Improving SMT with Morphology Knowledge for Baltic Languages

Click to edit Master subtitle style

Raivis SKADIŅŠa,b, Kārlis GOBAa and Valters ŠICS a
a *Tilde SIA, Latvia*
b *University of Latvia, Latvia*

Machine Translation and Morphologically-Rich Languages
University of Haifa, Israel, 23-27 January, 2011

Outline

- The exotic Baltic languages
- Problems we all know about
  - What Philipp & others said
  - Agreement, reordering
  - Sparseness
  - Alignment and evaluation
  - Morphology integration
- (Not) using factor models
  - English-Latvian
  - Lithuanian-English
- Interlude: Human evaluation

Baltic languages

	PIE	Latvian	Lithuanian
Nom.	pod-s	pēd-a	pėd-a
Voc.	pod	pēd-a	pėd-a
Acc.	pod-m	pēd-u	pėd-ą
Instr.	ped-eh	pēd-u	pėd-a
Dat.	ped-ey	pēd-ai	pėd-ai
Abl.	ped-es		
Gen.	ped-es	pēd-as	pėd-os
Loc.	ped-(i)	pēd-ā	pėd-oje

- >2000 morphosyntactic tags, >1000 observed
- Derivation vs inflection
- Similar problems as with Czech, German, etc.
  - Case/gender/number/definiteness
  - Quite elaborated participle system
  - Feature redundancy

Disambiguation

Gulf of Bothnia
Finland
Aland (Finland)
Sweden
Gulf of Finland
Hiiumaa (Estonia)
Tallinn
Russian Federation
Saaremaa (Estonia)
Baltic Sea
Estonia
Gulf of Riga
Gotland (Sweden)
Latvia
Riga
Oland (Sweden)
EUROPE
The Baltic States
Lithuania
N
Kaliningrad Oblast Russian Federation
Vilnius
Poland
Belarus

Latv(ij|i|j)a|**Latf**ia|**Let**(land|[oó]nia)|**Lett**(land|ország|oni[ae])|**Lot**yšk[áo]|**Łot**wa|**Läti** |An **Lait**via

Fancy letters: ā ē ī ū š ž č ķ ģ ņ ļ

Lietuva|**Liett**ua|**Lith**([uw]ania|áen)|**Lit**(uani[ea]|[vw]a|vánia|vanija|(au|ouw)en)|An **Liot**uáin

Fancy letters: ė ę ą ų ū

SMT as a product

- Quality
  - BLEU useful for development, but not for the end user
  - Efficient human evaluation
- Hardware requirements
  - Decoding speed
  - Memory
- Not just bare translation
  - Correct casing
  - Domain-specific translation
  - User feedback

English-Latvian

- English: dependency parser
- Latvian: morphological analysis, disambiguated
- Add lemmas and morphological tags as factors
  - pēdas ⍰ pēda | Nfsn
  - Tags also mark agreement for prepositions
- Alternatively,
  - Split words into stems and suffixes
  - pēdas ⍰ pēd | as
  - More ambiguous
  - -u ⍰ Amsa, Ampg, Afsa, Afpg, ...

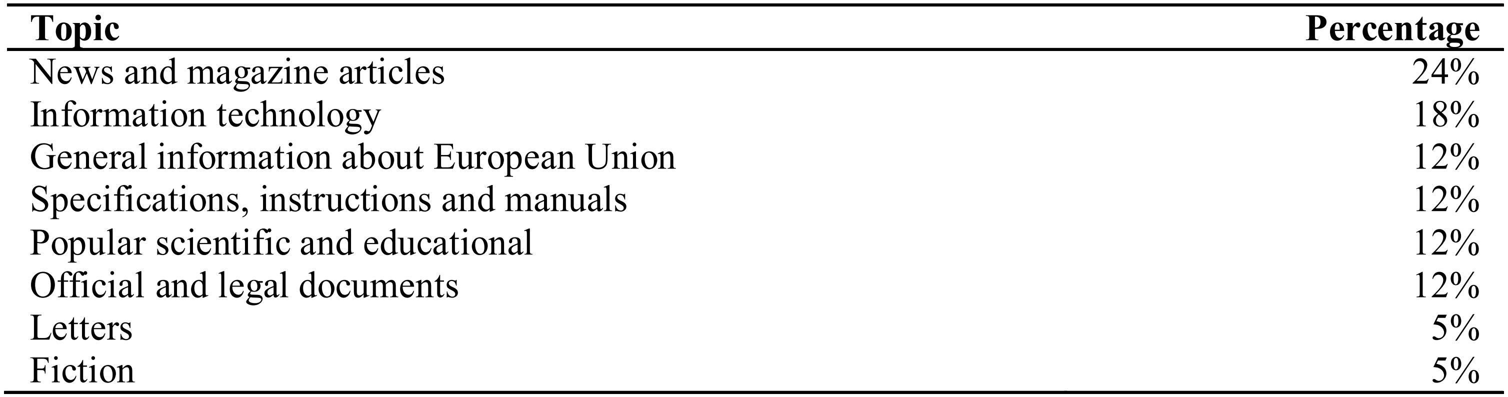
English-Latvian: Data

Bilingual corpus	**Parallel units**
Localization TM	1.29M
DGT-TM	1.06M
OPUS EMEA	0.97M
Fiction	0.66M
Dictionary data	0.51M
Total	**4.49M**
	Filtered: 3.23M

Monolingual corpus	**Words**
Parallel corpus	60M
News (web)	250M
Fiction	9M
Total	**319M**

Development data

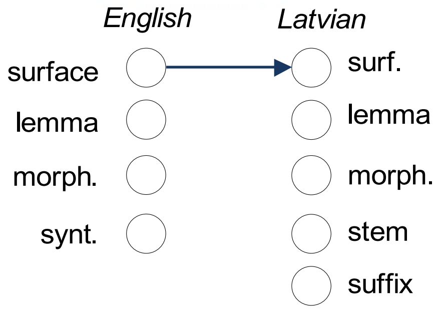
- Tuning and evaluation data
  - 1000 sentence tuning set + 512 sentence eval set
  - Domain/topic mixture



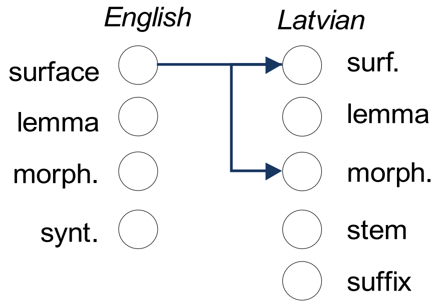
Topic	Percentage
News and magazine articles	24%
Information technology	18%
General information about European Union	12%
Specifications, instructions and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
Letters	5%
Fiction	5%

- Available in EN, LV, LT, ET, RO, SL, HR, DE, RU
  - ACCURAT project

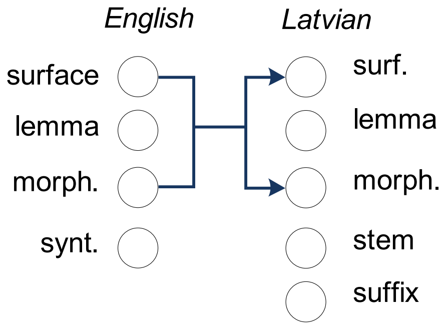
Factored translation



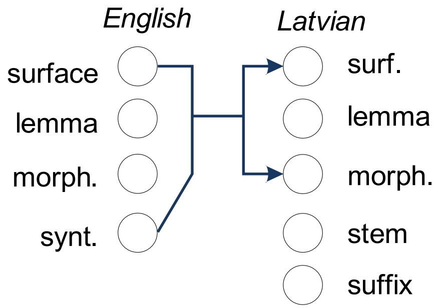
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



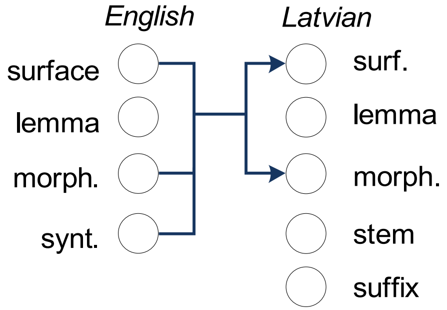
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



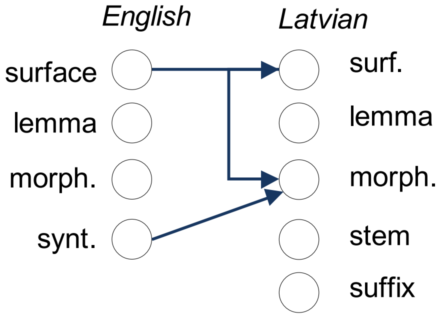
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



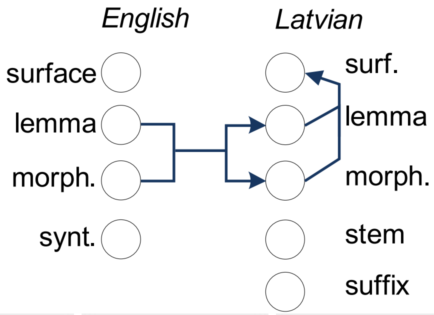
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



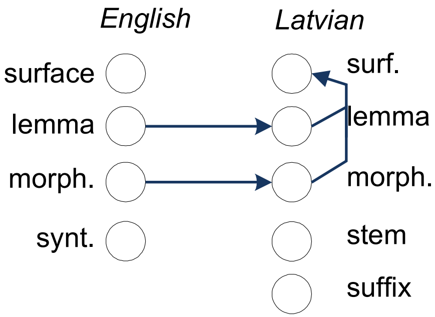
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



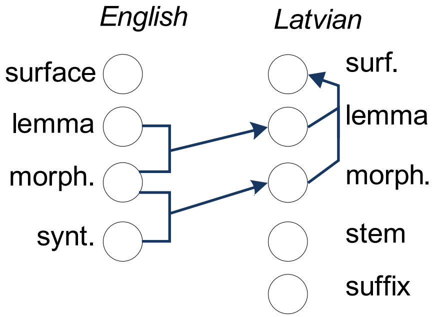
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix



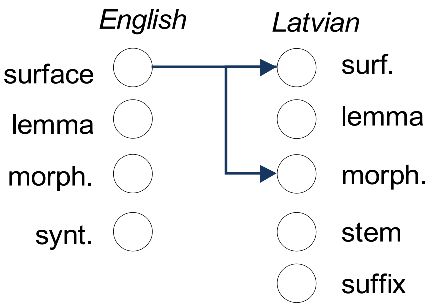
English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix

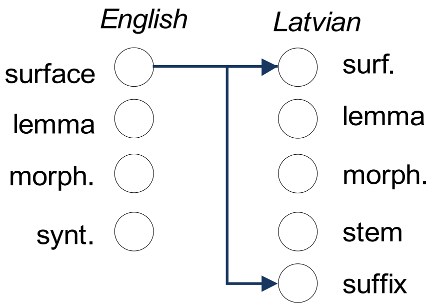


English
Latvian
surface
lemma
morph.
synt.
surf.
lemma
morph.
stem
suffix

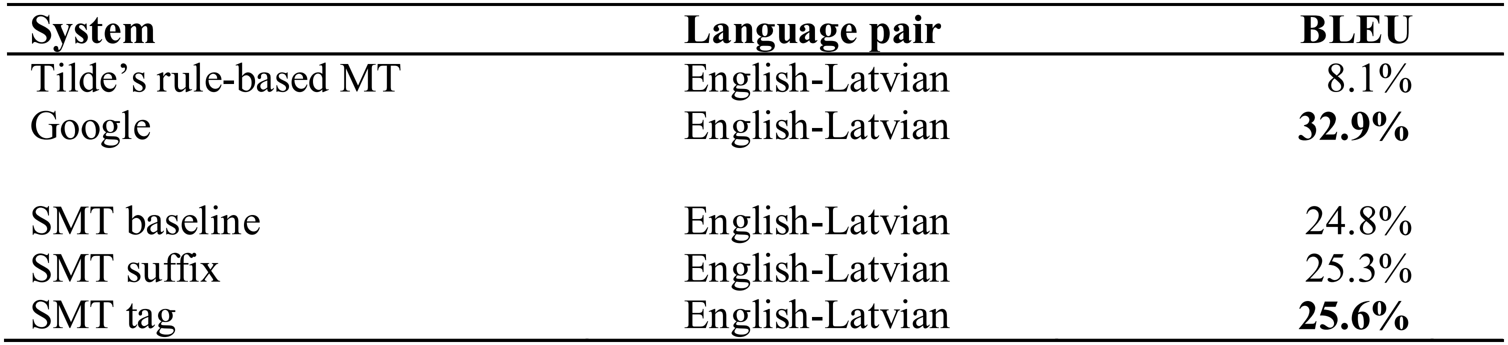
Final setup

- What Philipp said...
- 5-gram LM over surface forms
- 7-gram LM over morphology tags / suffixes





English-Latvian: Evaluation



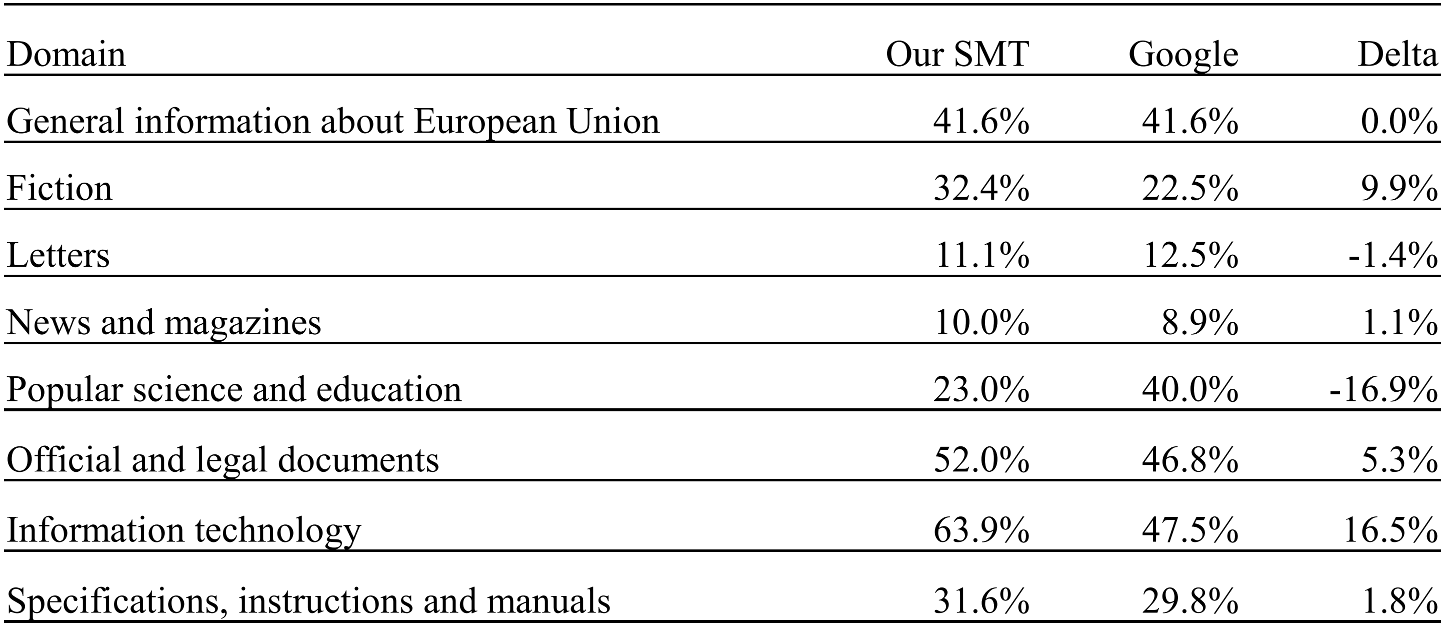
System	Language pair	BLEU
Tilde’s rule-based MT	English-Latvian	8.1%
Google	English-Latvian	**32.9%**
SMT baseline	English-Latvian	24.8%
SMT suffix	English-Latvian	25.3%
SMT tag	English-Latvian	**25.6%**

More data: Web scrapes

- Comparable corpora
- Heavily filtered
  - Alignment score
  - Length
  - Alphanumeric
  - Language detection
- 22.5M parallel units
- 0.9M left

System	**Language pair**	**BLEU**
SMT baseline	English-Latvian	33.0%
SMT w/ Web scrapes	English-Latvian	**35.0%**

Domain breakdown



Domain	Our SMT	Google	Delta
General information about European Union	41.6%	41.6%	0.0%
Fiction	32.4%	22.5%	9.9%
Letters	11.1%	12.5%	-1.4%
News and magazines	10.0%	8.9%	1.1%
Popular science and education	23.0%	40.0%	-16.9%
Official and legal documents	52.0%	46.8%	5.3%
Information technology	63.9%	47.5%	16.5%
Specifications, instructions and manuals	31.6%	29.8%	1.8%

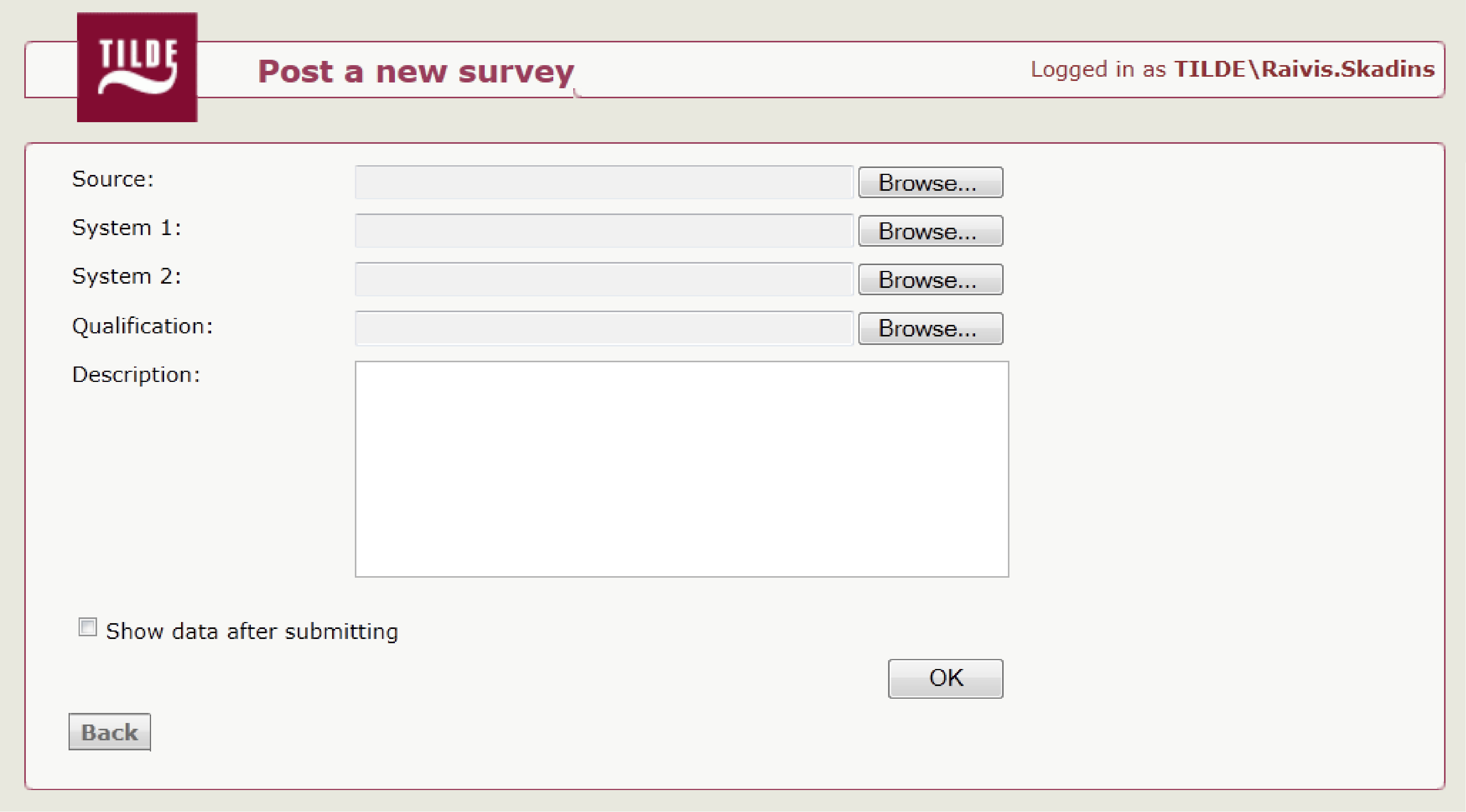
Human evaluation

- We got slightly better BLEU scores, **but is it really getting better?**
- Looking for practical methods simple, reliable, relatively cheap

Human evaluation

- Ranking of translated sentences relative to each other
  - Only 2 systems
  - The same 500 sentences as for BLEU

- Web based evaluation system
  - Upload source text and 2 target outputs
  - Send URL to many evaluators
  - See results



Post a new survey

Logged in as **TILDE\Raivis.Skadins**

Source:

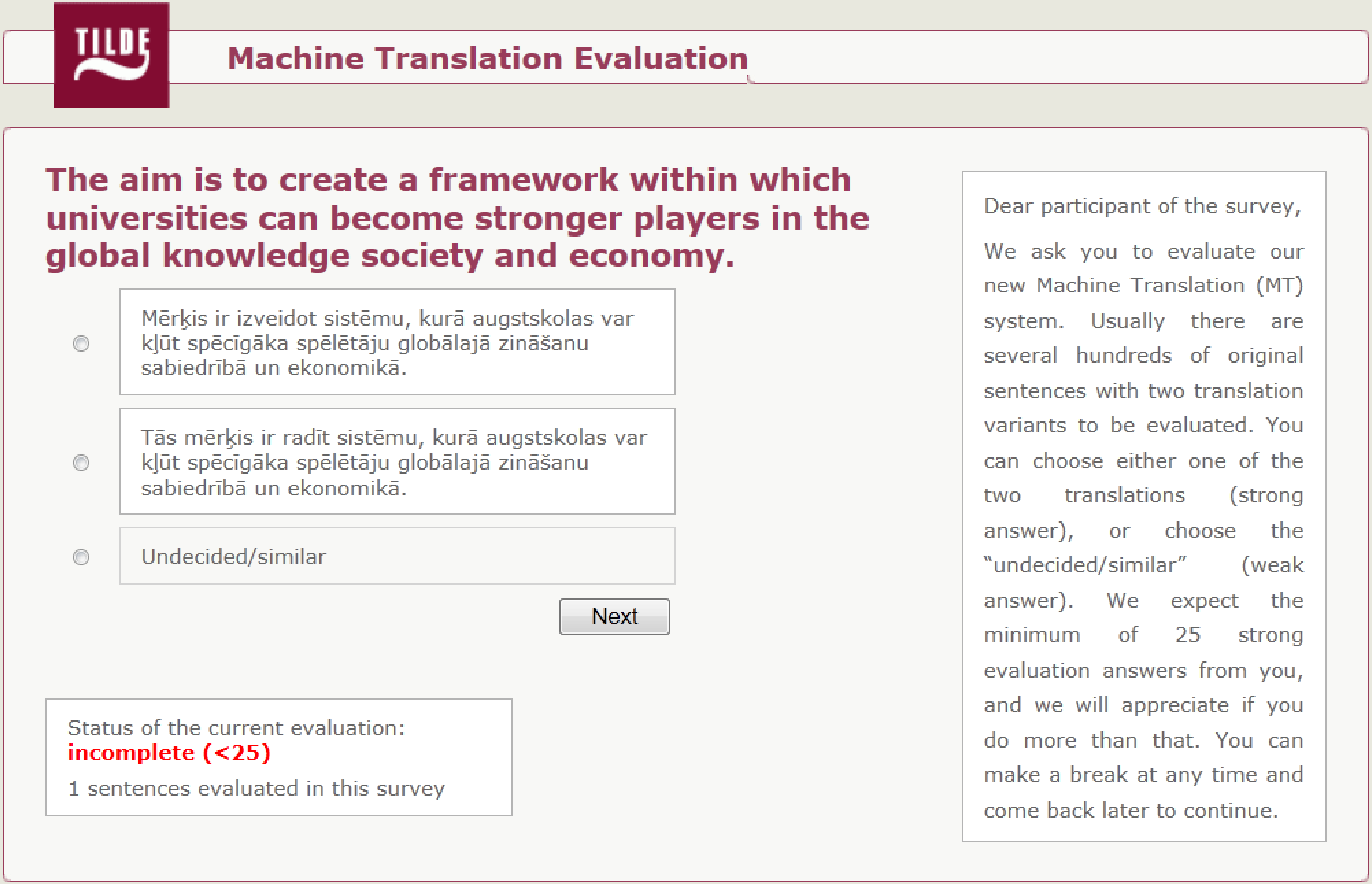
System 1:

System 2:

Qualification:

Description:

□ Show data after submitting



Machine Translation Evaluation

The aim is to create a framework within which universities can become stronger players in the global knowledge society and economy.

○ Mērķis ir izveidot sistēmu, kurā augstskolas var kļūt spēcīgāka spēlētāju globālajā zināšanu sabiedrībā un ekonomikā.

○ Tās mērķis ir radīt sistēmu, kurā augstskolas var kļūt spēcīgāka spēlētāju globālajā zināšanu sabiedrībā un ekonomikā.

○ Undecided/similar

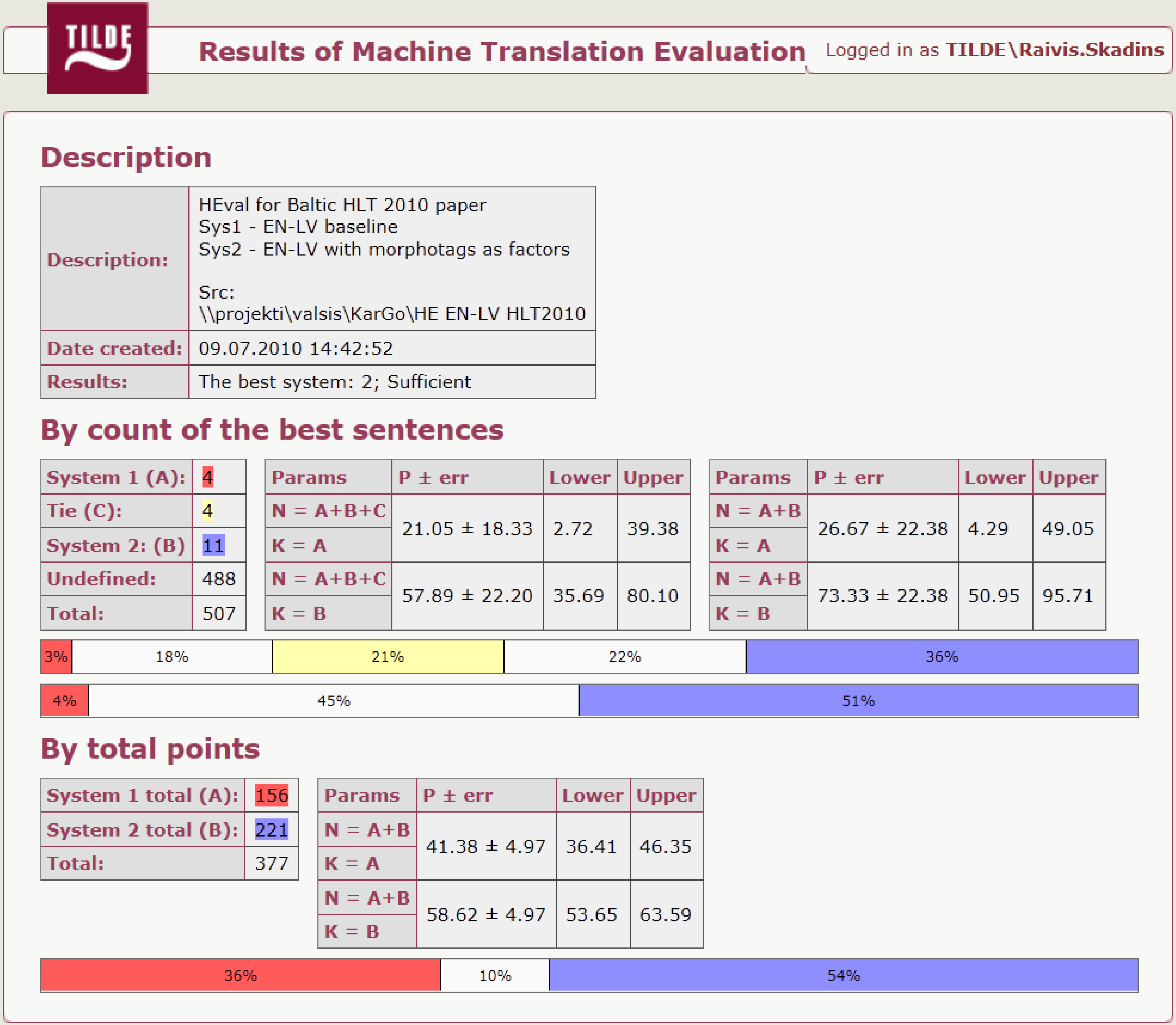
Status of the current evaluation:
incomplete (<25)
1 sentences evaluated in this survey

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.

Manual evaluation

- We calculate
  - Probability that system A is better than B: $p = \frac{A}{A+B} 100\%$
  - Confidence interval: $ci = z\sqrt{\frac{p(1-p)}{A+B}} 100\%$
- we calculate it
  - Based on all evaluations
  - Based on a sentence level



Results of Machine Translation Evaluation

Logged in as **TILDE\Raivis.Skadins**

Description

Description:	HEval for Baltic HLT 2010 paper Sys1 - EN-LV baseline Sys2 - EN-LV with morphotags as factors Src: \\projekti\valsis\KarGo\HE EN-LV HLT2010
Date created:	09.07.2010 14:42:52
Results:	The best system: 2; Sufficient

By count of the best sentences

System 1 (A):	4
Tie (C):	4
System 2: (B)	11
Undefined:	488
Total:	507

Params	P ± err	Lower	Upper

N = A+B+C
K = A
	21.05 ± 18.33	2.72	39.38

N = A+B+C
K = B
	57.89 ± 22.20	35.69	80.10

Params	P ± err	Lower	Upper

N = A+B
K = A
	26.67 ± 22.38	4.29	49.05

N = A+B
K = B
	73.33 ± 22.38	50.95	95.71

3%	18%	21%	22%	36%

4%	45%	51%

By total points

System 1 total (A):	156
System 2 total (B):	221
Total:	377

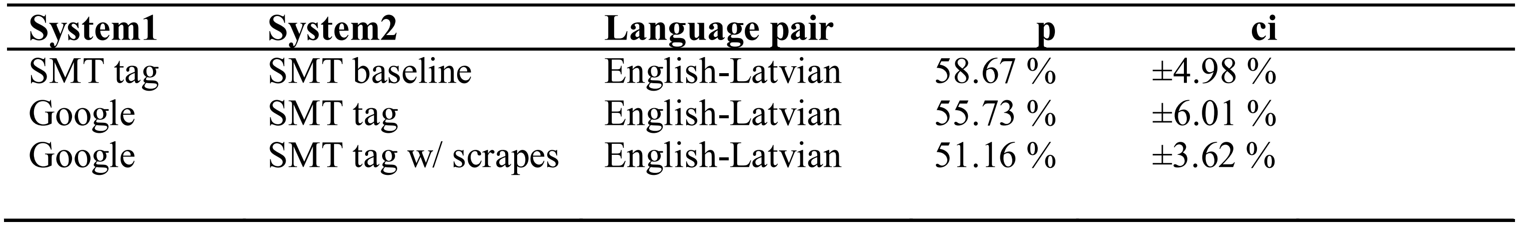
Params	P ± err	Lower	Upper

N = A+B
K = A
	41.38 ± 4.97	36.41	46.35

N = A+B
K = B
	58.62 ± 4.97	53.65	63.59

36%	10%	54%

English-Latvian: Human evaluation



System1	System2	Language pair	p	ci
SMT tag	SMT baseline	English-Latvian	58.67 %	±4.98 %
Google	SMT tag	English-Latvian	55.73 %	±6.01 %
Google	SMT tag w/ scrapes	English-Latvian	51.16 %	±3.62 %

Lithuanian-English

- No morphology tagger for Lithuanian
- Split
  - Stem and an optional suffix
  - Mark the suffix
  - vienas ⍰ vien #as
  - Suffixes correspond to endings, but are ambiguous
  - One would expect #ai ⍰ to/for
  - Prefixes – verb negation

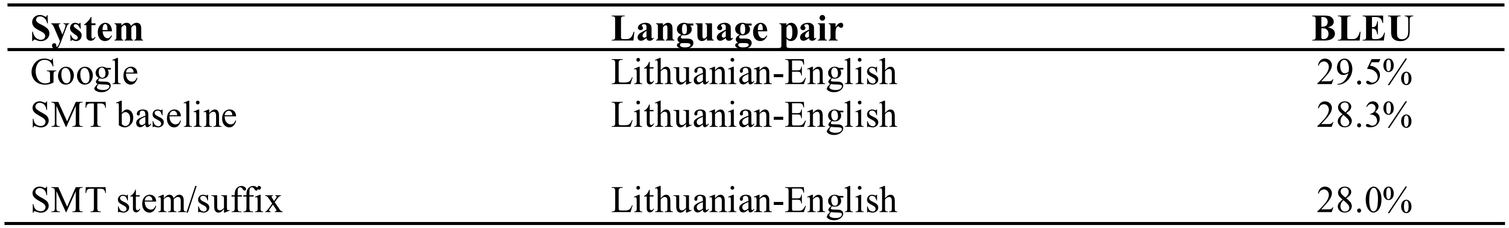
Lithuanian-English: Data

Bilingual corpus	Parallel units
Localization TM	1.56M
DGT-TM	0.99M
OPUS EMEA	0.84M
Dictionary data	0.38M
OPUS KDE4	0.05M
Total	**3.82M**
	Filtered: 2.71M

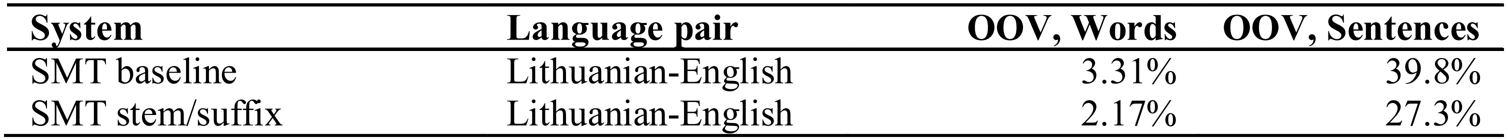
Monolingual corpus	Words
Parallel corpus	60M
News (WMT09)	440M
LCC	21M
Total	**521M**

Lithuanian-English

- Automatic evaluation

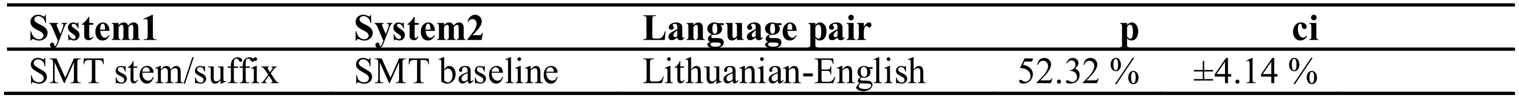


System	Language pair	BLEU
Google	Lithuanian-English	29.5%
SMT baseline	Lithuanian-English	28.3%
SMT stem/suffix	Lithuanian-English	28.0%



System	Language pair	OOV, Words	OOV, Sentences
SMT baseline	Lithuanian-English	3.31%	39.8%
SMT stem/suffix	Lithuanian-English	2.17%	27.3%

- Human evaluation



System1	System2	Language pair	p	ci
SMT stem/suffix	SMT baseline	Lithuanian-English	52.32 %	±4.14 %

Conclusions

- Translating to highly inflected language
  - Some success in predicting the right inflections by a LM
  - Things to try:
    - Two-step approach
    - Marking the relevant source features
- Translating from highly inflected language
  - Slight decrease in BLEU
  - Decrease in OOV rate
  - Human evaluation suggests users prefer lower OOV rate
  - Things to try:
    - Removing the irrelevant features