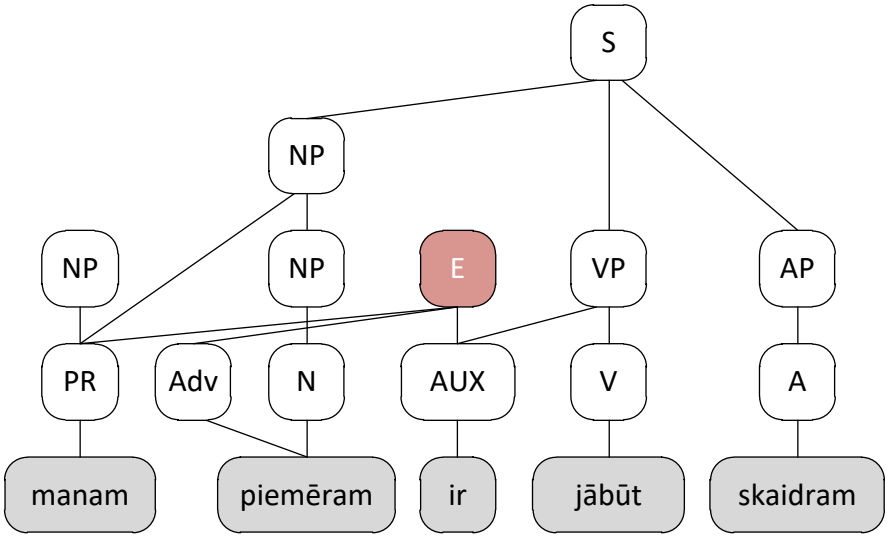
How to find the error?

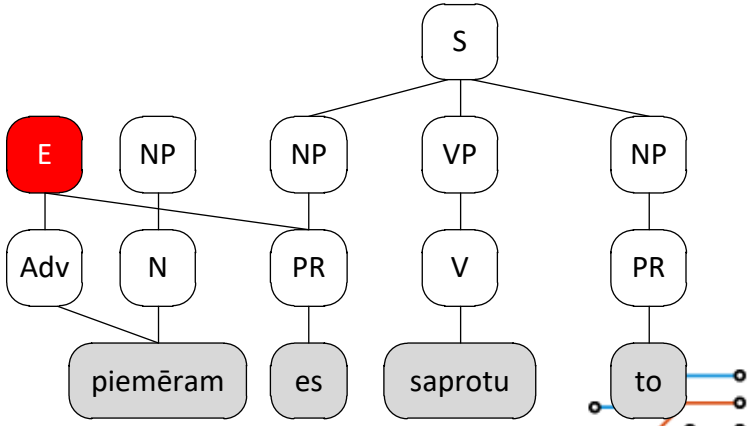
- Two types of rules
  - Regular rules that describe syntax
  - Rules that describe errors
- We parse the sentence with both at the same time

- There is an error, if
  - an error rule has been applied
  - fragment where it has been applied cannot be parsed with regular rules

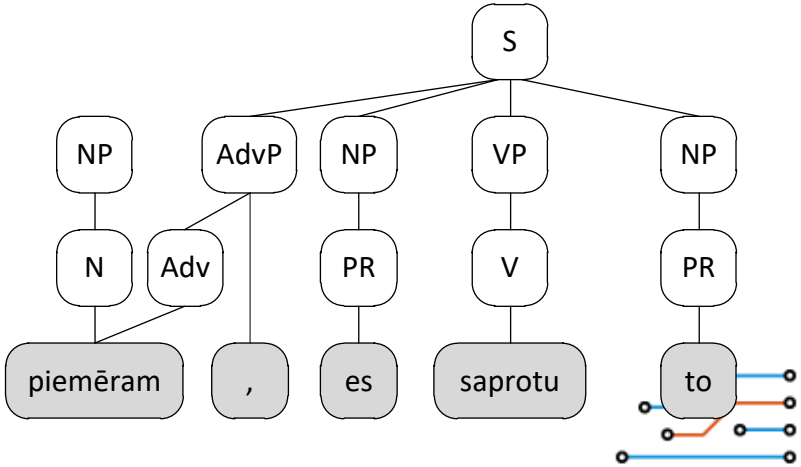
(Deksne & Skadiņš, 2011)

Example

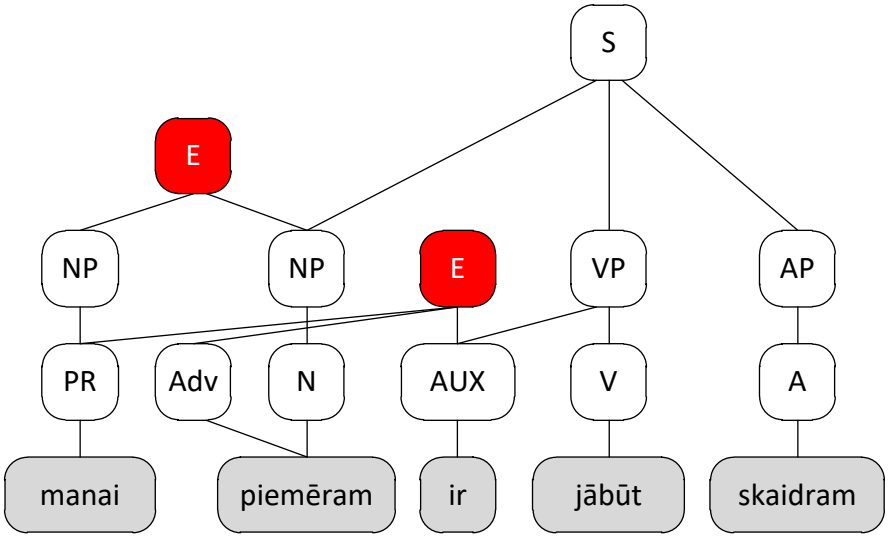

S
NP
NP
NP
E
VP
AP
PR
Adv
N
AUX
V
A
manam
piemēram
ir
jābūt
skaidram


S
E
NP
NP
VP
NP
Adv
N
PR
V
PR
piemēram
es
saprotu
to

Example



Example


S
E
NP
NP
E
VP
AP
PR
Adv
N
AUX
V
A
manai
piemēram
ir
jābūt
skaidram

Error rules

```
ERROR-1 -> attr:AP main:NP
   Disagree(attr:AP,main:NP, Case, Number, Gender)
GRAMMCHECK MarkAll
   attr:AP.Gender=main:NP.Gender
   attr:AP.Number=main:NP.Number
   SUGGEST(attr:AP+main:NP)
```

Error rules

```
ERROR-14 -> attr:N attr:G main:N
    attr:N.Case==genitive
    attr:N.Number==singular
    attr:G.AdjEnd==definite
    main:N.Number==plural
    Agree(attr:G, main:N, Case, Number, Gender)
    CapPattern fff
LEX Amerika savienots valsts
```

Rules

Rule type	Latvian	Lithuanian
Correct syntax rules	580	179
Error rules which depend on phrases described by correct syntax rules	263	72
Error rules which contain only terminal symbols	239	560
Total	**1082**	**811**

Evaluation

Corpus	Error type	Precision	Recall	F-measure
Lithuanian Balanced	all error types	0.898	0.412	0.564
	vocabulary errors	0.956	0.535	0.686
	incorrect usage of cases	0.734	0.259	0.383
Latvian Balanced	all error types	0.780	0.455	0.575
	punctuation in sub-clauses	0.757	0.643	0.695
	punctuation in participle clauses	0.617	0.671	0.643
Latvian Student papers (dev)	All error types	0.652	0.231	0.341
	punctuation in sub-clauses	0.706	0.586	0.641
	punctuation in participle clauses	0.656	0.560	0.604
Latvian Student papers (test)	all error types	0.753	0.203	0.320
	punctuation in sub-clauses	0.773	0.588	0.668
	punctuation in participle clauses	0.766	0.685	0.723

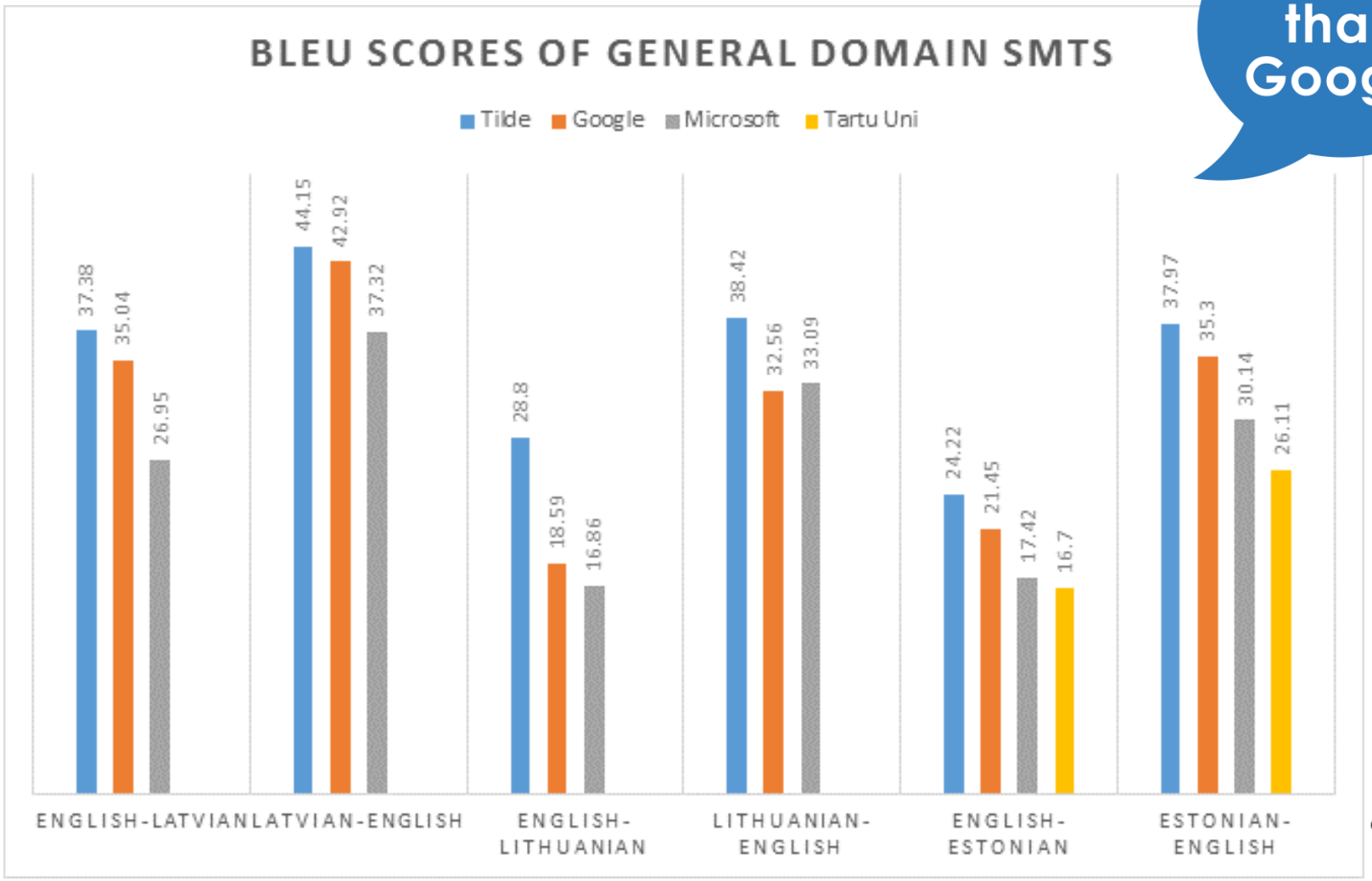
Machine Translation


Rule-based MT
Statistical MT
Neural MT

State-of-the-Art before neural MT?

Phrase-based statistical MT

Better than Google

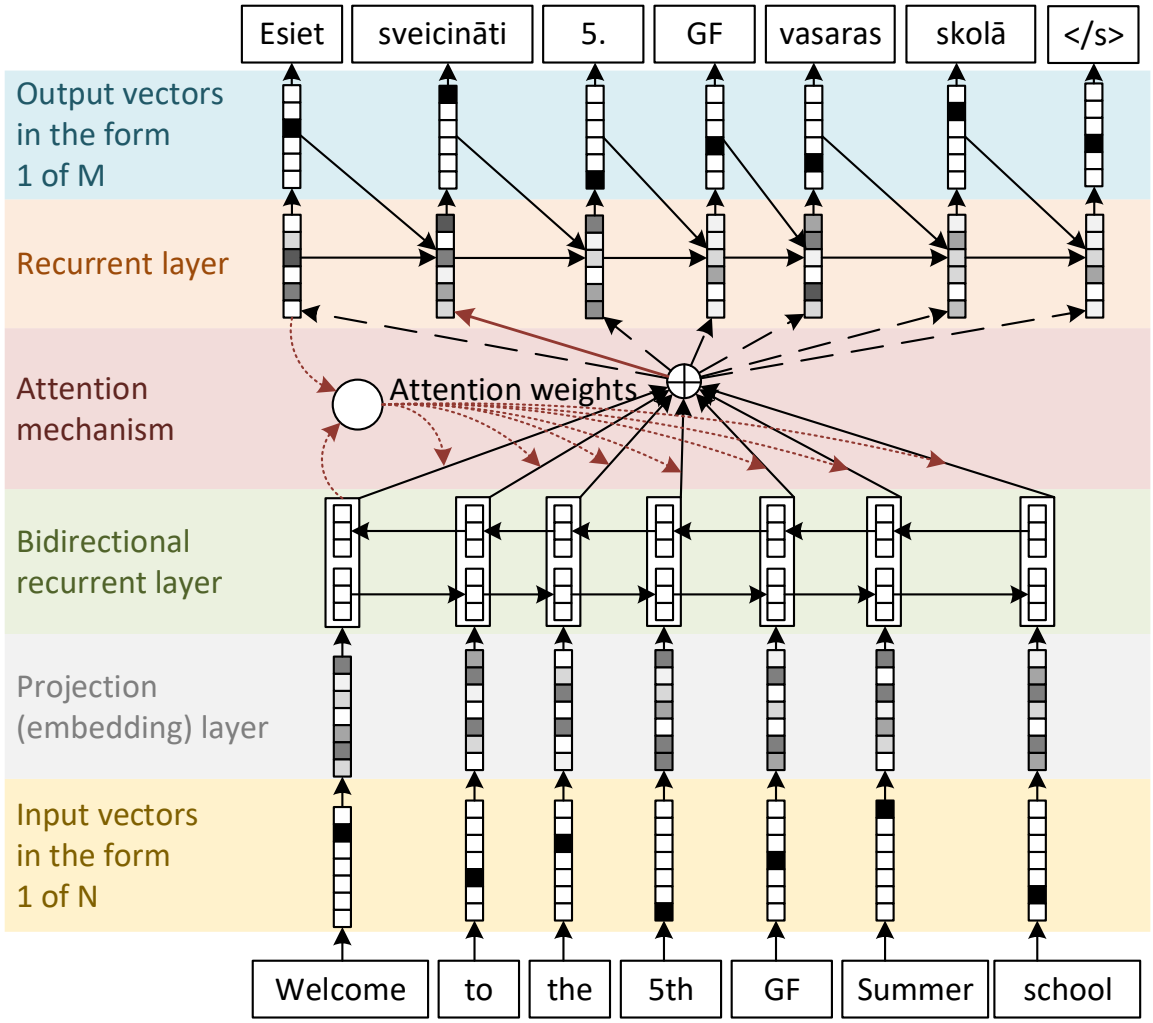


Dawn of the Neural MT

- New technology, 2015, 2016
- Very different architectures

- Many open questions
  - Is it good for Latvian and other under-resourced languages?
  - What is the quality?
  - Strengths and weaknesses?
  - Is it fast enough?
  - What infrastructure do we need?
  - etc.

Technology



- QT21 project
- Nematus and AmuNMT toolkits
- end-to-end NMT

- sub-word tokens (BPE)

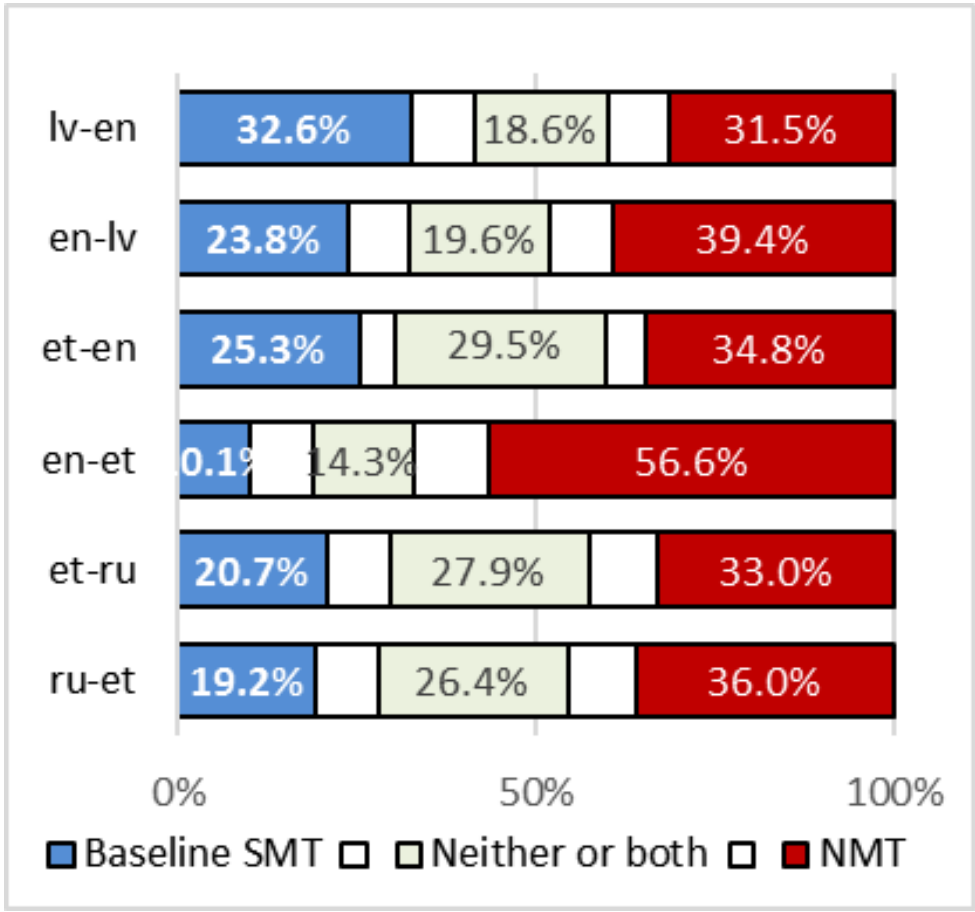
Training data

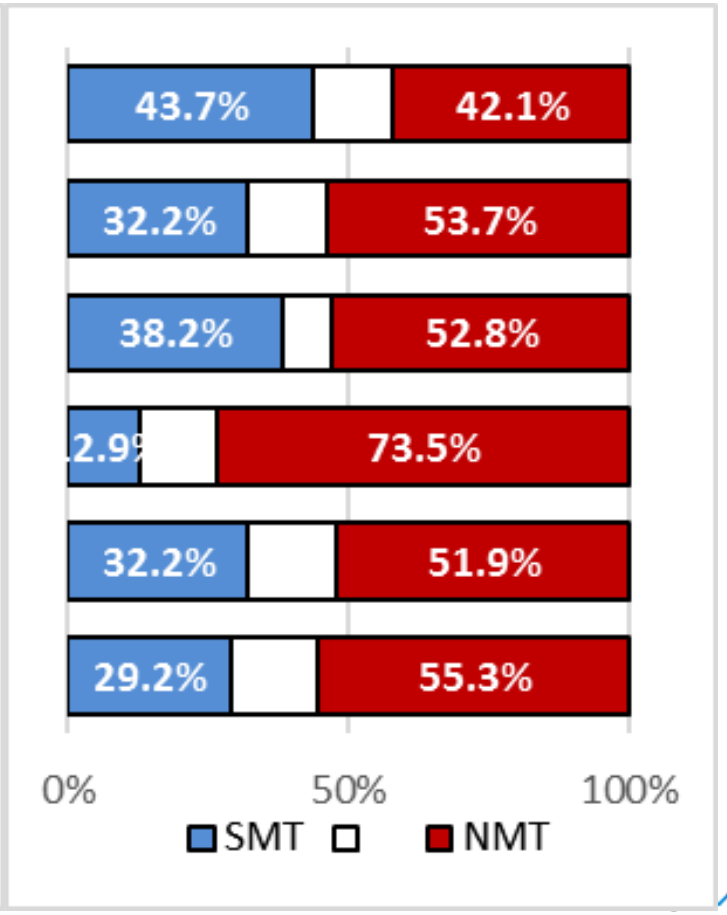
Language pairs	Sentences in parallel corpus	Sentences in monolingual corpus
General domain		
en-et	21 900 622	48 567 363
et-en	21 900 794	217 724 716
ru-et	4 179 198	48 606 392
et-ru	4 179 153	138 001 100
en-lv	7 477 785	74 741 452
lv-en	7 476 956	95 259 699
Pharmaceutical domain		
en-lv	316 443	309 182

Automatic evaluation (BLEU)

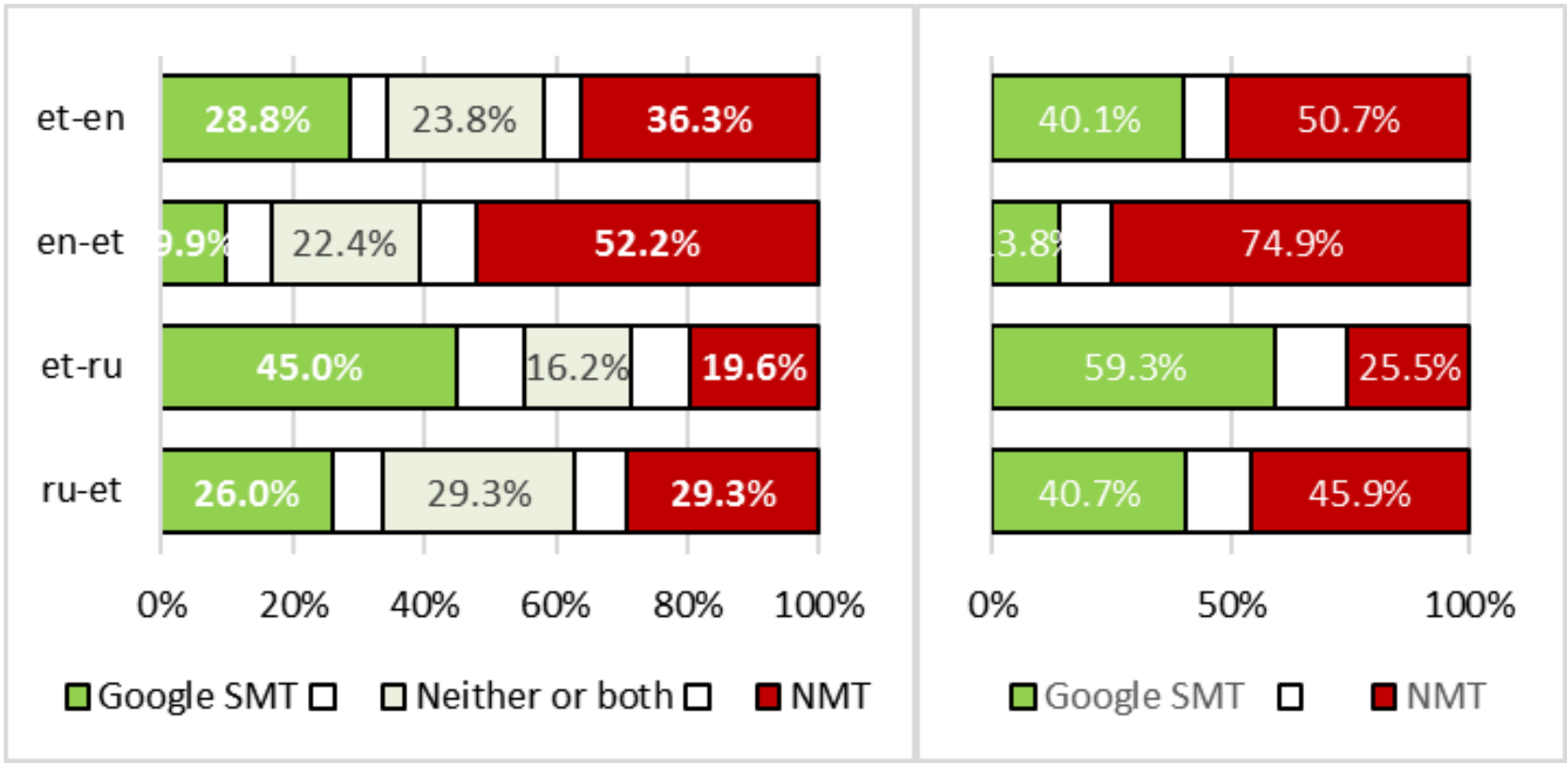
Language pair	System	BLEU
en-et	Baseline SMT	22.53 (20.39-24.95)
	Google Translate (SMT)	19.80 (18.00-21.60)
	NMT	**24.64** (22.76-26.54)
et-en	Baseline SMT	32.52 (30.55-34.53)
	Google Translate (SMT)	**40.57** (38.48-42.84)
	NMT	31.74 (29.91-33.45)
ru-et	Baseline SMT	09.87 (08.73-11.01)
	Google Translate (SMT)	**12.52** (11.03-14.01)
	NMT	09.02 (08.02-10.00)
et-ru	Baseline SMT	07.94 (07.07-08.82)
	Google Translate (SMT)	**14.74** (13.18-16.15)
	NMT	09.39 (08.33-10.46)
en-lv	Baseline SMT	32.57 (29.96-35.33)
	translate.tilde.com (SMT)	**37.54** (34.65-40.50)
	NMT	24.77 (22.94-26.72)
lv-en	Baselone SMT	28.79 (26.84-30.82)
	translate.tilde.com (SMT)	**43.76** (41.25-46.45)
	NMT	29.62 (27.62-31.44)

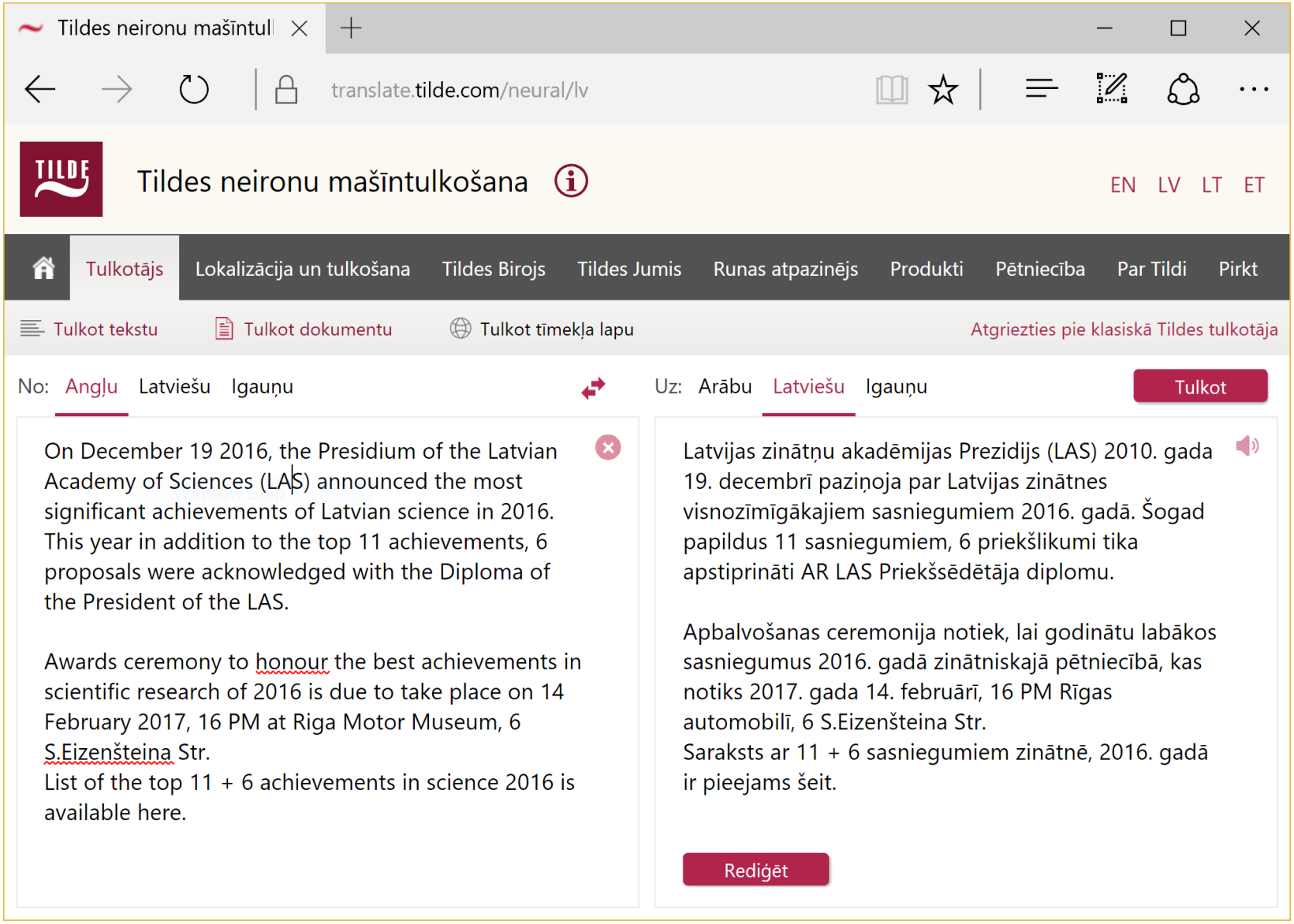
Human evaluation (system comparison)
SMT vs Neural MT





Human evaluation (system comparison)
Google Translate vs Neural MT




Tildes neironu mašīntul
translate.tilde.com/neural/lv
Tildes neironu mašīntulkošana
EN LV LT ET
Tulkotājs
Lokalizācija un tulkošana
Tildes Birojs
Tildes Jumis
Runas atpazinējs
Produkti
Pētniecība
Par Tildi
Pirkt
Tulkot tekstu
Tulkot dokumentu
Tulkot tīmekļa lapu
Atgriezties pie klasiskā Tildes tulkotāja
No: Angļu Latviešu Igauņu
Uz: Arābu Latviešu Igauņu
Tulkot
On December 19 2016, the Presidium of the Latvian Academy of Sciences (LAS) announced the most significant achievements of Latvian science in 2016. This year in addition to the top 11 achievements, 6 proposals were acknowledged with the Diploma of the President of the LAS.
Awards ceremony to honour the best achievements in scientific research of 2016 is due to take place on 14 February 2017, 16 PM at Riga Motor Museum, 6 S.Eizenšteina Str.
List of the top 11 + 6 achievements in science 2016 is available here.
Latvijas zinātņu akadēmijas Prezidijs (LAS) 2010. gada 19. decembrī paziņoja par Latvijas zinātnes visnozīmīgākajiem sasniegumiem 2016. gadā. Šogad papildus 11 sasniegumiem, 6 priekšlikumi tika apstiprināti AR LAS Priekšsēdētāja diplomu.
Apbalvošanas ceremonija notiek, lai godinātu labākos sasniegumus 2016. gadā zinātniskajā pētniecībā, kas notiks 2017. gada 14. februārī, 16 PM Rīgas automobilī, 6 S.Eizenšteina Str.
Saraksts ar 11 + 6 sasniegumiem zinātnē, 2016. gadā ir pieejams šeit.
Rediģēt

First Conclusions

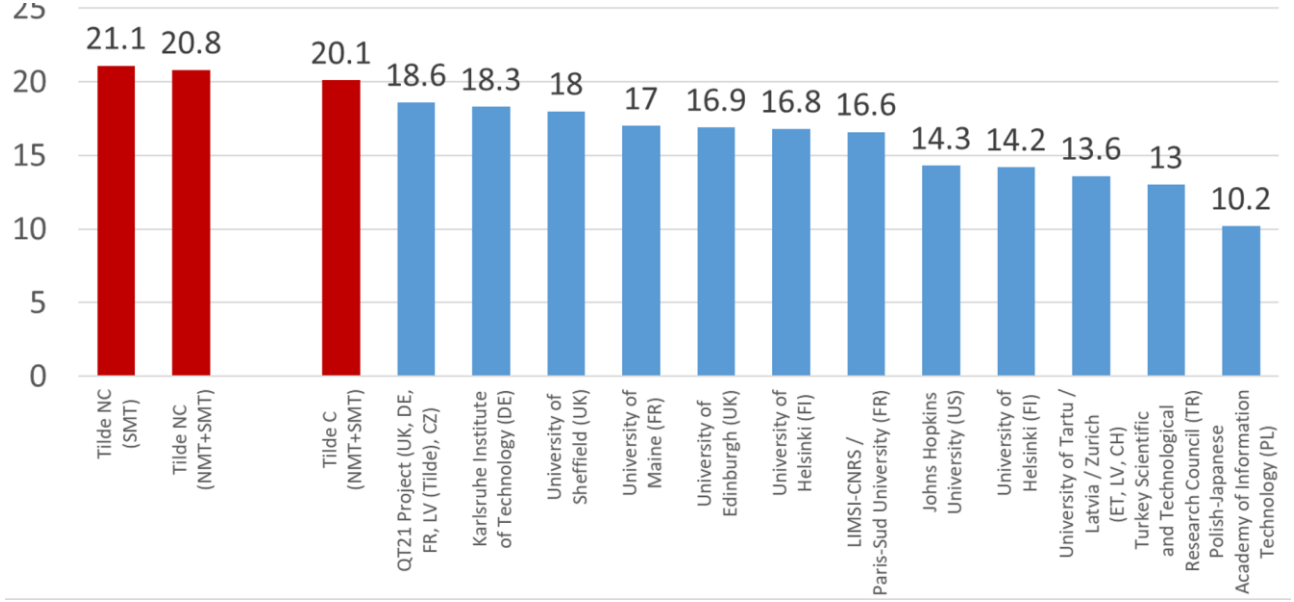
- In most cases Neural MT outperforms Statistical MT in human evaluation. It is true also for under-resourced languages like Latvian and Estonian

- Fluency is much better, word agreement is better, translates even unseen words
  but can hide semantic errors

- It is not a panacea, it is a field for new research and development

WMT 2017 Competition

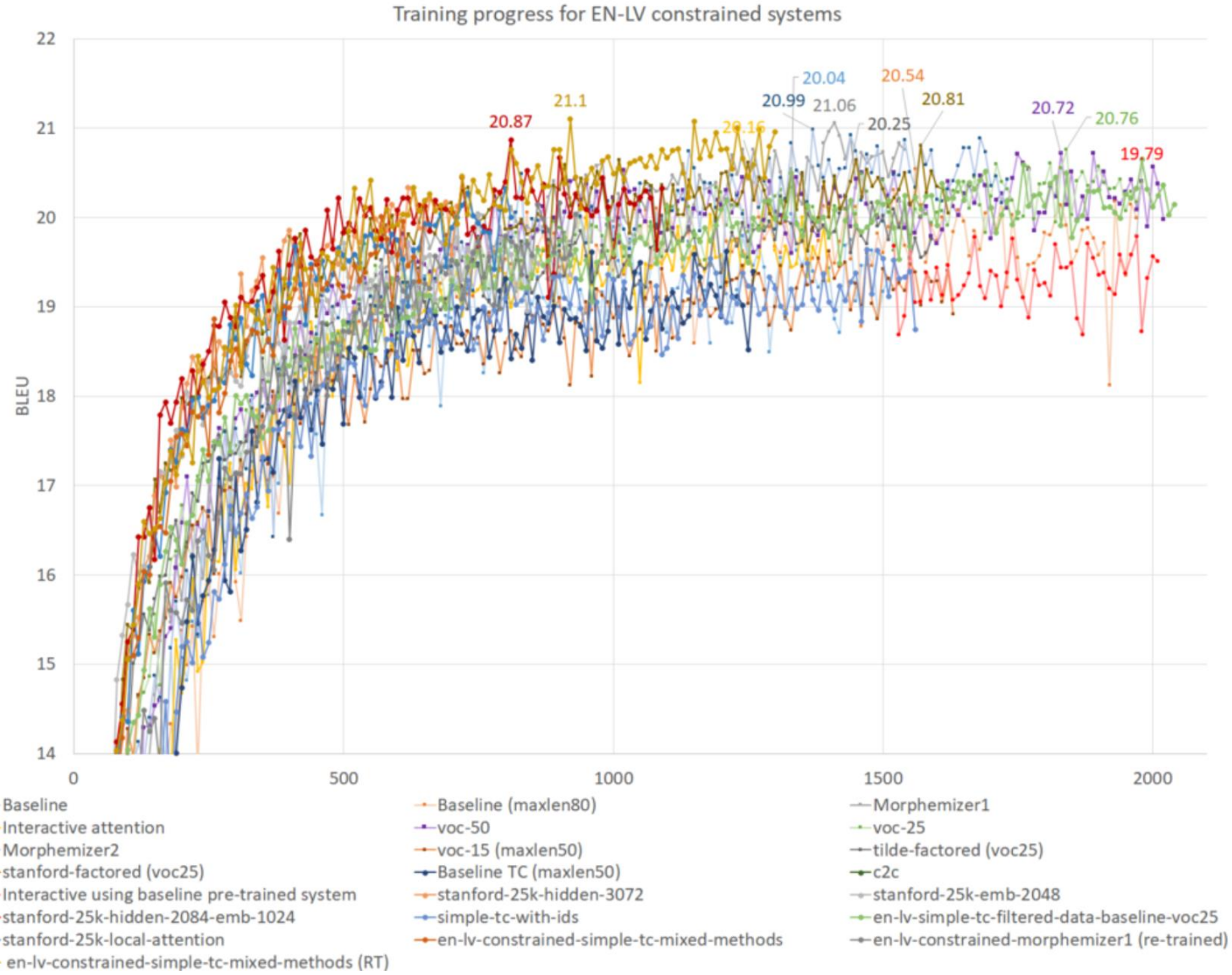
- Yearly competition of MT researchers
- Latvian – first time this year
- Both human and automatic evaluation



The winning system

- Nematus based NMT system
- Main improvements
  - data preprocessing and cleaning
  - special handling of numbers, ID etc. and rare words
  - hybrid with SMT
  - morphology aware sub-word units
  - factored NMT
  - back-translation of monolingual target language data
  - MLSTM recurrent neural network

  - A lot of experiments with different configurations (~ 55 trained NMT systems)


Training progress for EN-LV constrained systems
BLEU
22
21
20
19
18
17
16
15
14
0
500
1000
1500
2000
20.87
21.1
20.04
20.99
21.06
20.54
20.25
20.81
20.72
20.76
19.79
Baseline
Baseline (maxlen80)
Morphemizer1
Interactive attention
voc-50
voc-25
Morphemizer2
voc-15 (maxlen50)
tilde-factored (voc25)
stanford-factored (voc25)
Baseline TC (maxlen50)
c2c
Interactive using baseline pre-trained system
stanford-25k-hidden-3072
stanford-25k-emb-2048
stanford-25k-hidden-2084-emb-1024
simple-tc-with-ids
en-lv-simple-tc-filtered-data-baseline-voc25
stanford-25k-local-attention
en-lv-constrained-simple-tc-mixed-methods
en-lv-constrained-morphemizer1 (re-trained)
en-lv-constrained-simple-tc-mixed-methods (RT)

Tilde's Machine Translation Systems for WMT 2017

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics
Tilde, Vienības gatve 75A, Riga, Latvia
{marcis.pinnis,rihards.krislauks,daiga.deksne,toms.miks,valters.sics}@tilde.lv

NATIONAL DEVELOPMENT PLAN 2020
EUROPEAN UNION
European Regional Development Fund
INVESTING IN YOUR FUTURE

Data Filtering

Due to noise in the training data, we performed data filtering.

1. Long sentence filter *(>1500 symbols or >80 tokens)*
2. Sentence length ratio filter *(>0.3)*
3. Incorrect language filter *(using a language detection tool)*
4. Low content overlap filter *(using a cross-lingual alignment tool)*
5. Bad encoding filter *(e.g., corrupt symbols)*
6. Digit mismatch filter

Unique parallel/monolingual sentence counts

Scenario	Lang. pair	Before filtering	After filtering
Constrained	en-lv	1.92 M / 28.81 M	1.61 M / 27.75 M
	lv-en	1.92 M / 335.55 M	1.61 M / 330.23 M
Unconstrained	en-lv	15.78 M / 87.60 M	12.69 M / 81.68 M
	lv-en	15.78 M / 360.01 M	12.69 M / 351.99 M

Data Pre-processing

1. Normalisation of punctuation
2. Identification of non-translatable entities
3. Tokenisation *(Tilde's regular expression-based tokeniser)*
4. Truecasing *(first word only)*
5. Morphology-driven word splitting
6. Factorisation *(Tilde's perceptron tagger for Latvian, Stanford parser for English)*

Synthetic Data

Unknown Words as Placeholders

To make NMT models more robust to rare and unknown phenomena, we supplement the training data with sentence pairs where one to three content words are replaced with unknown word (UNK) placeholders.

Back-translation of Monolingual Data

For domain adaptation, we use a synthetic parallel corpus that is acquired by back-translation of in-domain monolingual data from the target language using a reverse NMT system.

Scenario	Lang. pair	<UNK> placeholder sent.	Back-translated sent.	Total
Constrained	en-lv	1.48 M	3.09 M	6.19 M
	lv-en	1.48 M	3.09 M	6.19 M
Unconstrained	en-lv	11.66 M	21.69 M	46.04 M
	lv-en	11.66 M	21.36 M	45.71 M

Introduction

We present Tilde's WMT 2017 MT systems that were ranked as **the best performing systems by automatic evaluation**.

Machine Translation Systems

- **SMT Systems**—Moses phrase-based systems, fast-align word alignment, 7-gram translation models, 5-gram KenLM language models, trained on the Tilde MT platform.
- **NMT Systems**—Nematus NMT systems with **MLSTM recurrent units, morphology-driven word splitting**, vocabulary size of 25,000 for constrained systems and 50,000 for unconstrained systems, decoding beam size of 12, ensembles of 5 to 7 models, back-translated data used to train final systems.
- **NMT-SMT hybrid systems**—rare words (e.g., person names, location names, different scripts, etc.) are replaced with unknown word place-holders, sentences are translated with NMT systems, after which rare words are translated with SMT systems. In unconstrained systems, a named entity data base is used to improve person name translation quality.

Example of the NMT-SMT Hybrid Translation Process

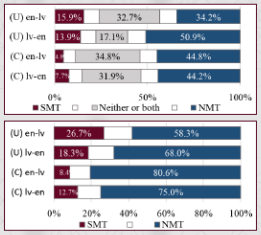
Translation step	Example sentence
Source text	Šodien skatieties **Ikauniecēs-Admidiņas** startu Rio spēlēs.
Pre-processed text	Šodien skat@@ ieties **I@@ kaun@@ iec@@ ēs - Ad@@ mi@@ di@@ ņas** start@@ u Rio spēlē@@ s .
Text with identified rare words	Šodien skat@@ ieties **ϐIDϐ - ϐIDϐ** start@@ u Rio spēlē@@ s .
NMT translation	watch the **ϐIDϐ - ϐIDϐ** start at the Rio Games today .
Moses XML with untranslated rare words	<nmt translation="watch the"> šodien skatieties </nmt> **Ikauniecēs** <nmt translation="-"> -</nmt> **Admidiņas** <nmt translation="start at the Rio Games today"> šodien startu Rio spēlēs</nmt> <nmt translation="."> .</nmt>
Moses XML with identified un-translated person names	<nmt translation="watch the"> šodien skatieties </nmt> **<ne translation="Ikauniece" prob="1.0"> Ikauniecēs</ne> <nmt translation="-"> -</nmt> <ne translation="Admidina||Admidins" prob="0.95||0.05"> Admidiņas</ne>** <nmt translation="start at the Rio Games today"> šodien startu Rio spēlēs</nmt> <nmt translation="."> .</nmt>
SMT translation	watch the **Ikauniece - Admidina** start at the Rio Games today .
Post-processed translation	Watch the **Ikauniece-Admidina** start at the Rio Games today.
NMT only transl. (for comparison)	Today, look at the start of the **Isolence-Admidias** in the Rio Games.

Evaluation

Automatic evaluation (submitted systems are marked in bold)

Scenario	Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER
Constrained	en-lv	SMT	12.98±0.62	0.5086	0.6642
		NMT	†19.49±0.79	0.5478	0.5877
		Hybrid	**†19.52±0.82**	**0.5482**	**0.5853**
	lv-en	SMT	15.47±0.59	0.5219	0.6606
		NMT	†20.01±0.67	0.5494	0.6088
		Hybrid	**†20.06±0.63**	**0.5496**	**0.6081**
Unconstrained	en-lv	**SMT**	**20.43±0.86**	0.5491	0.6126
		NMT	20.04±0.78	0.5563	0.5832
		Hybrid	20.08±0.78	**0.5567**	**0.5827**
	lv-en	SMT	19.05±0.63	0.5515	0.6233
		NMT	†22.02±0.63	0.5677	0.5838
		Hybrid	**†22.06±0.66**	0.5683	**0.5833**

Manual (comparative) **evaluation**



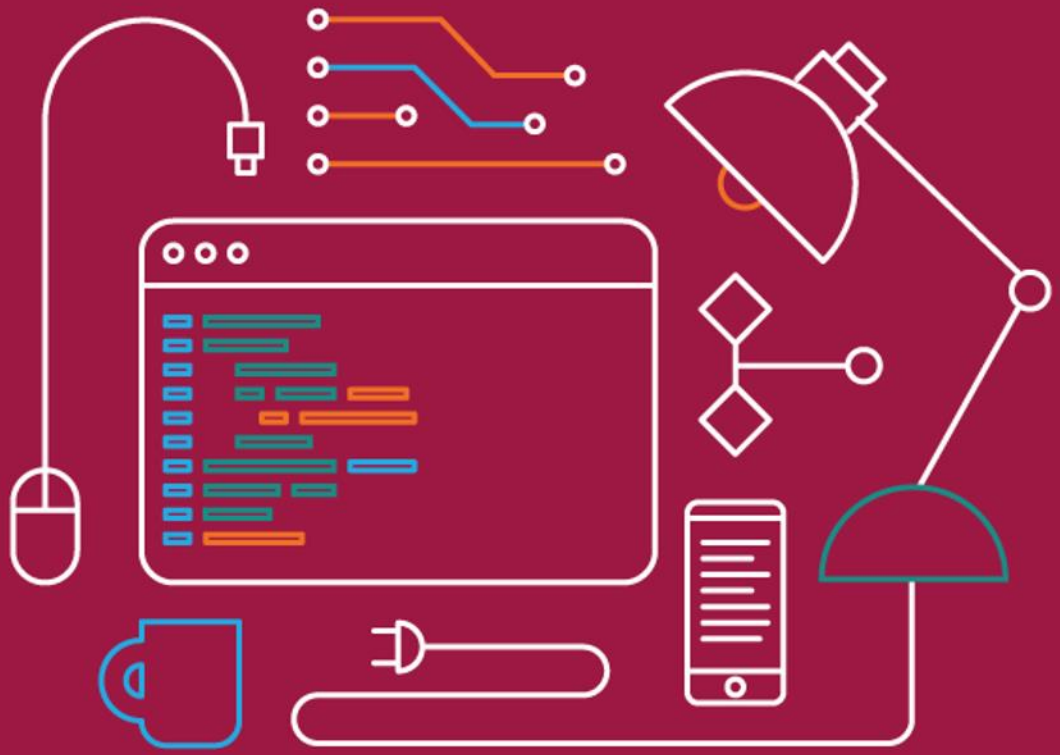
Comparison of the best constrained system submissions

Lang. pair	System	BLEU (CS)	BEER 2.0	CharacTER
en-lv	Tilde (hybrid)	**†19.52±0.79**	0.5482	0.5853
	QT21 combination	18.03±0.71	0.5403	0.6455
	KIT primary	17.72±0.69	0.5428	0.6051
lv-en	Tilde (hybrid)	**†20.06±0.65**	**0.5496**	**0.6081**
	UEDIN NMT	19.08±0.65	0.5462	0.6308
	JHU SMT	16.95±0.60	0.5281	0.6485

Get the poster in PDF:

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

- (Pinnis et al., 2017)



Thank you!

Questions, Discussions

References

- Deksne, D., & Skadiņšš, R. (2011). CFG Based Grammar Checker for Latvian. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011* (p. 275 278). Riga.
- Deksne, D., Skadiņa, I., & Skadiņš, R. (2014). Extended CFG Formalism for Grammar Checker and Parser Development. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing, 15th International Conference, CICLing 2014, Proceedings, Part I* (pp. 237–249). Kathmandu, Nepal: Springer. http://doi.org/10.1007/978-3-642-54906-9
- Pinnis, M., Krišlauks, R., Miks, T., Deksne, D., Šics, V. (2017). Tilde's Machine Translation Systems for WMT 2017.