Moses at the European Commission

Francis Morton Tyers

17th July 2013

Outline

Outline

Introduction

Background:

- My PhD grant ran out in December and I was looking for work
- I saw an job posting for working on SMT at the EC and applied
- Worked there from March 2013 to July 2013

Disclaimer:

- I do not work and have never worked for the European Commission

Structure of the talk

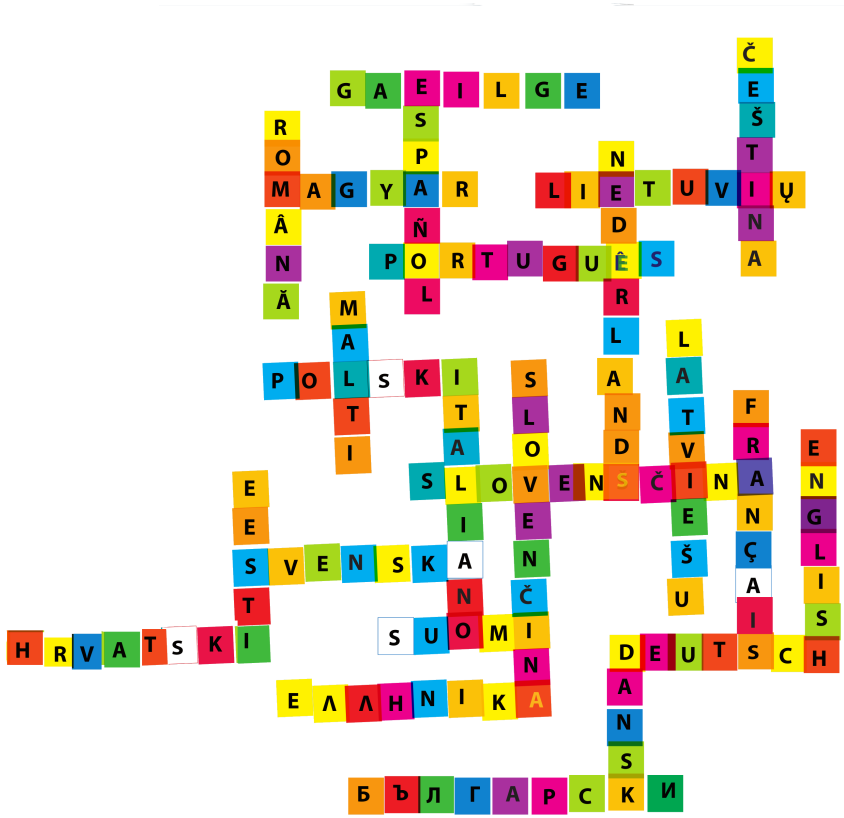
I was asked to talk about Moses at the European Commission (EC).

- **Introduction**
  - Languages and translation
  - History of MT at the EC
  - The MT@EC project
- **Experiments and development**
  - Incremental training
  - Word order
  - Morphology
  - Placeholders
- **Concluding remarks**

The objective of the talk is to answer the question ‘what is being done with Moses inside the European Commission?’

Languages at the European Commission

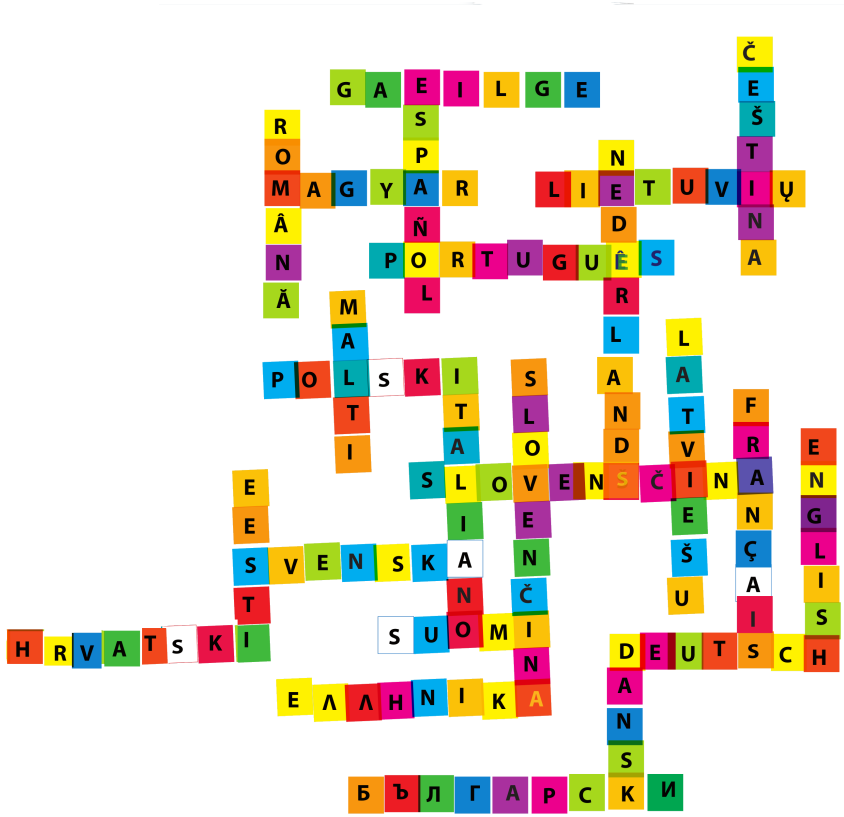
- Slavic: Bulgarian, Croatian, Czech, Polish, Slovak, Slovenian
- Romance: **French**, Italian, Portuguese, Romanian, Spanish
- Germanic: Danish, Dutch, **English**, German, Swedish
- Finno-Ugric: Estonian, Finnish, Hungarian
- Baltic: Latvian, Lithuanian
- Hellenic: Greek
- Semitic: Maltese
- Celtic: Irish



One language, one department (except Irish)

Languages at the European Commission

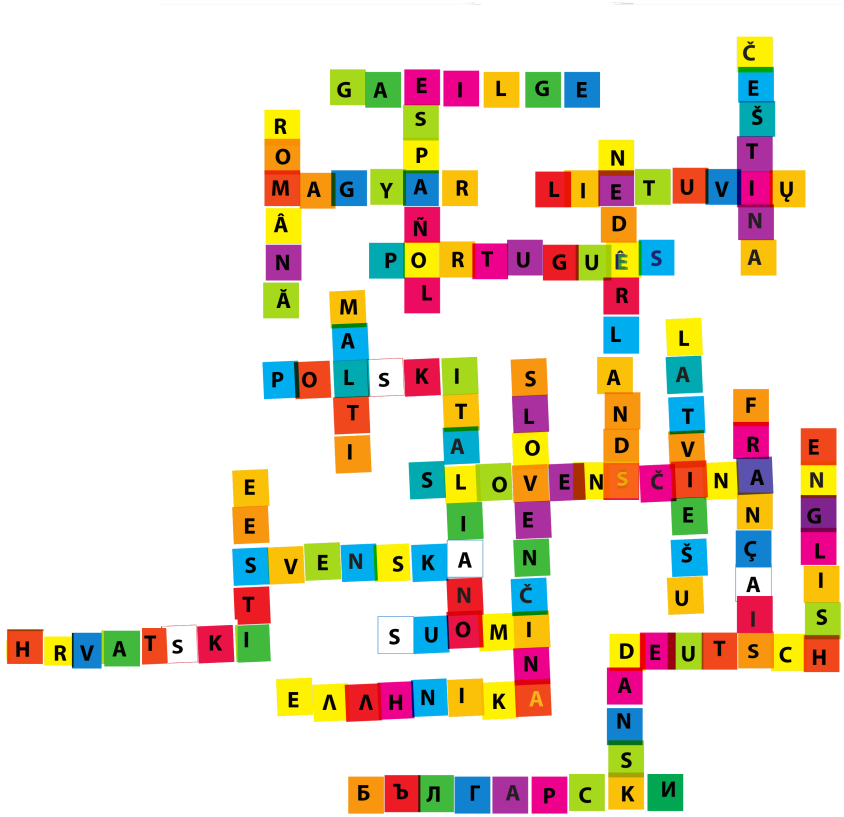
- Slavic: Bulgarian, Croatian, Czech, Polish, Slovak, Slovenian
- Romance: **French**, Italian, Portuguese, Romanian, Spanish
- Germanic: Danish, Dutch, **English**, German, Swedish
- Finno-Ugric: Estonian, Finnish, Hungarian
- Baltic: Latvian, Lithuanian
- Hellenic: Greek
- Semitic: Maltese
- Celtic: Irish



One language, one department (except Irish)

Languages at the European Commission

- Slavic: Bulgarian, Croatian, Czech, Polish, Slovak, Slovenian
- Romance: **French**, Italian, Portuguese, Romanian, Spanish
- Germanic: Danish, Dutch, **English**, German, Swedish
- Finno-Ugric: Estonian, Finnish, Hungarian
- Baltic: Latvian, Lithuanian
- Hellenic: Greek
- Semitic: Maltese
- Celtic: Irish



One language, one department (except Irish)

What kind of text is translated?

Commission Implementing Regulation (EU) No 401/2012

http://tinyurl.com/ecxmashat

(eng) Textile articles that have a utilitarian function are excluded from Chapter 95, even when they have a festive design (see also the Harmonised System Explanatory Notes to heading 95.05, point (A), last paragraph). Classification under subheading 95051090 as other articles for Christmas festivities is therefore excluded.

658

(ces) Textilní výrobky, které mají užitkovou funkci, jsou vyloučeny z kapitoly 95, i když mají slavnostní design (viz též vysvětlivky k harmonizovanému systému k číslu 9505, písm. A), poslední odstavec). Zařazení do podpoložky 95051090 jako ostatní vánoční výrobky je proto vyloučeno.

Some history of MT at the EC

- ECMT (European Commission Machine Translation): a SYSTRAN-based system used since 1976; development and maintenance activities stopped in 2006; service discontinued in 2010 as a result of a ruling from the EU court (overturned 2013)
- A lot of customisation work was put into the system, hand coding multiword units

```
-AB1CL3AUSULA DE ASSUN1C6AO DE A D3IVIDA
FA100N1004.1....1EASSUMPTION OF DEBT CLAUSE
```

Why a bespoke system?

- Avoid vendor lock-in
  - Like what happened with SYSTRAN
- Confidentiality
  - EC documents are public, but perhaps not in the moment of translation
- Domain-specific
  - Take advantage of existing data …

Workflows

Translators working in DGT:

- A document arrives for translation
- It gets sent to planning and EURAMIS[1]
  - In EURAMIS the text is extracted
- The text is sent to MT@EC to make a TMX with MT
- The translator gets the original TMX and the TMX with the MT, imports them into Trados
- The text is translated, and sent back to EURAMIS

Other users in the Commission:

- Web form allows translation of documents and text snippets
- Mostly to save translators time (gisting for other users)

[1] The EU-wide translation memory

MT@EC: Project management

The MT@EC project is fairly big, development is split into three groups:

- Data:
  - Extracting data from EURAMIS (European Advanced Multilingual Information System)
  - Basically a big translation memory database
  - The files are “exported” in text format.
- Engines
  - The team takes the data and build translation models with Moses.
- Interface
  - Web services to integrate the system with the end-user applications (Trados, web interface, etc.)

Project management: Engines





The group:

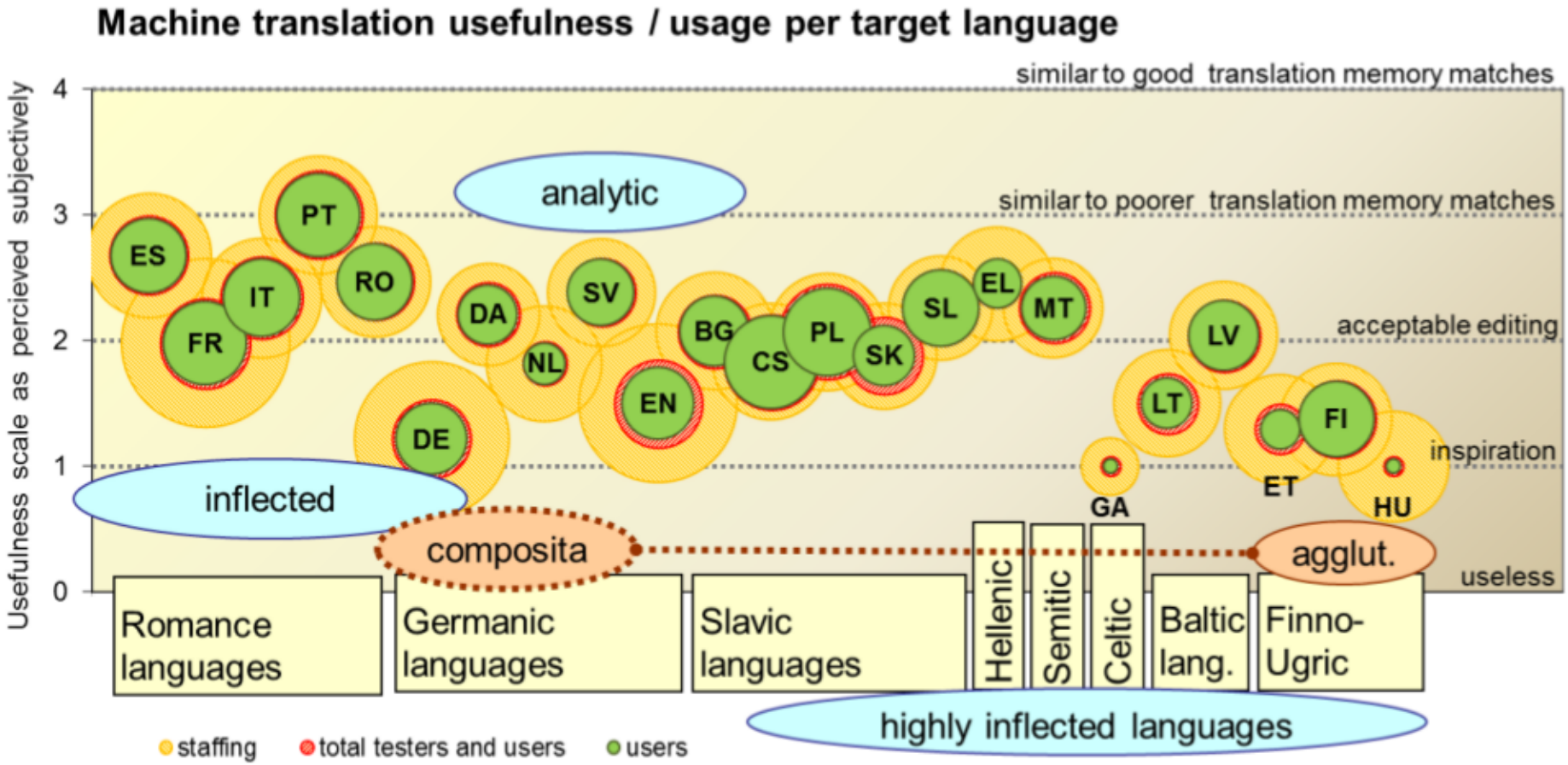
- Andreas Eisele (European Commission)
- Micha Jellinghaus (Fujitsu)
- Tom Vanallemersch (Fujitsu)
- László Tihanyi (IRIS)
- ?

How much training data is there?

For MT training:

- For most language pairs, there are around 10 million training segments
- For more recent languages (Irish, Croatian), around 300,000

User satisfaction



From 2012, graphic by Daniel Kluvanec

Outline

Infrastructure

The backbone of the system is Moses, with KenLM for language modelling. Tuning: MERT; and ttable pruning: Johnson et al. (2007)

Training:

- Set of python scripts to wrap around the training script
  - Avoids temporary files by using named pipes, and compresses on the fly — disk space is really expensive.
- The training process is automated, but each language pair needs to be started separately
- Training all the pairs takes around 2–3 weeks on around 4 servers
- Each model comes to around 5Gb

Other stuff:

- There is a translation ‘cache’ in SQLite which input segments are checked against before translating.

Engine generations

First generation (May, 2011):

- Prototype

Second generation:

- More data

Third generation (January, 2013):

- More data
- Input normalisation
  - Fixing typos
  - Fixing punctuation errors (e.g. extra spaces)

Fourth generation (July, 2013):

- More data
- Croatian
- Pivot translation
- Placeholders

Input normalisation

- **German**: replace incorrect beta character by “ß”
- **Greek**: correct some frequent abbreviations mixing Latin and Greek characters, e.g. “ ” (3 Greek letters) instead of “E K” (Latin E, Greek Omikron, Latin K)
- **Italian**: correct grave accents on some common words, e.g. “piu‘” -> “più”
- **Dutch**: correct capitalisation of IJ, e.g. “IJsland” instead of “Ijsland” repair incorrectly encoded characters etc.

Experiments

Priorities:

- Increasing acceptability of translations
  - Particularly for low-scoring language pairs or pairs with low acceptability

Experiments:

- Incremental training
- Word order
- Morphology
- Training-data expansion
- Placeholders

Most results are ‘negative’…

Incremental training

Question:

- Thousands of new segments translated daily.
- Training a whole system takes several days.
- Can reusing old alignments reduce training time ?

Motivation:

- Objective of the MT@EC project from the very beginning[2]

Approaches:

- **MGIZA++**: `http://www.kyloo.net/software/doku.php/mgiza:forcealignment`
- **Moses**: `http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc33`

[2] See Spyros Polis "Machine Translation at th European Commission" (Translingual Europe, 2010).

Incremental training (MGIZA++) /1

Setup:

- Language pair: English–Romanian
- Initial engine: 100k sentences
- Increment: 10k sentences
- Test corpus: 100 sentences

System	BLEU
Initial	16.4
Incremented	15.8
Retrained	16.6

Further investigation needed!

Incremental training (MGIZA++) /2

Setup:

- Language pair: English–Portuguese
- Initial engine: 50k sentences
- Increments: 500 – 16,000 sentences
- Test corpus: 300 sentences

Size	Increment	Retrained
50,000 (-)	34.88	34.88
50,500 (+500)	34.96	35.27
51,000 (+1,000)	34.91	34.94
52,000 (+2,000)	34.91	34.99
54,000 (+4,000)	35.00	35.19
58,000 (+8,000)	35.17	35.62
66,000 (+16,000)	35.49	35.70

Incremental training (MGIZA++) /2

Setup:

- Language pair: English–Portuguese
- Initial engine: 50k sentences
- Increments: 500 – 16,000 sentences
- Test corpus: 300 sentences

Size	Increment	Retrained
50,000 (-)	34.88	34.88
50,500 (+500)	34.96	35.27
51,000 (+1,000)	34.91	34.94
52,000 (+2,000)	34.91	34.99
54,000 (+4,000)	35.00	35.19
58,000 (+8,000)	35.17	35.62
66,000 (+16,000)	35.49	35.70

Incremental training (MGIZA++) /3

Setup:

- Language pairs: English–{Portuguese, French, Polish, German, Hungarian}
- Initial engine: 50k sentences
- Increments: 500 – 16,000 sentences
- Test corpus: 300 sentences

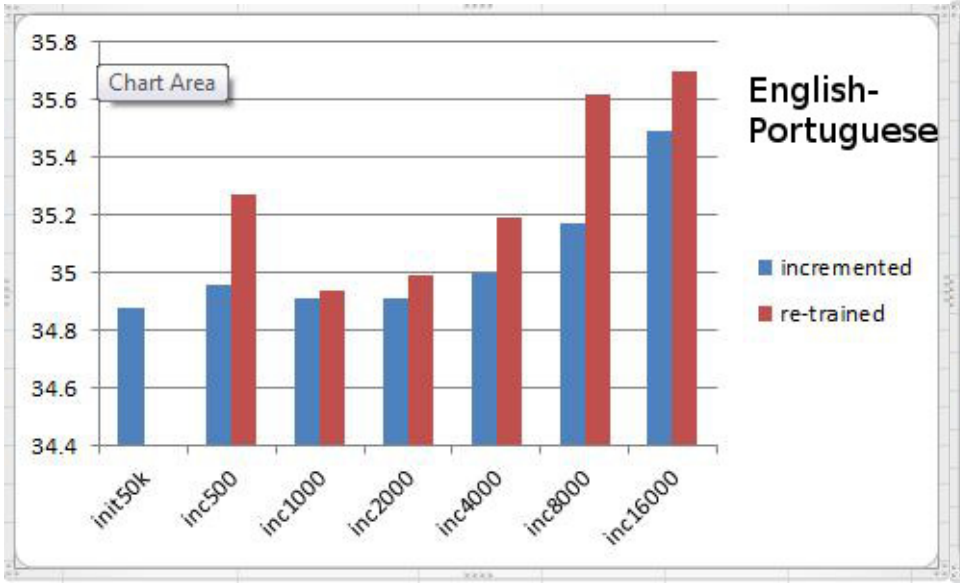
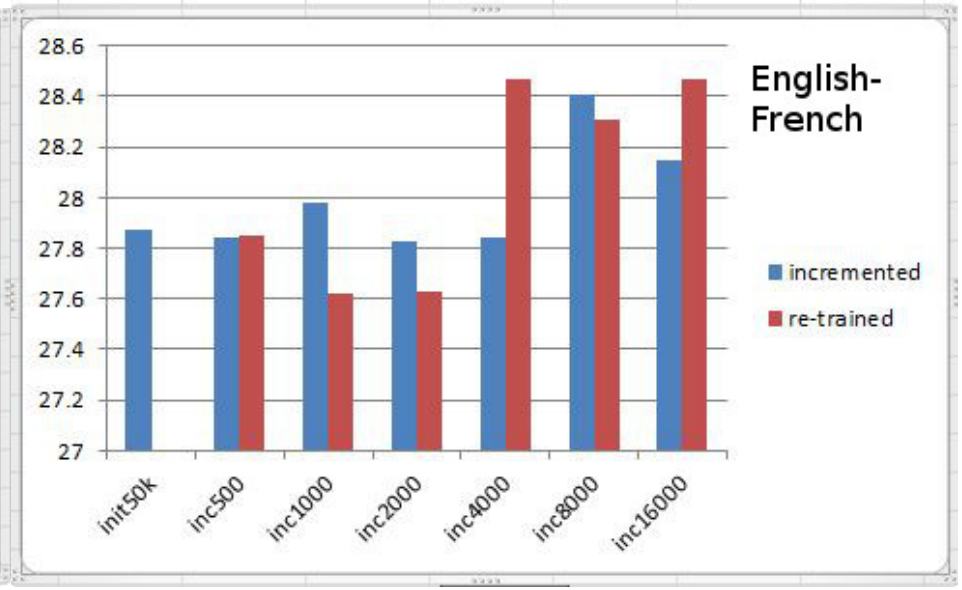
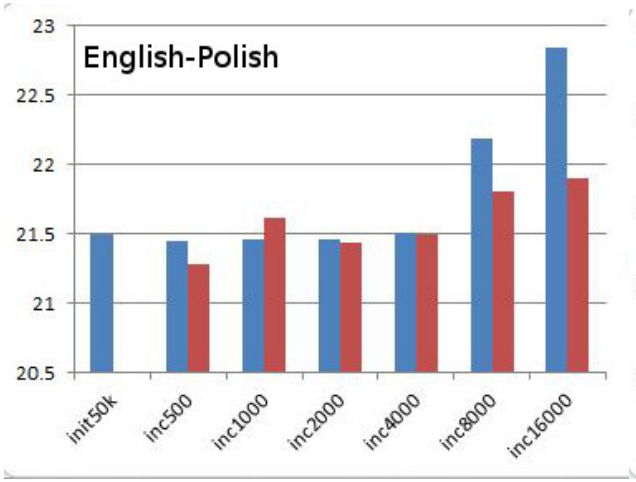
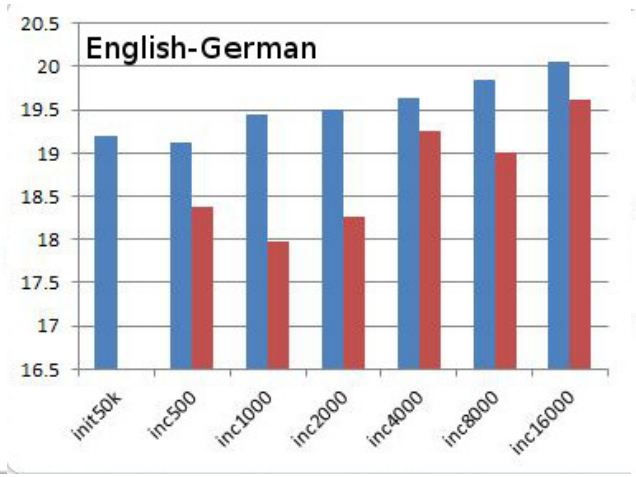
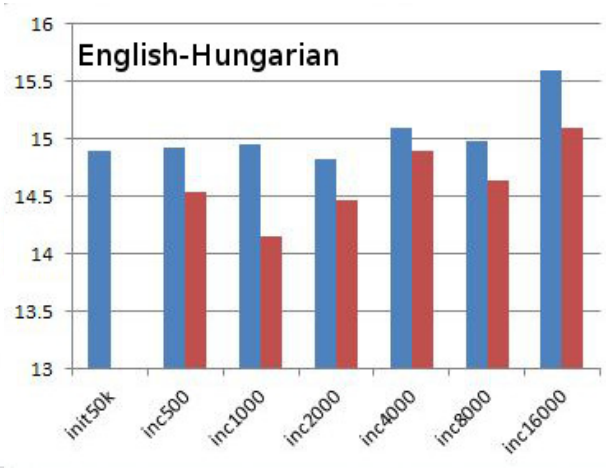


Chart Area
English-Portuguese
incremented
re-trained
35.8
35.6
35.4
35.2
35
34.8
34.6
34.4
init50k
inc500
inc1000
inc2000
inc4000
inc8000
inc16000


English-French
incremented
re-trained
28.6
28.4
28.2
28
27.8
27.6
27.4
27.2
27
init50k
inc500
inc1000
inc2000
inc4000
inc8000
inc16000


English-Polish
23
22.5
22
21.5
21
20.5
init50k
inc500
inc1000
inc2000
inc4000
inc8000
inc16000

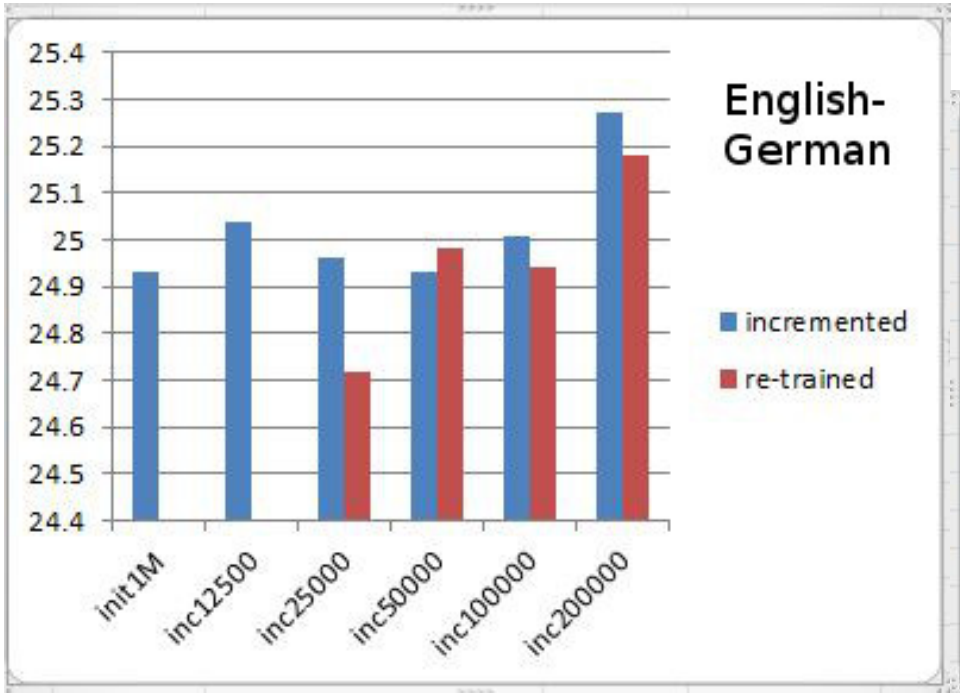

English-German
20.5
20
19.5
19
18.5
18
17.5
17
16.5
init50k
inc500
inc1000
inc2000
inc4000
inc8000
inc16000

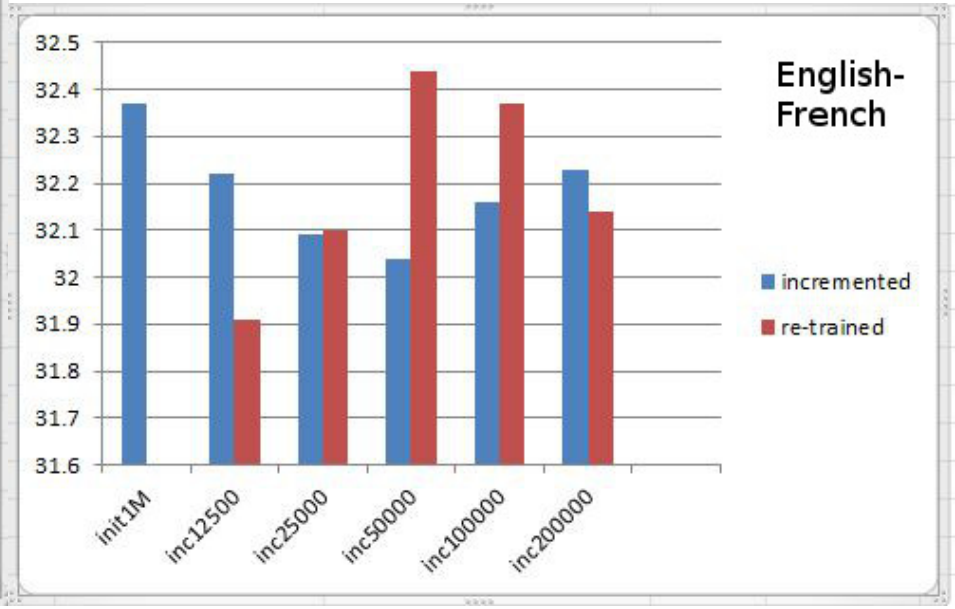

English-Hungarian
16
15.5
15
14.5
14
13.5
13
init50k
inc500
inc1000
inc2000
inc4000
inc8000
inc16000

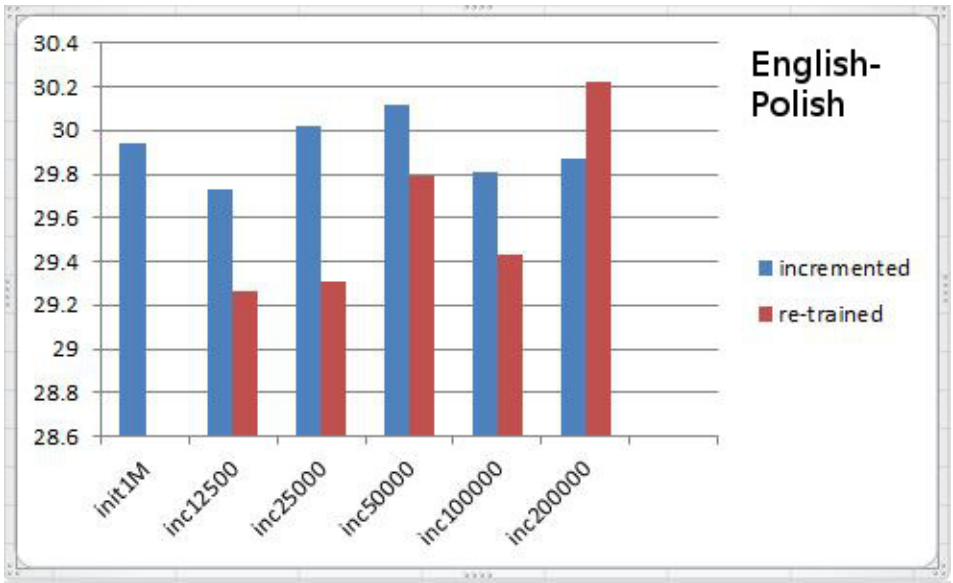
Incremental training (MGIZA++) /4

Setup:

- Language pairs: English–{German, French, Polish}
- Initial engine: 1m sentences
- Increments: 12.5k – 200k
- Test corpus: 300 sentences


English-German
25.4
25.3
25.2
25.1
25
24.9
24.8
24.7
24.6
24.5
24.4
init1M
inc12500
inc25000
inc50000
inc100000
inc200000
incremented
re-trained


English-French
32.5
32.4
32.3
32.2
32.1
32
31.9
31.8
31.7
31.6
init1M
inc12500
inc25000
inc50000
inc100000
inc200000
incremented
re-trained

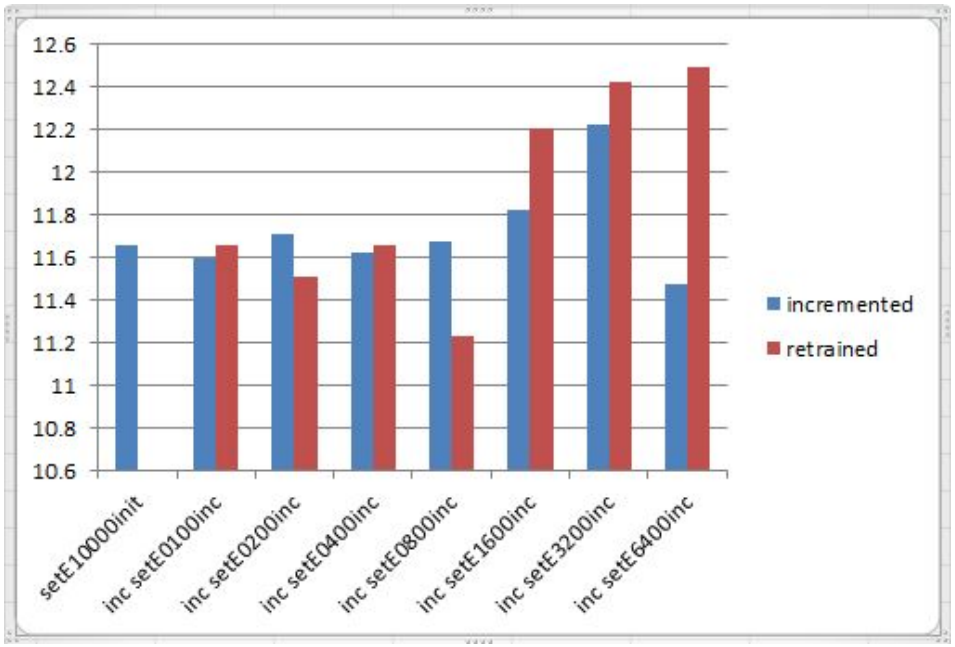

English-Polish
30.4
30.2
30
29.8
29.6
29.4
29.2
29
28.8
28.6
init1M
inc12500
inc25000
inc50000
inc100000
inc200000
incremented
re-trained

sadface

Incremental training (Moses) /5

Setup:

- Language pairs: English–Hungarian
- Initial engine: 10k
- Increments: 100 – 6.4k
- Test corpus: 300 sentences



Incremental training: Conclusions

- Mixed bag: Performance was variable
- Experiments directed at finding a combination that worked, and not directly comparable
- Depends too much on language pair and amount of training data

But:

- Why do the results vary so much between language pairs ?

Word order

Motivation:

- Results for languages with different word order are worse
- In principle: All languages should be equal

Word order (English → Hungarian) /1

Problem:

- Word-order differences between English and Hungarian
- Not between ‘constituents’, but inside

Example:

‘The meaning of the sentence.’

A	mondat	jelentés	-e
The	sentence	meaning	of

Resources:

- Berkeley parser (English)
- Parses are simplified
- A simple perl script:
  - preposition NP $\rightarrow$ NP preposition
    - in the house $\rightarrow$ the house in
  - possessive NP $\rightarrow$ NP possessive
    - in my house $\rightarrow$ house my in
  - the NP_1 of the NP_2 $\rightarrow$ the NP_2 NP_1 of
    - the meaning of the sentence $\rightarrow$ the sentence meaning of

Training corpus:

English	**Hungarian**
the sentence meaning of	a mondat jelentése
…	…

Word order (English → Hungarian) /2

Setup:

- Training: 100,000
- Testing: 1,000

Results:[3]

Original	Reordered
15.00	17.00

[3] These results are approximate.

Morphology

Motivation:

- To decrease data sparsity for morphologically-more-complex languages
- $\rightarrow$ Fewer unknown words and better statistics for known words

Approaches:

- Morpheme splitting
- Word-form simplification

Papers:

- Dyer, Muresan, and Resnik. “Generalizing Word Lattice Translation”. ACL2008

Morphology (Finnish → English) /1

"Tullitariffeja ja kauppaa koskeva yleissopimus 1994 rakentuu:"

Rf.: The General Agreement on Tariffs and Trade 1994 shall consist of:
Google: "On Tariffs and Trade in 1994 built on:"

Resources:

- Open Morphology for Finnish:
  `http://code.google.com/p/omorfi/`

Approaches:

- With fewest-splits segmentation
- Using lattice input

Morphology (Finnish → English) /2

Setup:

- 100,000 training sentences
- 2,000 test sentences (standard test set)
- Process corpus with morphological analyser, taking the output of the segmenter:
  - When there is ambiguous segmentation, select the segmentation with fewest splits
- Generate two phrase tables:
  - Surface forms
  - Segmented forms
- Use `decoding-graph-back-off` to back off to segmented forms

Morphology (Finnish → English) /3

- Input:
  - Tullitariffeja ja kauppaa koskeva yleissopimus 1994 rakentuu:
- Segmented:
  - Tullitariffe >j >a ja kauppa >a koskeva yleis sopimus 1994 rakentuu :
  - Gloss: Customs+tarrif PL PAR and trade PAR regarding general agreement 1994 builds :

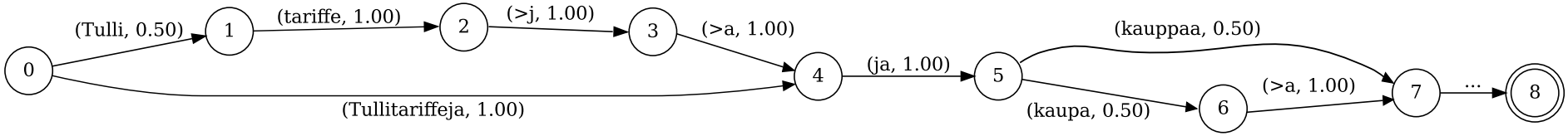
Training corpus:

English	**Finnish**
…on trade and tarrifs …	tullitariffe >j >a ja kauppa >a koskeva … …

Results:

Reduction in BLEU score

Morphology (Finnish → English) /4



Setup:

- Same setup as previous
- Used lattice input
- Weights: Surface form gets 0.5, segmented forms split the remaining 0.5 between them

Results:

Segmentation fault on line 100 of the test corpus :(– Didn't get around to debugging

Morphology (Latvian → English) /1

Motivation:

- What can be achieved with a very rudimentary morphological analyser?

Example:

- Certain languages include information in word forms that is not necessary when translating to another language that doesn't express this information
  - If we simplify/normalise forms, can we improve translation performance ?
    - e.g. change inflected forms for some words to their canonical form

Morphology (Latvian → English) /2

Latvian	**English**
Moderate inflection	Little inflection
Adjectives inflect for: comparison, gender, number, case, definiteness	Adjectives inflect for: comparison

Resources:

- Training data: 100,000 sentence subset of EC internal data
- Apertium morphology of Latvian[4]
  - Around 80% coverage of Latvian side of training data
- Rules to remove case/gender/number from Latvian adjectives
  - Only simplified in “safe” (unambiguous) cases
  - Gender altered to masculine, number to singular, case to nominative and definiteness to indefinite

[4] https://svn.code.sf.net/p/apertium/svn/incubator/apertium-lvs

Training corpus

Before:

Latvian	English
uz lielu , vecu koku	to the big old tree
uz lielus , vecus kokus	to the big old trees
…	…

After:

Latvian	English
uz liels , vecs koku	to the big old tree
uz liels , vecs kokus	to the big old trees
…	…

The case/number of the noun is unaltered, but the adjectives are simplified.

Morphology (Latvian → English) /3

Setup:

- Training: 100,000 sentences
- Testing: 10,000 sentences
- Apertium morphology of Latvian[5]
- Hand-written rules for simplifying adjectives

Results:

Without simplification	With simplification
28.4	29.1

Qualitative:

- Difficult to make a full qualitative evaluation because of many factors involved,
- Looked around 1,000 sentences: Some improvements and some regressions

[5] https://svn.code.sf.net/p/apertium/svn/languages/apertium-lvs

Training data expansion

Motivation:

- Can we improve SMT by synthetically creating training data by taking advantage of an existing RBMT system ?

Papers:

- Toral. (2012) ‘Pivot-based Machine Translation between Statistical and Black Box systems’. EAMT2012
- …

Training data expansion (English → Croatian) /1

Motivation:

- Croatian became official language of the EU on the 1st July
- Translation had started before this date
- Much more data for Slovenian (a closely-related language)

Resources:

- `apertium-hbs-slv`: A rule-based system between Slovenian and Serbo-Croatian (all three national standards).
- Existing EU data for English–Slovenian

Training data expansion (English → Croatian) /2

Setup:

- Full training set for English–Croatian: 500,000 segments
- Translated segments English–Croatian (via Slovenian): 2m segments

Results:

- Lower BLEU score
- RBMT system not mature enough
- Vocabulary coverage of the test set already very good

Placeholders

Motivation:

- Certain codes, references should be treated as a single unit, and should not get split up/reordered all over the place

Approach:

- Regular expressions for replacing numeral expressions / dates

> **Commission Implementing Regulation (EU) No 401/2012**
> Textile articles that have a utilitarian function are excluded from Chapter 95, even when they have a festive design (see also the Harmonised System Explanatory Notes to heading 95.05, point (A), last paragraph). Classification under subheading 95051090 as other articles for Christmas festivities is therefore excluded.

Results:

- Implemented as part of general improvements in the system

Comments and recap /1

- The environment at the EC is focussed on using existing results to improve a working system.
- Many sets of results are hard to reproduce, or the ideas are very fragile to the data set and experimental setup.

Opinion:

- Things would be helped with HOWTOs
- Homogeneity of linguistic resources
  - Piles upon piles of perl scripts changing input/output formats is not convenient for a production environment

Comments and recap /2

What might a HOWTO look like ?

- Self contained
- ‘Toy’ system
- Check that the setup works before extending

Homogeneity of linguistic resources:

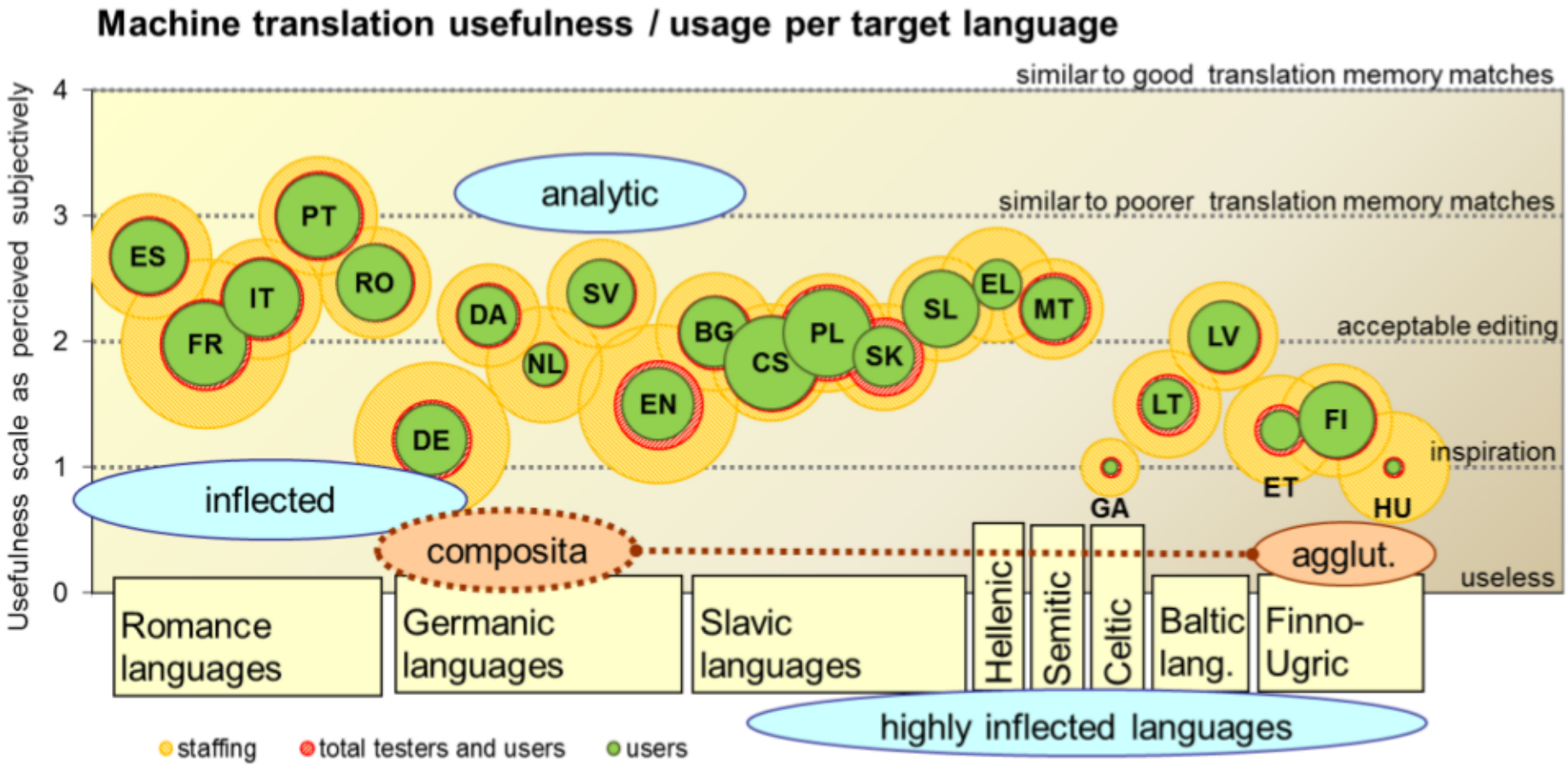
- Same input / output
- Not in conversion scripts!

Outline

Challenges

- The system always gets compared with Google
  - For in-domain we are much better, but general domain is another story
- The project is not a permanent feature
  - Continued funding depends on ‘results’ (or goodwill)
- Not allowed to share data
  - This would be cool if it could be arranged, perhaps an EC task at WMT ?

Future directions



No. of translators (staff, testers and regular users of machine translation) per target language grouped by morphology characteristics and language families

Future directions

- What would it take to get Hungarian to the level of Portuguese ?
  - If linguistic data is to be included, it will need to be made homogenous...
  - 5-person project, 24 languages...
  - In many cases, the state-of-the-art set of linguistic resources for each language has its own incompatible toolchain.
- How about non-EC languages?
  - Kimmo Rossi: “We have no possibility to support any work on Tetun, as we need to concentrate scarce resources on EU languages and some major world languages (ZH, JP, RU...)”
- Offer the service to national government bodies

How can ‘we’ help ?

Three things:

- Keep doing what we’re doing!
- Try working with more languages at the same time
- HOWTOs

Contacts

- Andreas Eisele
  - Andreas.EISELE@ec.europa.eu
- Micha Jellinghaus
  - Michael.JELLINGHAUS@ext.europa.eu
- László Tihanyi
  - Laszlo.TIHANYI@ext.europa.eu

Thanks · Gracias · Merci · Danke · Hvala · Tak · Bedankt · Kiitos · Köszönöm · Go raibh maith agat · Grazie · Paldies · Ačiū · Grazzi · Obrigado · Mulţumesc · Tack · Ďakujem · Děkuji · Благодаря · Gràcies · Giitu · Aitäh · Ευχαριστώ · Eskerrik asko · Gràcies · Dziękuję