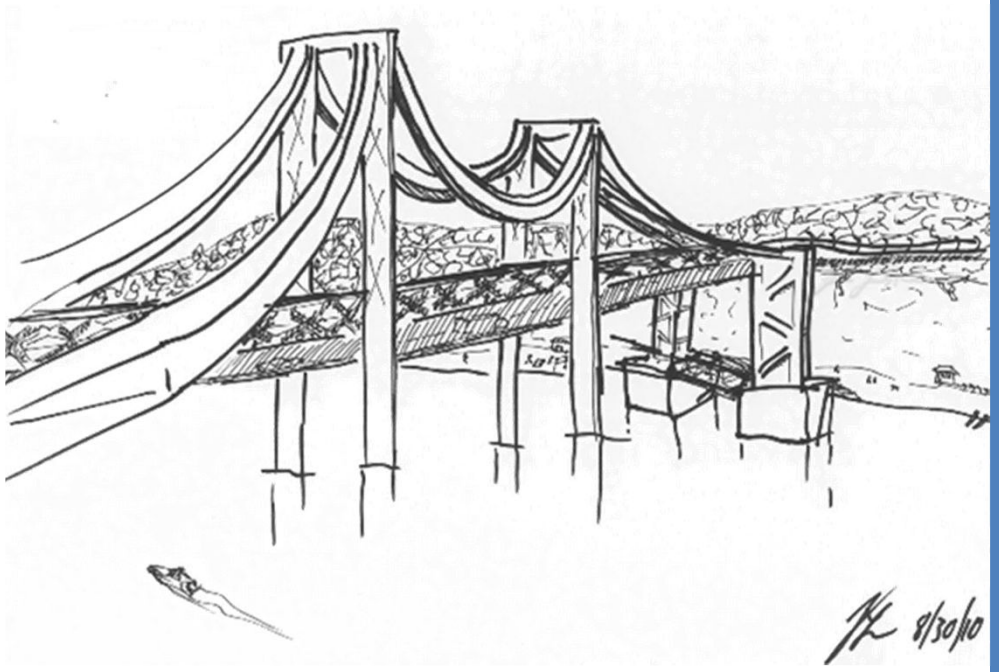
technologies
for
smaller
languages

andrejs vasiljevs
andrejs@tilde.com

- Language technology developer
- Localization service provider
- Leadership in smaller languages
- Offices in Riga (Latvia), Tallinn (Estonia) and Vilnius (Lithuania)
- 120 employees
- Strong R&D team
- 9 PhDs and candidates



statistical machine translation

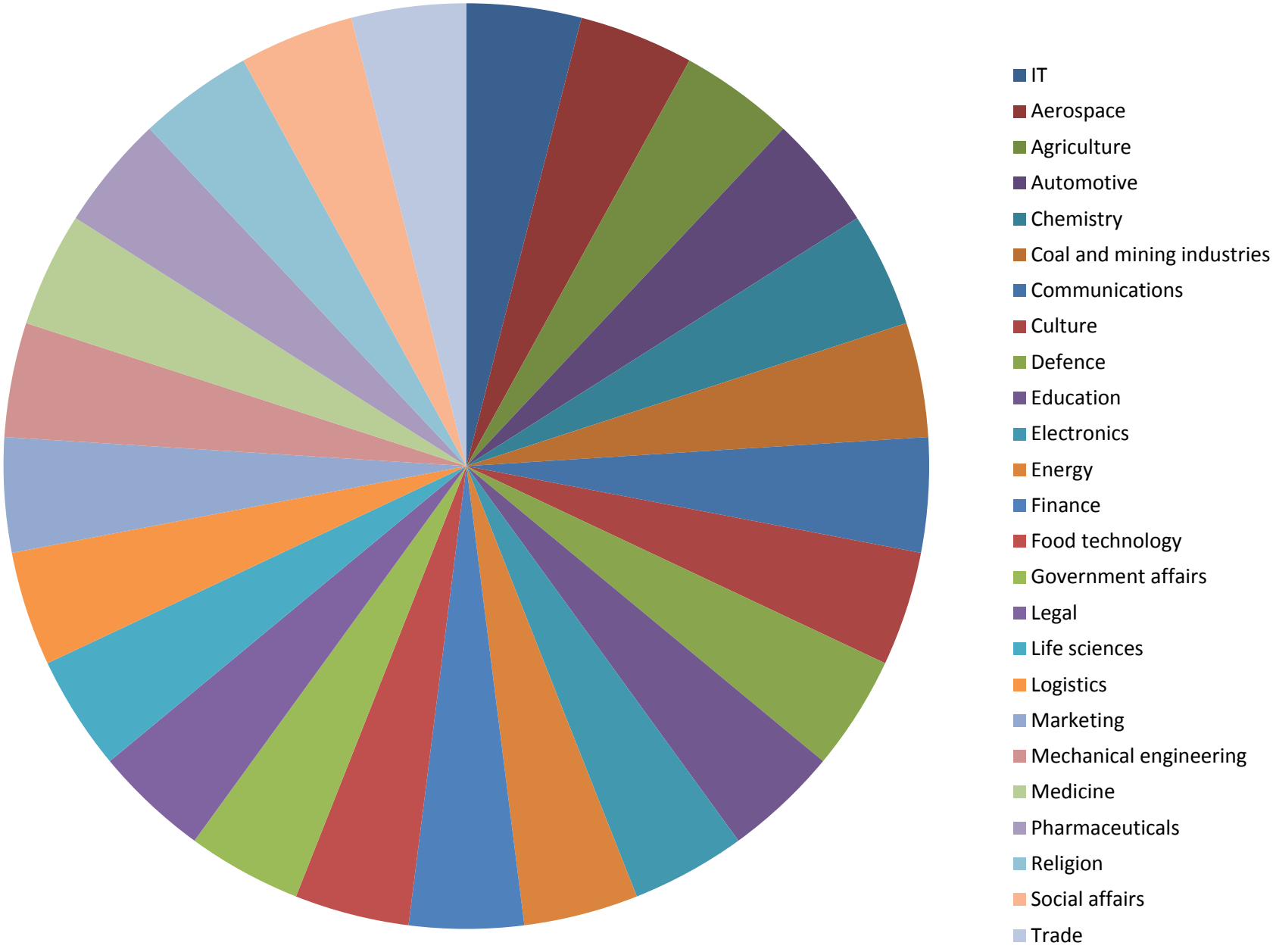
bridging the language barriers

making multi-lingual web accessible


DATA

15 largest languages

50%


IT
Aerospace
Agriculture
Automotive
Chemistry
Coal and mining industries
Communications
Culture
Defence
Education
Electronics
Energy
Finance
Food technology
Government affairs
Legal
Life sciences
Logistics
Marketing
Mechanical engineering
Medicine
Pharmaceuticals
Religion
Social affairs
Trade

domains

how to tailor MT for smaller languages and special domains

?

the new paradigm for user-tailored machine translation

```
[ttable-file]
0 0 5 /.../unfactored/model/phrase-table.0-0.gz
% ls steps/1/LM_toy_tokenize.1* | cat
steps/1/LM_toy_tokenize.1
steps/1/LM_toy_tokenize.1.DONE
steps/1/LM_toy_tokenize.1.INFO
steps/1/LM_toy_tokenize.1.STDERR
steps/1/LM_toy_tokenize.1.STDERR.digest
steps/1/LM_toy_tokenize.1.STDOUT
% train-model.perl \
--corpus factored-corpus/proj-syndicate \
--root-dir unfactored \
--f de --e en \
--lm 0:3:factored-corpus/surface.lm:0
% moses -f moses.ini -lmodel-file "0 0 3
../lm/europarl.srilm.gz"
use-berkeley = true
alignment-symmetrization-method = berkeley
berkeley-train = $moses-script-
dir/ems/support/berkeley-train.sh
berkeley-process = $moses-script-
dir/ems/support/berkeley-process.sh
berkeley-jar = /your/path/to/berkeleyaligner-
2.1/berkeleyaligner.jar
berkeley-java-options = "-server -mx30000m -ea"
berkeley-training-options = "-Main.iters 5 5 -
EMWordAligner.numThreads 8"
berkeley-process-options = "-
EMWordAligner.numThreads 8"
berkeley-posterior = 0.5
tokenize
in: raw-stem
out: tokenized-stem
default-name: corpus/tok
pass-unless: input-tokenizer output-tokenizer
template-if: input-tokenizer IN.$input-
extension OUT.$input-extension
template-if: output-tokenizer IN.$output-
extension OUT.$output-extension
parallelizable: yes
working-dir = /home/pkoehn/experiment
```

LetsMT!
takes the
complexity out of
machine translation

load & share data

click to build

use immediately

Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

EASY

- ✔ Frees you from hardware and software infrastructure
- ✔ Store corpora and engines in one place
- ✔ Access to many free public texts and translation engines
- ✔ Instantly increase productivity with CAT plug-in for localization process and web browser widget

IS LETS MT FOR YOU?

- ✔ Localization & translation service providers
- ✔ Holders of linguistic resources
- ✔ Companies with a need to translate large amounts of information
- ✔ Those who don't want to trust their resources to public systems

THE BASICS

Translate

Translate now using available systems

Build

Build your own machine translation system

Store and Share

Upload, organize, store, and share your corpora

Currently on Let's MT

- 11 trained systems
- 97 languages
- 10 public corpora
- 495 million parallel sentences in public corpora

News

- LetsMT! presented at ICT-PSP info week in Zagreb
- LetsMT! partner Moravia cooperates with M4Loc
- LetsMT! presented at SlaviCorp2010: Corpora of Slavic Languages
- LetsMT! presented at the first META-NET Forum
- LetsMT! 2010 Annual Public Report available

Project partners

- Tilde
- University of Edinsburgh
- University of Zagreb
- University of Copenhagen
- Uppsala University
- Moravia

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

EASY

- ✔ Frees you from hardware and software infrastructure
- ✔ Store corpora and engines in one place
- ✔ Access to many free public texts and translation engines
- ✔ Instantly increase productivity with CAT plug-in for localization process and web browser widget

IS LETS MT FOR YOU?

- ✔ Localization & translation service providers
- ✔ Holders of linguistic resources
- ✔ Companies with a need to translate large amounts of information
- ✔ Those who don't want to trust their resources to public systems

THE BASICS

Translate

Translate now using available systems

Build

Build your own machine translation system

Store and Share

Upload, organize, store, and share your corpora

Currently on Let's MT

- 11 trained systems
- 97 languages
- 10 public corpora
- 495 million parallel sentences in public corpora

News

- LetsMT! presented at ICT-PSP info week in Zagreb
- LetsMT! partner Moravia cooperates with M4Loc
- LetsMT! presented at SlaviCorp2010: Corpora of Slavic Languages
- LetsMT! presented at the first META-NET Forum
- LetsMT! 2010 Annual Public Report available

Project partners

- Tilde
- University of Edinsburgh
- University of Zagreb
- University of Copenhagen
- Uppsala University
- Moravia

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

Data for SMT training

Let's MT! language statistics of Top 20 languages

Data is represented in million (M) or thousand (k) of sentences per language or language pair.

Total parallel size: **1 178 081 012** sentences, total parallel count: **234**.

	Mono	en	es	pt	nl	lt	it	cz	da	lv	sv	hu	et	fi	fr	pl	sk	de	sl	el	cs	mt	ro	bg
en	313.8M	<1k	23.4M	20.7M	16.8M	16.3M	15.9M	15M	14.3M	13.8M	13.5M	11.1M	10.7M	10.2M	10.2M	8.9M	8M	7.9M	7.6M	6.1M	5.6M	4.6M	4.3M	3.5I
es	89.8M	23.4M		5.4M	6.6M	5M	6.4M		6.2M	5M	6.4M	5M	5M	6.4M	6.7M	5M	4.9M	6.4M	5M	5.2M	4.4M	2.7M	2.4M	2.4I
pt	77.3M	20.7M	5.4M		5M	5M	6.6M		6.2M	4.9M	4.9M	4.9M	4.9M	6.1M	7.2M	3.7M	3.8M	6.5M	3.8M	5.2M	4.7M	2.7M	2M	2.6I
nl	75.9M	16.8M	6.6M	5M		4.7M	6.5M		6.3M	4.7M	4.9M	4.7M	4.7M	6.1M	6.7M	4.7M	3.6M	6.4M	3.5M	5.2M	4.5M	2.4M	1.9M	2.4I
lt	87.6M	16.3M	5M	5M	4.7M		5.1M		4.7M	6.3M	4.7M	5.3M	5.3M	4.7M	5.9M	5.3M	5.2M	5.1M	5.2M	4.4M	5M	3M	2.6M	2.2I
it	82.1M	15.9M	6.4M	6.6M	6.5M	5.1M			6.1M	5M	6.3M	4.7M	4.7M	6M	6.7M	5.1M	5M	6.4M	5M	5.1M	4.5M	2.8M	2.5M	2.4I
cz	15M	15M																						
da	74.4M	14.3M	6.2M	6.2M	6.3M	4.7M	6.1M		<1k	4.6M	6.1M	4.7M	3.5M	4.8M	6.2M	4.7M	4.6M	6.4M	4.6M	5.1M	4.5M	2.4M	2.2M	1.8I
lv	85.5M	13.8M	5M	4.9M	4.7M	6.3M	5M		4.6M		4.6M	5.3M	5.3M	4.7M	5.8M	5.2M	5.1M	5.1M	5.1M	4.4M	4.9M	2.9M	2.6M	2.7I
sv	69.8M	13.5M	6.4M	4.9M	4.9M	4.7M	6.3M		6.1M	4.6M		4.7M	4.7M	6.1M	6.6M	3.6M	3.6M	6.3M	3.6M	5.1M	4.5M	2.5M	1.9M	2.4I
hu	62M	11.1M	5M	4.9M	4.7M	5.3M	4.7M		4.7M	5.3M	4.7M		5.2M	4.7M	5.9M	5.2M	5.1M	3.9M	5.2M	4.4M	4.9M	2.9M	2.6M	2.7I
et	60.7M	10.7M	5M	4.9M	4.7M	5.3M	4.7M		3.5M	5.3M	4.7M	5.2M		4.7M	5.1M	5.3M	5.1M	5.1M	5.2M	4.4M	4.9M	2.9M	2.6M	2.7I
fi	71.5M	10.2M	6.4M	6.1M	6.1M	4.7M	6M		4.8M	4.7M	6.1M	4.7M	4.7M		6.2M	4.7M	4.6M	6.4M	4.6M	5.1M	4.4M	2.4M	2.2M	2.4I
fr	81.8M	10.2M	6.7M	7.2M	6.7M	5.9M	6.7M		6.2M	5.8M	6.6M	5.9M	5.1M	6.2M		5.9M	5.7M	6.7M	5.8M	5.2M	4.8M	3.6M	3M	2.7I
pl	56.6M	8.9M	5M	3.7M	4.7M	5.3M	5.1M		4.7M	5.2M	3.6M	5.2M	5.3M	4.7M	5.9M		4.1M	5.1M	4M	4.4M	5M	3M	2.2M	2.7I
sk	53M	8M	4.9M	3.8M	3.6M	5.2M	5M		4.6M	5.1M	3.6M	5.1M	5.1M	4.6M	5.7M	4.1M		5M	4.1M	4.3M	5M	2.9M	2.2M	2.7I
de	74.3M	7.9M	6.4M	6.5M	6.4M	5.1M	6.4M		6.4M	5.1M	6.3M	3.9M	5.1M	6.4M	6.7M	5.1M	5M		5M	5.3M	4.5M	2.9M	2.5M	1.9I
sl	53.3M	7.6M	5M	3.8M	3.5M	5.2M	5M		4.6M	5.1M	3.6M	5.2M	5.2M	4.6M	5.8M	4M	4.1M	5M		4.3M	5M	3M	2.2M	2.7I
el	59.8M	6.1M	5.2M	5.2M	5.2M	4.4M	5.1M		5.1M	4.4M	5.1M	4.4M	4.4M	5.1M	5.2M	4.4M	4.3M	5.3M	4.3M		4.1M	2.2M	2.1M	2.4I
cs	52.5M	5.6M	4.4M	4.7M	4.5M	5M	4.5M		4.5M	4.9M	4.5M	4.9M	4.9M	4.4M	4.8M	5M	5M	4.5M	5M	4.1M		2.7M	2.5M	2.6I
mt	40M	4.6M	2.7M	2.7M	2.4M	3M	2.8M		2.4M	2.9M	2.5M	2.9M	2.9M	2.4M	3.6M	3M	2.9M	2.9M	3M	2.2M	2.7M		2.3M	2.4I
ro	34.7M	4.3M	2.4M	2M	1.9M	2.6M	2.5M		2.2M	2.6M	1.9M	2.6M	2.6M	2.2M	3M	2.2M	2.2M	2.5M	2.2M	2.1M	2.5M	2.3M		2.4I
bg	36.6M	3.5M	2.4M	2.6M	2.4M	2.2M	2.4M		1.8M	2.7M	2.4M	2.7M	2.7M	2.4M	2.7M	2.7M	2.7M	1.9M	2.7M	2.4M	2.6M	2.4M	2.4M	

Let's MT!

Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

EASY

- ✔ Frees you from hardware and software infrastructure
- ✔ Store corpora and engines in one place
- ✔ Access to many free public texts and translation engines
- ✔ Instantly increase productivity with CAT plug-in for localization process and web browser widget

IS LETS MT FOR YOU?

- ✔ Localization & translation service providers
- ✔ Holders of linguistic resources
- ✔ Companies with a need to translate large amounts of information
- ✔ Those who don't want to trust their resources to public systems

THE BASICS

Translate

Translate now using available systems

Build

Build your own machine translation system

Store and Share

Upload, organize, store, and share your corpora

Currently on Let's MT

- 11 trained systems
- 97 languages
- 10 public corpora
- 495 million parallel sentences in public corpora

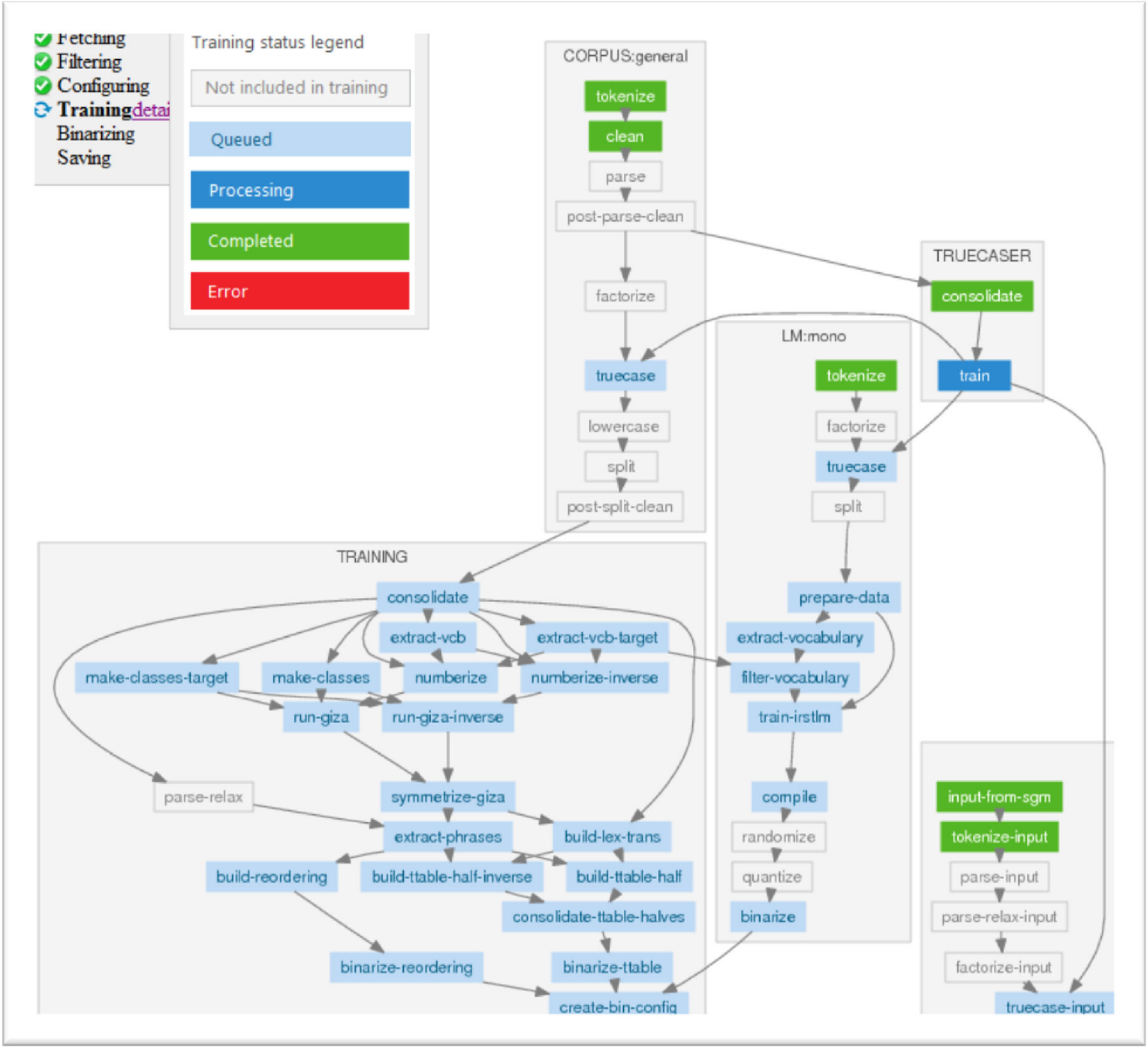
News

- LetsMT! presented at ICT-PSP info week in Zagreb
- LetsMT! partner Moravia cooperates with M4Loc
- LetsMT! presented at SlaviCorp2010: Corpora of Slavic Languages
- LetsMT! presented at the first META-NET Forum
- LetsMT! 2010 Annual Public Report available

Project partners

- Tilde
- University of Edinsburgh
- University of Zagreb
- University of Copenhagen
- Uppsala University
- Moravia

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more


Fetching
Filtering
Configuring
Training detai
Binarizing
Saving
Training status legend
Not included in training
Queued
Processing
Completed
Error
CORPUS:general
tokenize
clean
parse
post-parse-clean
factorize
truecase
lowercase
split
post-split-clean
TRUECASER
consolidate
train
LM:mono
tokenize
factorize
truecase
split
prepare-data
extract-vocabulary
filter-vocabulary
train-irstlm
compile
randomize
quantize
binarize
TRAINING
consolidate
extract-vcb
extract-vcb-target
make-classes-target
make-classes
numberize
numberize-inverse
run-giza
run-giza-inverse
parse-relax
symmetrize-giza
extract-phrases
build-lex-trans
build-reordering
build-ttable-half-inverse
build-ttable-half
consolidate-ttable-halves
binarize-reordering
binarize-ttable
create-bin-config
input-from-sgm
tokenize-input
parse-input
parse-relax-input
factorize-input
truecase-input

Build your own machine translation system!

With LetsMT you can easily build and run your own custom machine translation systems. Simply upload your own corpora and/or choose to use any of the publicly available corpora. Train your systems and use them for all your translation needs.

EASY

- ✔ Frees you from hardware and software infrastructure
- ✔ Store corpora and engines in one place
- ✔ Access to many free public texts and translation engines
- ✔ Instantly increase productivity with CAT plug-in for localization process and web browser widget

IS LETS MT FOR YOU?

- ✔ Localization & translation service providers
- ✔ Holders of linguistic resources
- ✔ Companies with a need to translate large amounts of information
- ✔ Those who don't want to trust their resources to public systems

THE BASICS

Translate

Translate now using available systems

Build

Build your own machine translation system

Store and Share

Upload, organize, store, and share your corpora

Currently on Let's MT

- 11 trained systems
- 97 languages
- 10 public corpora
- 495 million parallel sentences in public corpora

News

- LetsMT! presented at ICT-PSP info week in Zagreb
- LetsMT! partner Moravia cooperates with M4Loc
- LetsMT! presented at SlaviCorp2010: Corpora of Slavic Languages
- LetsMT! presented at the first META-NET Forum
- LetsMT! 2010 Annual Public Report available

Project partners

- Tilde
- University of Edinsburgh
- University of Zagreb
- University of Copenhagen
- Uppsala University
- Moravia

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

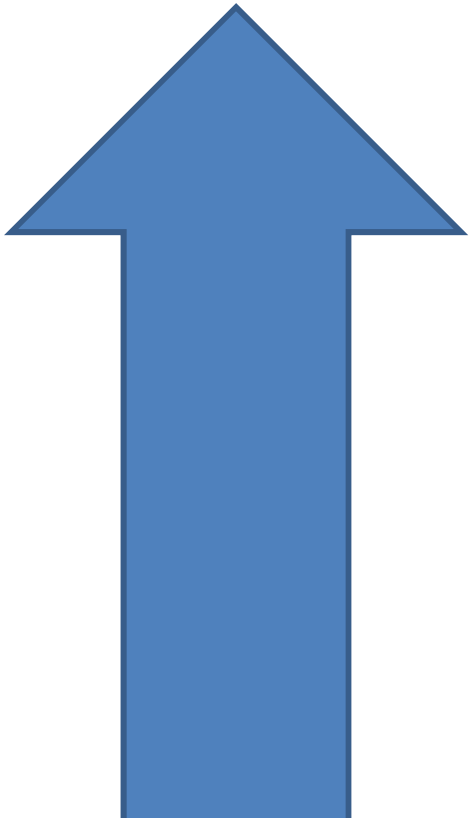
integration in translation tools

technology

- Moses - open source MT platform
- highly customized for performance, robustness and scalability
- cloud-based infrastructure
- technologies for processing and alignment of multilingual texts
- language specific modules
- APIs for integration into applications

24%-

32.9%*



productivity

* Skadiņš R., Puriņš M., Skadiņa I., Vasiļjevs A., *Evaluation of SMT in localization to under-resourced inflected language*, in Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, p. 35-40, May 30-31, 2011, Leuven, Belgium

language service market

31.44

billion USD

market

quality

surpasses

Google Translate*

**generic and domain specific systems for Baltic languages provides better quality in both automatic (BLEU) and human evaluation*

human users

professional translator

international organizations

universities & researchers

casual users

machine users

dictionary & translation tools

e.g. Tildes Birojs

mobile translators

e.g. Moravia and Tilde mobile apps

content analytics

e.g. SemLab newssentiment.eu

website translation tools

e.g. UFZG browser widget

user testimonies

The customized LetsMT! engine provided better results within our scenario in comparison with the results achieved using Google Translate

The estimated productivity increase was about 25% when compared with translating without MT

The engine output suggested the correct and required terminology – this has been found as the most useful benefit

industry & research collaboration

Tilde / Coordinator
LATVIA

University of Edinburgh
UK

Uppsala University
SWEDEN

Copehagen University
DENMARK

University of Zagreb
CROATIA

Moravia
CZECH REPUBLIC

SemLab
NETHERLANDS

build your own
MT engine
!

TILDE