HOW TO GET MORE DATA FOR UNDER-RESOURCED LANGUAGES AND DOMAINS?

Andrejs Vasiļjevs

Tilde

FLaReNet Venice Forum

27.05.2011

WORLD SUMMIT ON THE INFORMATION SOCIETY PLAN OF ACTION

- Encourage the development of content and to put in place technical conditions in order to facilitate the presence and use of **all world languages** on the Internet;
- Governments, through public/private partnerships, should promote technologies and R&D programmes in such areas as translation, iconographies, voice-assisted services and the development of necessary hardware and a variety of software models, [..], electronic dictionaries, terminology and thesauri, multilingual search engines, machine translation tools, internationalized domain names, content referencing as well as general and application software.

MT AND THE FUTURE OF SMALLER LANGUAGES

- Survival of smaller languages depends on the outcome of the race between development of Machine Translation and proliferation of larger languages

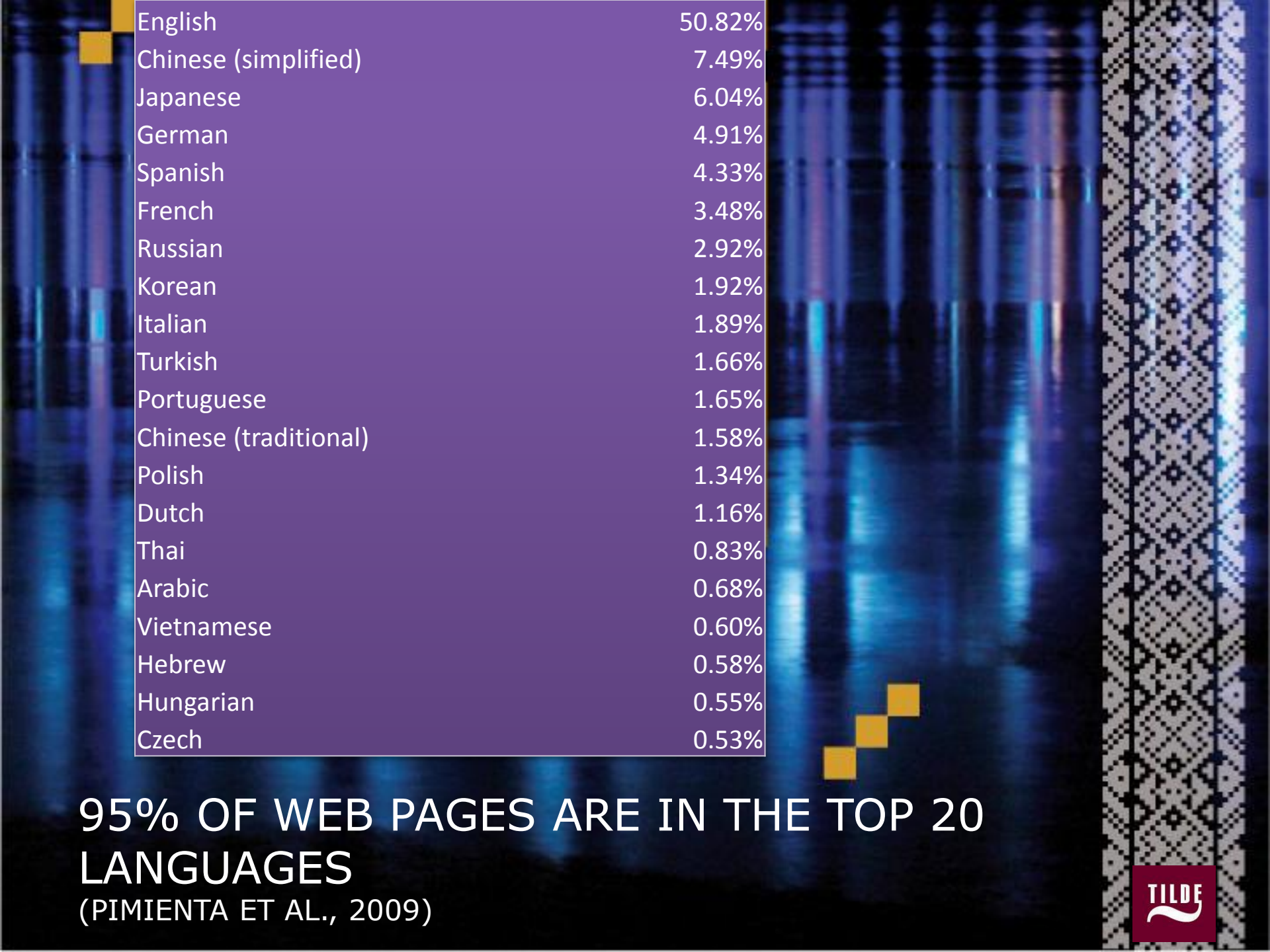
  (Alvin Toffler)

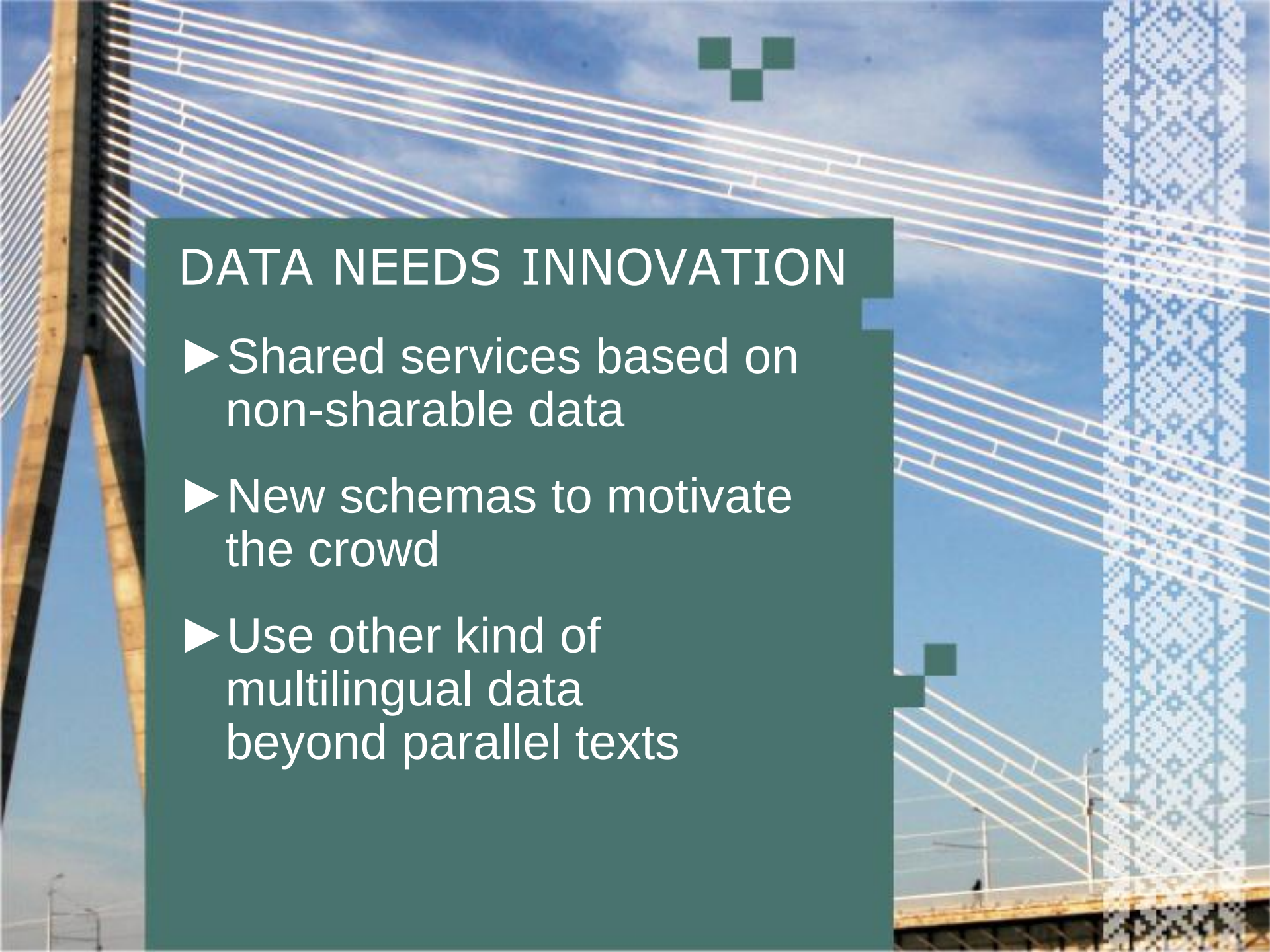
SMALLER LANGUAGES NEED MORE DATA

- Smaller languages often have a complex morphological structure and free word order.
- Much larger volumes of training data are needed to learn this complexity by statistical methods.

- *English:* tree, trees
- *Latvian:* koks, koku, kokam, kokā, koki, kokiem, kokos, kociņš, kociņu, kociņam, kociņā, kociņ, kociņi, kociņos, kociņiem

- *English:* run
- *Latvian:* skriet, skrien, skrēja, skries, skrienam, skrējām, skriesim, skrienat, skrieniet, skrējāt, skriesiet


English 50.82%
Chinese (simplified) 7.49%
Japanese 6.04%
German 4.91%
Spanish 4.33%
French 3.48%
Russian 2.92%
Korean 1.92%
Italian 1.89%
Turkish 1.66%
Portuguese 1.65%
Chinese (traditional) 1.58%
Polish 1.34%
Dutch 1.16%
Thai 0.83%
Arabic 0.68%
Vietnamese 0.60%
Hebrew 0.58%
Hungarian 0.55%
Czech 0.53%
95% OF WEB PAGES ARE IN THE TOP 20 LANGUAGES
(PIMIENTA ET AL., 2009)
TILDE



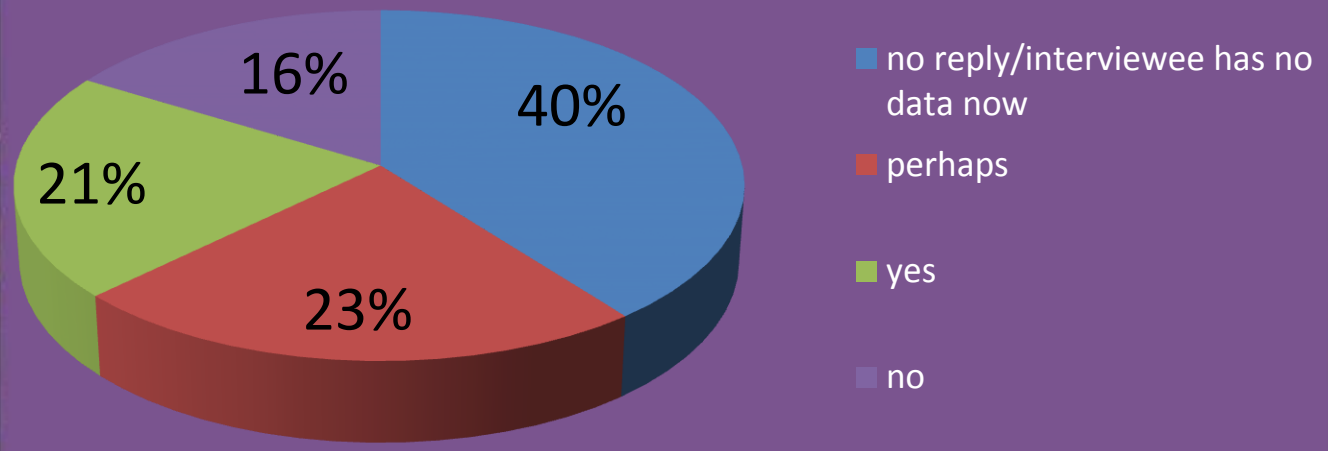
DATA NEEDS INNOVATION

- Shared services based on non-sharable data
- New schemas to motivate the crowd
- Use other kind of multilingual data beyond parallel texts

WHEN THE WEB IS NOT ENOUGH

- Parallel data accessible on the web is just a fraction of all translated texts
- Lot of parallel data reside in the local systems of corporations, public and private institutions,desktops of individual users
- Not all the data can be shared:
  - Confidentiality
  - Competitiveness
  - Personal data
- How to benefit from the non-shared data?

USER SURVEY: WILLINGNESS TO SHARE DATA



Let's MT!

- Put users in control of their data
- Fully public or fully private should not the only choice
- Data can be used for MT generation without exposing it
- Empower users to create custom MT engines from their data

SHARED SERVICES FROM NON-SHARABLE DATA

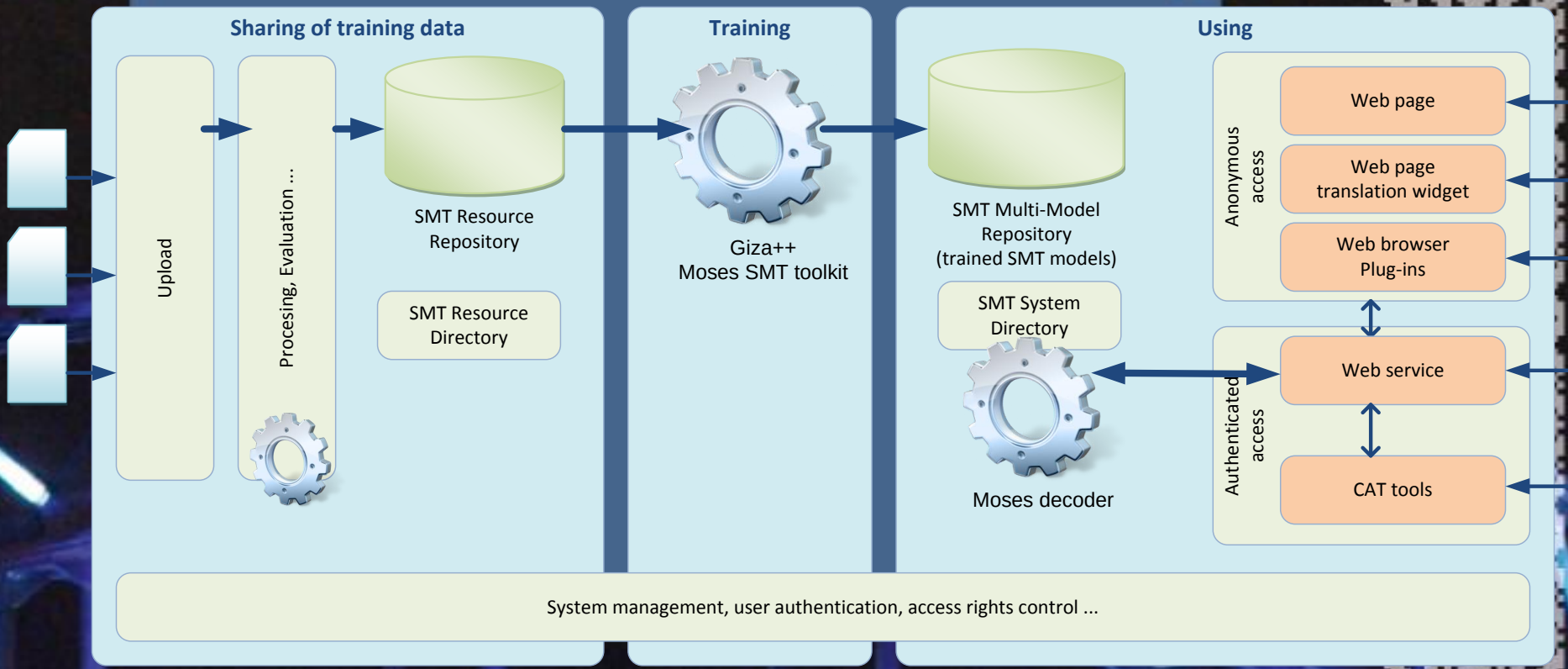
Let's MT!

- Sustainable user-driven MT factory on the cloud
- Services for data collection, MT generation, customization and running of variety of user-tailored MT systems.

LetsMT! Project

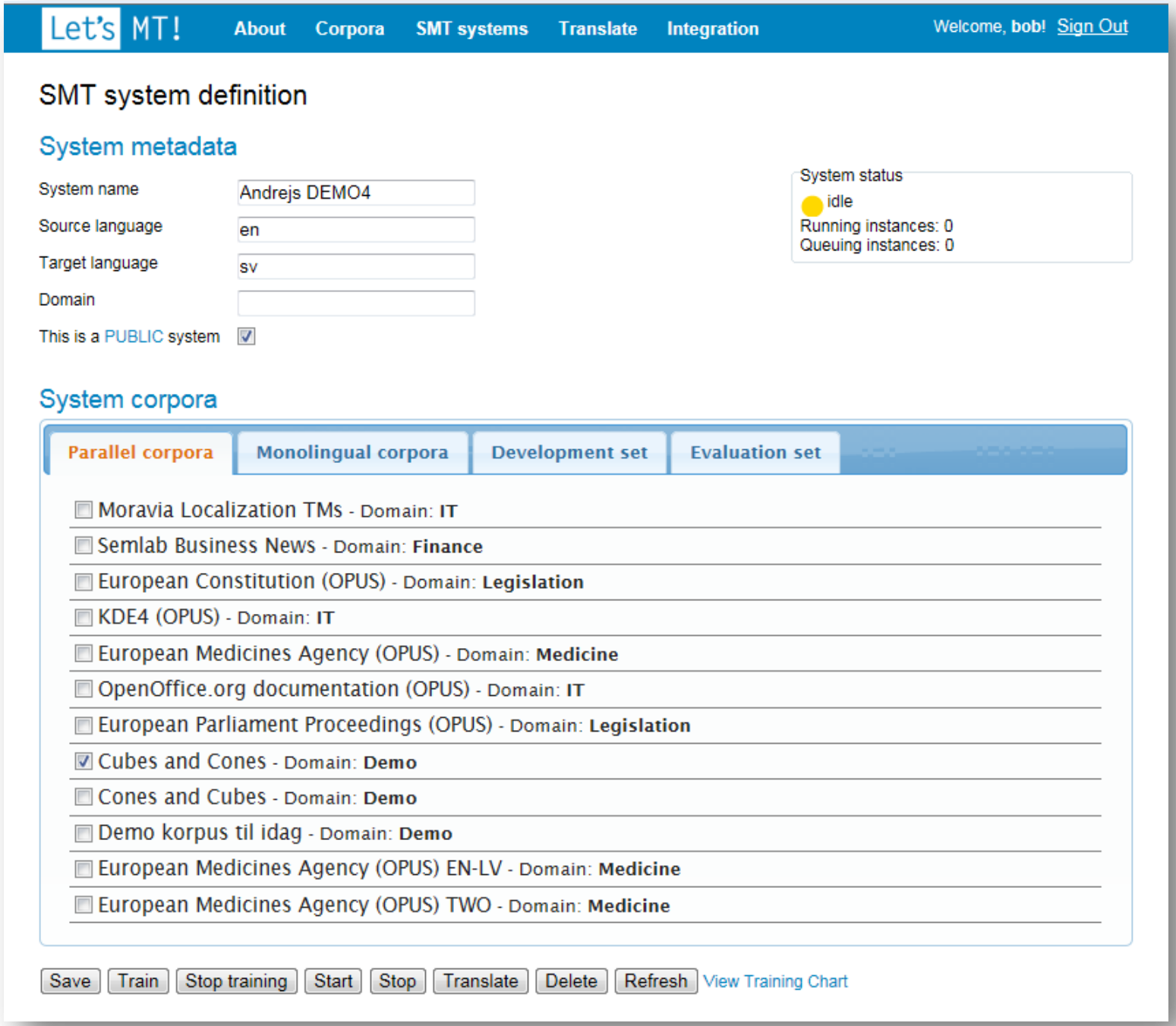
Let's MT!

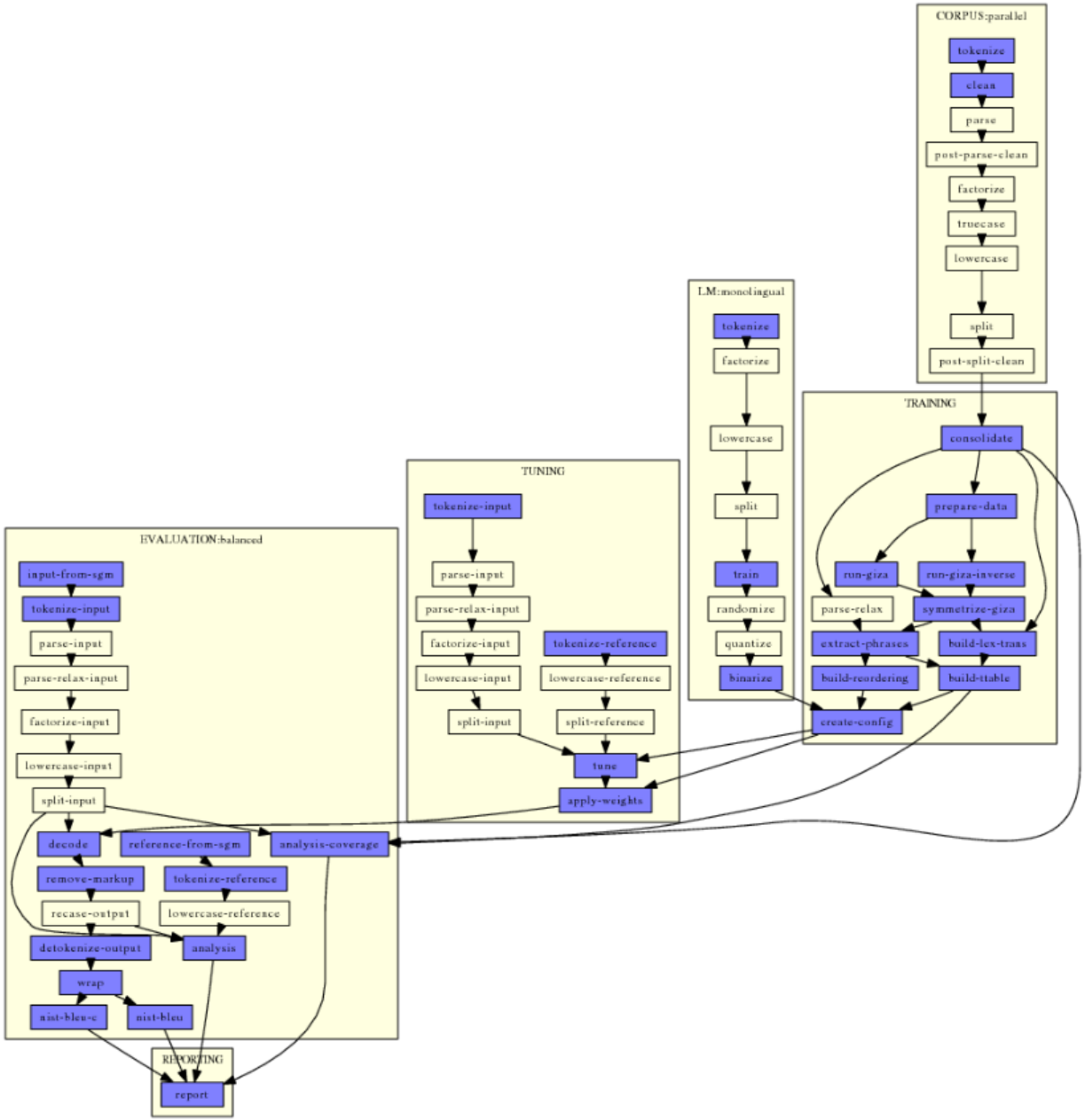
Let's MT!

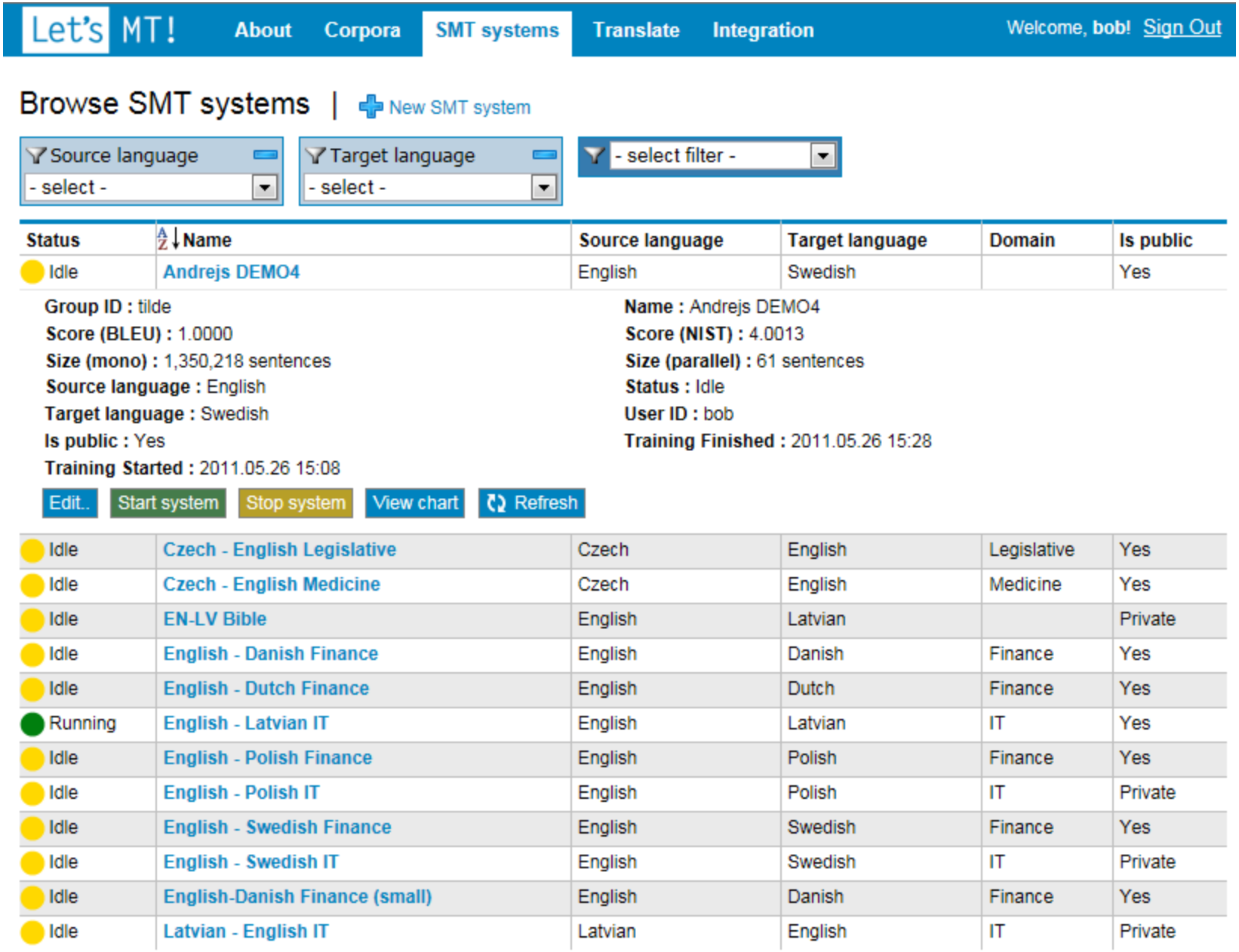


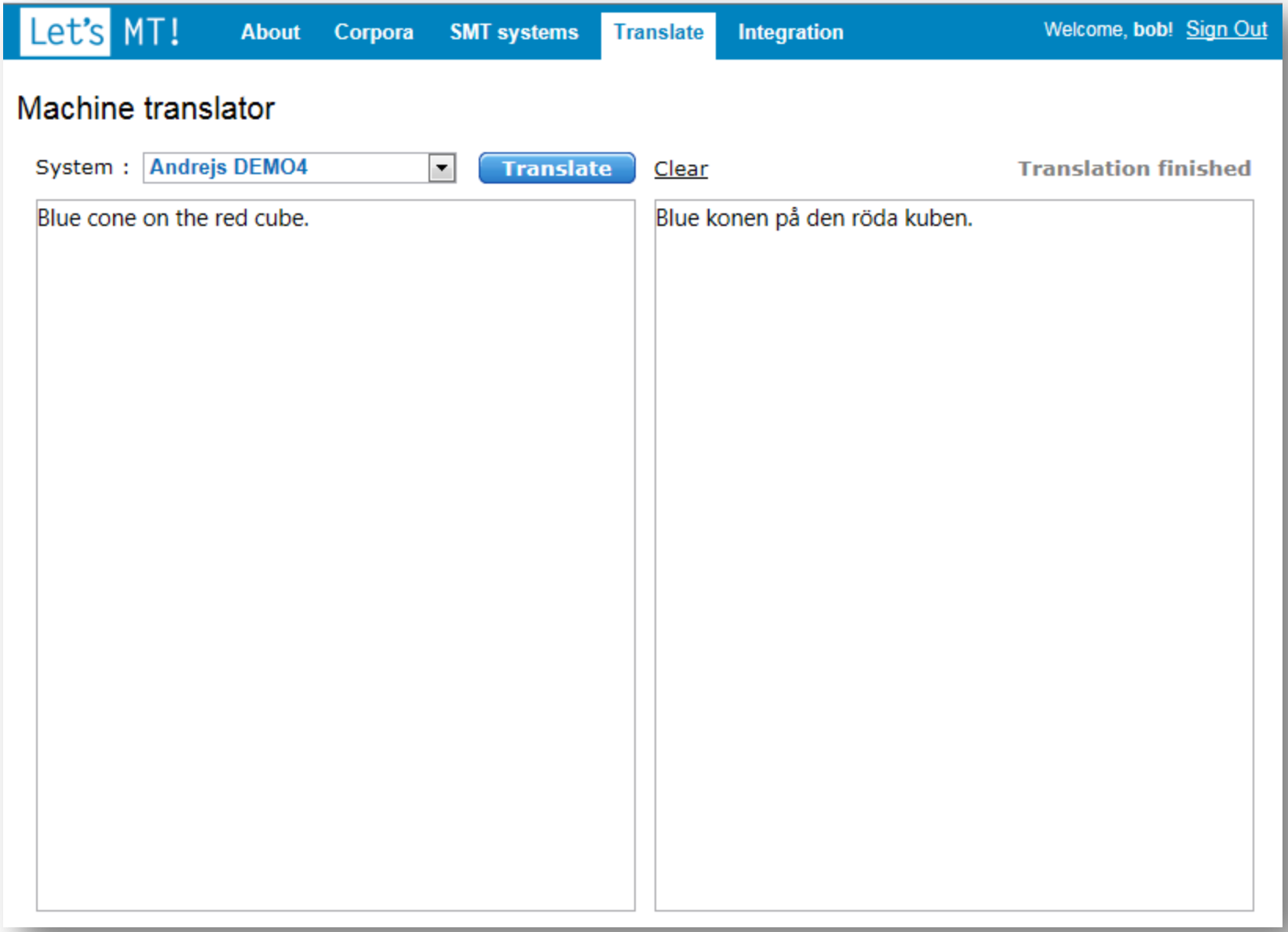
CLOUD BASED MT FACTORY

Let's MT!


Let's MT!
About
Corpora
SMT systems
Translate
Integration
Welcome, bob! Sign Out
SMT system definition
System metadata
System name
Andrejs DEMO4
Source language
en
Target language
sv
Domain
This is a PUBLIC system
System status
idle
Running instances: 0
Queuing instances: 0
System corpora
Parallel corpora
Monolingual corpora
Development set
Evaluation set
Moravia Localization TMs - Domain: IT
Semlab Business News - Domain: Finance
European Constitution (OPUS) - Domain: Legislation
KDE4 (OPUS) - Domain: IT
European Medicines Agency (OPUS) - Domain: Medicine
OpenOffice.org documentation (OPUS) - Domain: IT
European Parliament Proceedings (OPUS) - Domain: Legislation
Cubes and Cones - Domain: Demo
Cones and Cubes - Domain: Demo
Demo korpus til idag - Domain: Demo
European Medicines Agency (OPUS) EN-LV - Domain: Medicine
European Medicines Agency (OPUS) TWO - Domain: Medicine
Save
Train
Stop training
Start
Stop
Translate
Delete
Refresh
View Training Chart


CORPUS:parallel
tokenize
clean
parse
post-parse-clean
factorize
truecase
lowercase
split
post-split-clean
LM:monolingual
tokenize
factorize
lowercase
split
train
randomize
quantize
binarize
TRAINING
consolidate
prepare-data
run-giza
run-giza-inverse
parse-relax
symmetrize-giza
extract-phrases
build-lex-trans
build-reordering
build-ttable
create-config
TUNING
tokenize-input
parse-input
parse-relax-input
factorize-input
lowercase-input
split-input
tokenize-reference
lowercase-reference
split-reference
tune
apply-weights
EVALUATION:balanced
input-from-sgm
tokenize-input
parse-input
parse-relax-input
factorize-input
lowercase-input
split-input
decode
reference-from-sgm
analysis-coverage
remove-markup
tokenize-reference
recase-output
lowercase-reference
detokenize-output
analysis
wrap
nist-bleu-c
nist-bleu
REPORTING
report


Let's MT!
About
Corpora
SMT systems
Translate
Integration
Welcome, bob! Sign Out
Browse SMT systems | New SMT system
Source language
- select -
Target language
- select -
- select filter -
Status
Name
Source language
Target language
Domain
Is public
Idle
Andrejs DEMO4
English
Swedish
Yes
Group ID : tilde
Score (BLEU) : 1.0000
Size (mono) : 1,350,218 sentences
Source language : English
Target language : Swedish
Is public : Yes
Training Started : 2011.05.26 15:08
Name : Andrejs DEMO4
Score (NIST) : 4.0013
Size (parallel) : 61 sentences
Status : Idle
User ID : bob
Training Finished : 2011.05.26 15:28
Edit..
Start system
Stop system
View chart
Refresh
Idle
Czech - English Legislative
Czech
English
Legislative
Yes
Idle
Czech - English Medicine
Czech
English
Medicine
Yes
Idle
EN-LV Bible
English
Latvian
Private
Idle
English - Danish Finance
English
Danish
Finance
Yes
Idle
English - Dutch Finance
English
Dutch
Finance
Yes
Running
English - Latvian IT
English
Latvian
IT
Yes
Idle
English - Polish Finance
English
Polish
Finance
Yes
Idle
English - Polish IT
English
Polish
IT
Private
Idle
English - Swedish Finance
English
Swedish
Finance
Yes
Idle
English - Swedish IT
English
Swedish
IT
Private
Idle
English-Danish Finance (small)
English
Danish
Finance
Yes
Idle
Latvian - English IT
Latvian
English
IT
Private


Let's MT!
About
Corpora
SMT systems
Translate
Integration
Welcome, bob! Sign Out
Machine translator
System :
Andrejs DEMO4
Translate
Clear
Translation finished
Blue cone on the red cube.
Blue konen på den röda kuben.

MOTIVATING THE CROWD

- Collaborative translation using Collaborative Translation Framework tool by Microsoft Research
- User involvement in MT improvement
- Collaboration with libraries
- Translation game in collaboration with social network draugiem.lv

HISTORICAL MAPS OF ĢEOREFERENCĒŠAN

The oldest of the iesienamaj cards are *Ducatu Curlandi et Livonia Museum of*, dating back to 1720. for years, the most recent Latvian railway map, 1935.

Ģeoreferencējam material as ancient maps are interesting because of the development of creative thinking, but piņķerīg. In General, the card can be grouped into three groups, respectively, by the way, what information they contain, as well as for binding, as they prepare for binding:

READ MORE»

Comments 5:, categories: **digital library, items, General**

DIGITAL COLLECTION "LATVIA LOST"

February 23, 2011, author: **Arthur radonski**

This week, the Latvian National Library public presents its latest digitization project – **"lost Latvia"**. This article provides examples to familiarize yourself with some details of the project go backstage.

Project origins

The idea of "the lost Latvia ' occurred four years ago, when the NLL decided to enlist in the digitization and other cultural memory institutions and collect a variety of libraries, museums and private

HAVE A LOOK

» Digital Library
Isaiah Berlin and the Riga of His Time
Catalog» LNB
NLL website»
En» periodicals.

TWITTER SMS»

RT @silvenij-LNB_lv : @ Lost Latvian project (http://j.mp/f3RuqE) can be simple lost.
Extremely interesting and valuable resource!
6 days ago

Top digital also http://bit.ly/kfMh2m library (viLA_lv@)
6 days ago

FOLLOW THE NLL ON TWITTER»

LAST COMMENTS

Kristin on historic map of ģeoreferencēšan
Signe Valtiņ on historic map of ģeoreferencēšan
fein on historic map of ģeoreferencēšan
Mr. Serge on historic map of ģeoreferencēšan

TAG

poll copyright BibCamp-4 digitization costs non-conference conservation SZF history

BLOG NEWS IN YOUR E-MAIL

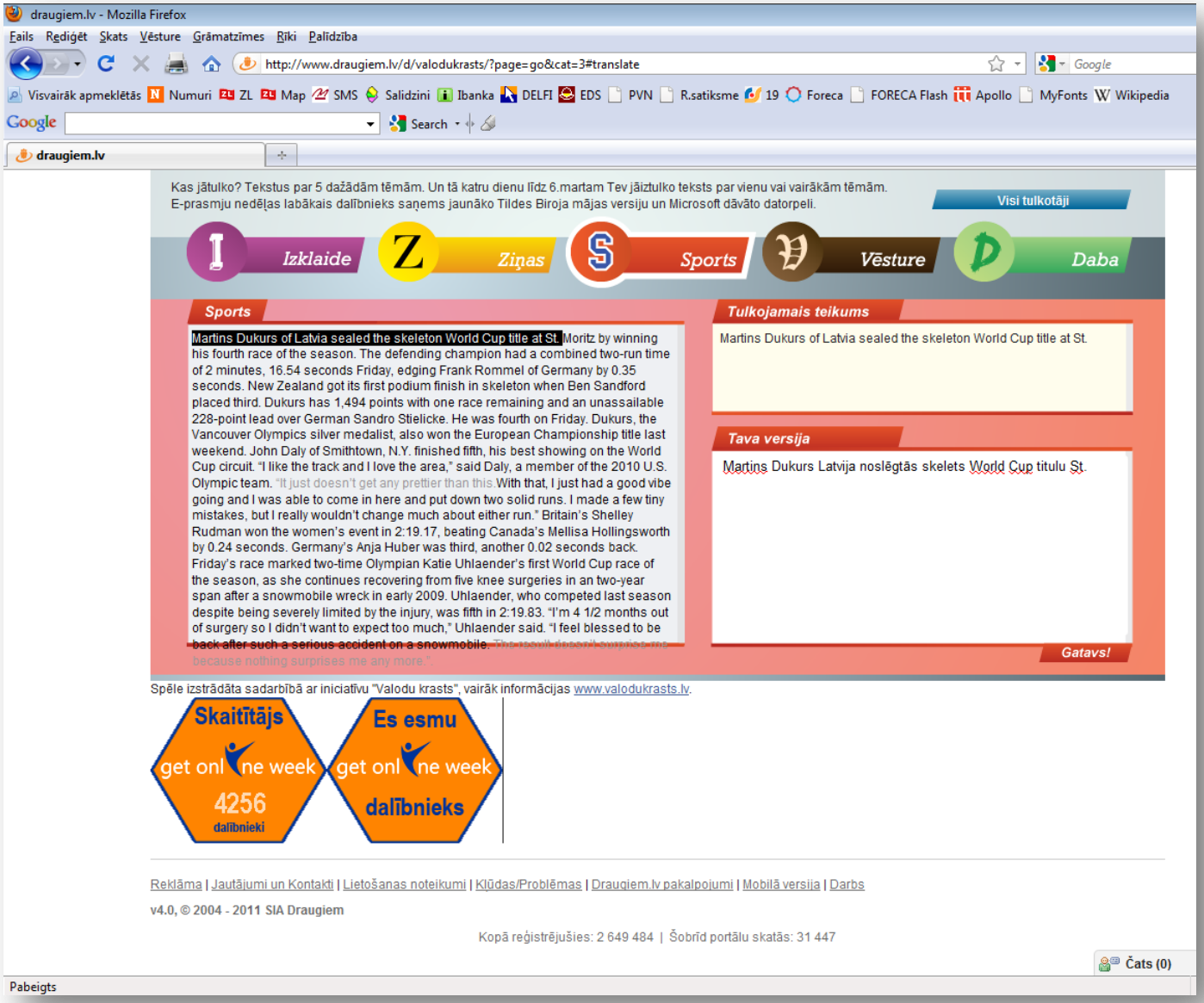
Type here your e-mail address

Abonēt

THE PROJECT THE LANGUAGE OF THE COAST OF , MAŠĪNTULKOTĀJ

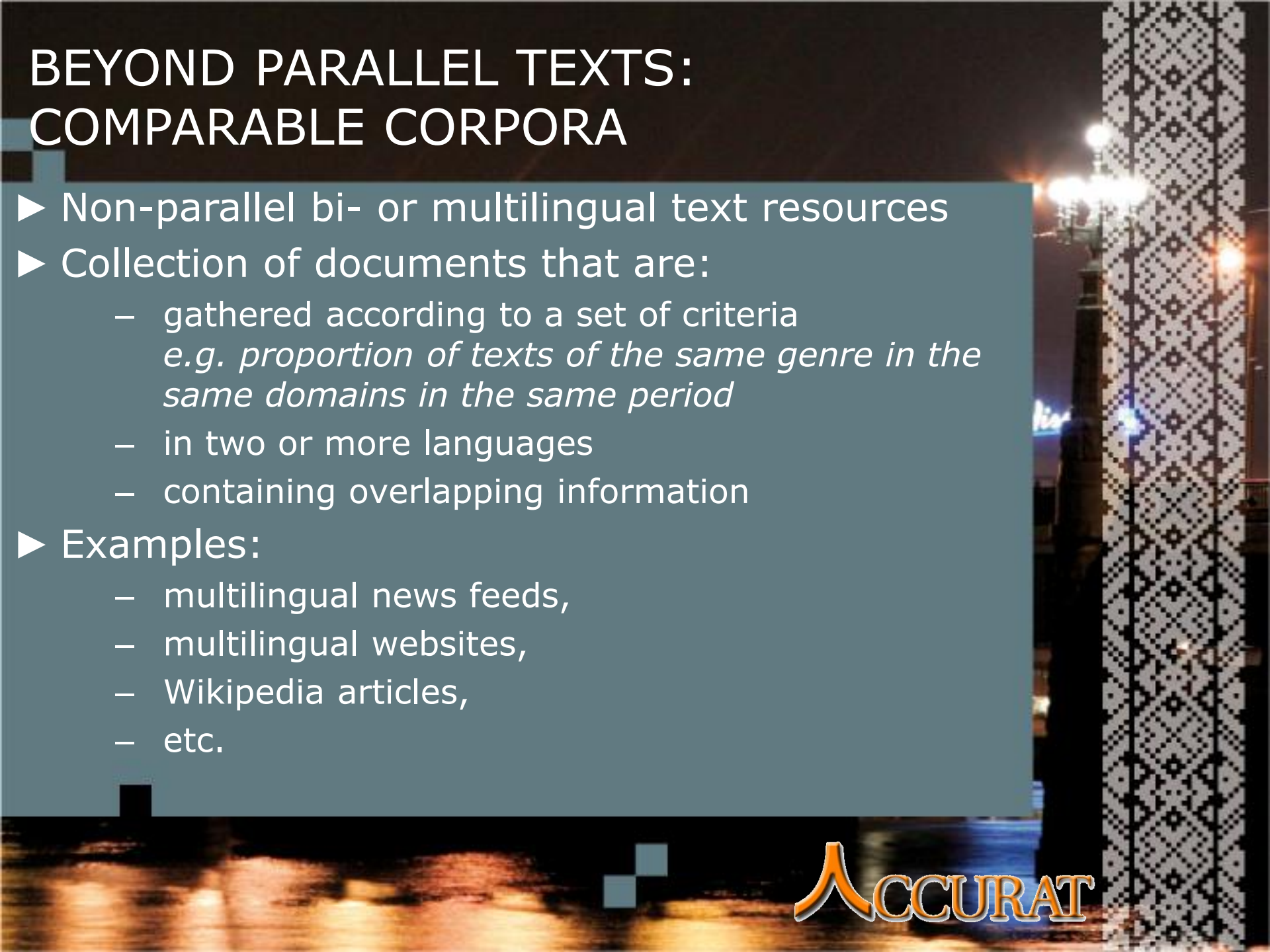
ARTICLE TOPICS

Digital library (64)
Information resources (7)
Interesting materials (5)
News (19)
Items (10)
Readers (9)
Legislation (6)
An overview of the measures to the NLL (3)
Services (1)
Reflections (27)
PR (15)
Projects (9)
Cooperation (5)
Content (20)
Technologies (26)
General (32)


draugiem.lv - Mozilla Firefox
Failis Redigēt Skats Vēsture Grāmatzīmes Rīki Palīdzība
http://www.draugiem.lv/d/valodukrasts/?page=go&cat=3#translate
draugiem.lv
Kas jātulko? Tekstus par 5 dažādām tēmām. Un tā katru dienu līdz 6.martam Tev jāiztulko teksts par vienu vai vairākām tēmām.
E-prasmju nedēļas labākais dalībnieks saņems jaunāko Tildes Biroja mājas versiju un Microsoft dāvāto datorpeli.
Visi tulkotāji
Izklaide
Ziņas
Sports
Vēsture
Daba
Sports
Martins Dukurs of Latvia sealed the skeleton World Cup title at St. Moritz by winning his fourth race of the season. The defending champion had a combined two-run time of 2 minutes, 16.54 seconds Friday, edging Frank Rommel of Germany by 0.35 seconds. New Zealand got its first podium finish in skeleton when Ben Sandford placed third. Dukurs has 1,494 points with one race remaining and an unassailable 228-point lead over German Sandro Stielicke. He was fourth on Friday. Dukurs, the Vancouver Olympics silver medalist, also won the European Championship title last weekend. John Daly of Smithtown, N.Y. finished fifth, his best showing on the World Cup circuit. "I like the track and I love the area," said Daly, a member of the 2010 U.S. Olympic team. "It just doesn't get any prettier than this.With that, I just had a good vibe going and I was able to come in here and put down two solid runs. I made a few tiny mistakes, but I really wouldn't change much about either run." Britain's Shelley Rudman won the women's event in 2:19.17, beating Canada's Mellisa Hollingsworth by 0.24 seconds. Germany's Anja Huber was third, another 0.02 seconds back. Friday's race marked two-time Olympian Katie Uhlaender's first World Cup race of the season, as she continues recovering from five knee surgeries in an two-year span after a snowmobile wreck in early 2009. Uhlaender, who competed last season despite being severely limited by the injury, was fifth in 2:19.83. "I'm 4 1/2 months out of surgery so I didn't want to expect too much," Uhlaender said. "I feel blessed to be back after such a serious accident on a snowmobile. This won't [illegible] because nothing surprises me any more."
Tulkojamais teikums
Martins Dukurs of Latvia sealed the skeleton World Cup title at St.
Tava versija
Martins Dukurs Latvija noslēgtās skelets World Cup titulu St.
Gatavs!
Spēle izstrādāta sadarbībā ar iniciatīvu "Valodu krasts", vairāk informācijas www.valodukrasts.lv.
Skaitītājs
get onl ne week
4256
dalībnieki
Es esmu
get onl ne week
dalībnieks
Reklāma | Jautājumi un Kontakti | Lietošanas noteikumi | Kļūdas/Problēmas | Draugiem.lv pakalpojumi | Mobilā versija | Darbs
v4.0, © 2004 - 2011 SIA Draugiem
Kopā reģistrējušies: 2 649 484 | Šobrīd portālu skatās: 31 447
Čats (0)
Pabeigts

BEYOND PARALLEL TEXTS: COMPARABLE CORPORA

- Non-parallel bi- or multilingual text resources
- Collection of documents that are:
  - gathered according to a set of criteria
    e.g. proportion of texts of the same genre in the same domains in the same period
  - in two or more languages
  - containing overlapping information
- Examples:
  - multilingual news feeds,
  - multilingual websites,
  - Wikipedia articles,
  - etc.



KEY RESEARCH QUESTIONS

- How to measure comparability?
- How to collect comparable corpora?
- How to extract linguistic data for MT from comparable corpora?
- How to get most out of the data to improve SMT and RBMT?
- How to evaluate effect of our methods?

ACCURAT

ACCURAT PROJECT OBJECTIVES

- Comparability metrics:
  automated measurement of the comparability of source and target language documents in comparable corpora
- Methods and tools for automatic acquisition of comparable corpora from the Web
- Extraction of lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT
- Implementation in real-life usage scenarios

ACCURAT

LET'S HELP SMALLER LANGUAGES TO GET MORE DATA!

Andrejs Vasiljevs, andrejs@tilde.com

letsmt.eu

Project supported by EU CIP ICT-PSP Programme

accurat-project.eu

Project supported by EU CIP ICT-PSP Programme

tilde.com