Translation Quality in MOLTO use-cases

Jussi Rautio, Maarit Koponen

- All evaluations were made by native or near-native level speakers of each language, 34 evaluators in all (plus 11 translation studies students in UHEL for Finnish tourist phrasebook)

Tourist Phrasebook

- 13 languages with 139 sentences: **BUL, CAT, DAN, DUT, FIN, FRE, GER, ITA, NOR, POL, RON, SPA, SWE**
- Two evaluators were given translation suggestions by **GF, Google, Bing** and **Systran** (in ITA, POL, GER, FRE, DUT, SWE, SPA) in a random order
- Evaluators chose the suggestion they deemed the best and post-edited if needed
- BLEU, NIST, TER, WER and PER were calculated from the two+ references obtained

Phrasebook TER Results

	gf_TER	google_TER	bing_TER	systran_TER
avg	**0,084**	**0,258**	**0,263**	**0,265**
SWE	0,017	0,194	0,235	0,290
SPA	0,035	0,242	0,176	0,202
CAT	0,040	0,309	0,292	n/a
ITA	0,050	0,290	0,203	0,237
GER	0,052	0,262	0,209	0,227
FIN	0,053	0,330	0,377	n/a
FRE	0,078	0,251	0,252	0,201
DUT	0,083	0,176	0,226	0,174
NOR	0,085	0,220	0,248	n/a
POL	0,094	0,303	0,340	0,526
DAN	0,100	0,163	0,174	n/a
RON	0,175	0,287	0,349	n/a
BUL	0,232	0,331	0,331	n/a

TER (Translation Error Rate) measures the amount of editing needed to get the reference translation:
Lower score = Better
Score of 0 = Exact match

Notes on the scores

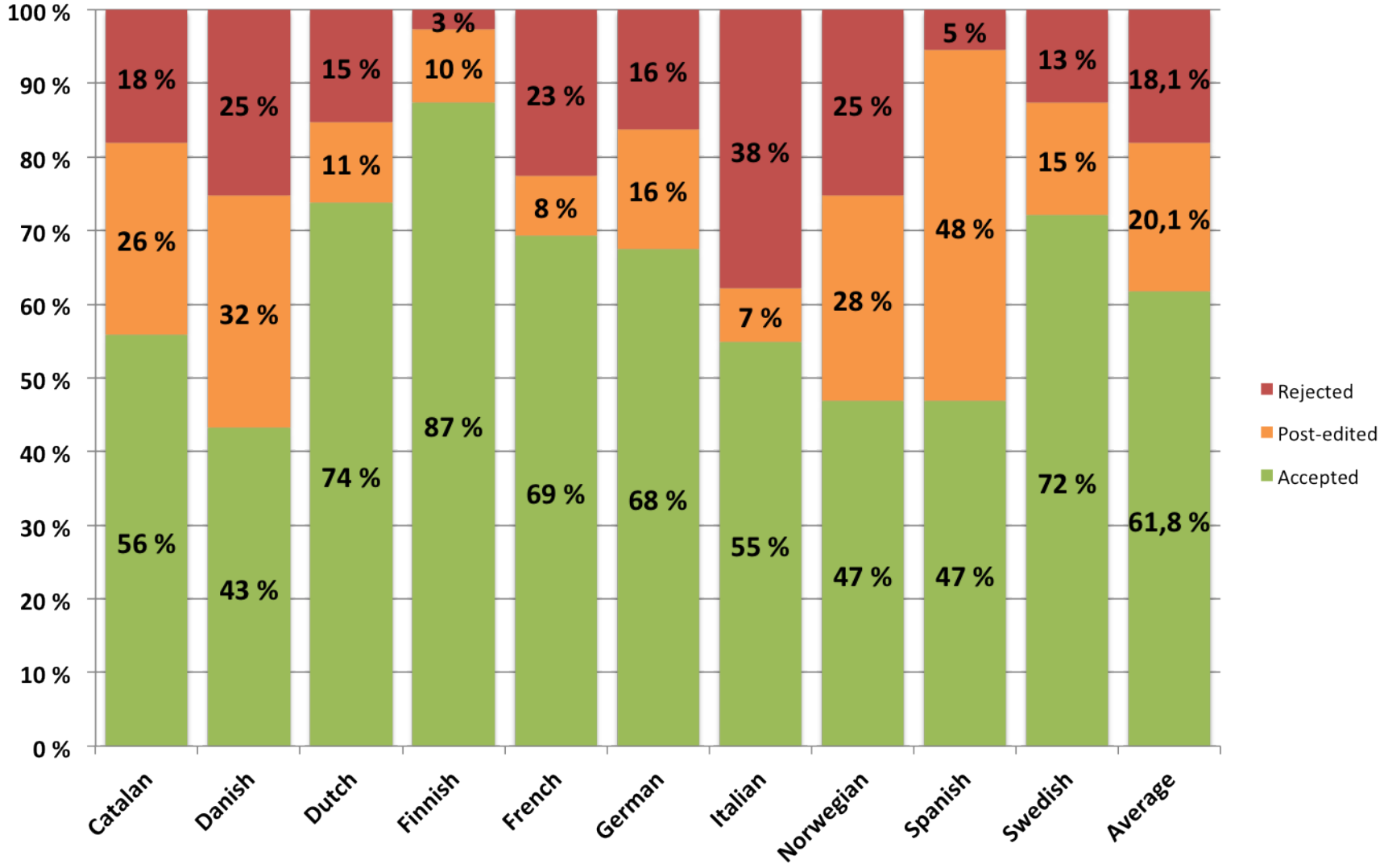
- GF gets the best average score in all the metrics, especially TER

	BLEU	NIST	TER	WER	PER
GF	**0,829**	**8,616**	**0,084**	**0,202**	**0,163**
Google	0,561	6,790	0,258	0,397	0,333
Bing	0,540	6,717	0,263	0,397	0,333
Systran	0,485	6,136	0,290	0,436	0,384

We also studied which suggestions the evaluators chose as such or for post-editing

- Accepted: One or both of the evaluators chose the GF suggestion as such
- Edited: One or both of the evaluators chose the GF suggestion for post-editing
- Rejected: Both evaluators chose a suggestion from another MT or translated from scratch

	Accepted %	Edited %	Rejected %
SWE	91 %	3 %	6 %
SPA	83 %	4 %	14 %
ITA	81 %	3 %	16 %
CAT	80 %	12 %	8 %
FIN	77 %	16 %	7 %
FRE	75 %	14 %	11 %
GER	73 %	12 %	15 %
DUT	69 %	4 %	27 %
POL	65 %	23 %	12 %
NOR	61 %	22 %	17 %
DAN	57 %	22 %	21 %
RON	47 %	28 %	24 %
BUL	28 %	44 %	28 %
avg	68 %	16 %	16 %


100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
18 %
26 %
56 %
25 %
32 %
43 %
15 %
11 %
74 %
3 %
10 %
87 %
23 %
8 %
69 %
16 %
16 %
68 %
38 %
7 %
55 %
25 %
28 %
47 %
5 %
48 %
47 %
13 %
15 %
72 %
18,1 %
20,1 %
61,8 %
Catalan
Danish
Dutch
Finnish
French
German
Italian
Norwegian
Spanish
Swedish
Average
Rejected
Post-edited
Accepted

More notes on the scores

- Bulgarian and Romanian got the lowest scores: This is because GF did not have the pro-drop forms which the evaluators preferred. This should be easy to fix.
- GF got good scores in Finnish in which Google and Bing got the lowest ones.
- GF translations were significantly more preferred than the other MTs

ACE-in-GF

- The same setup as in the phrasebook, although just Google as a second choice (as Bing and Systran were seen to be much worse with the phrasebook)
- 10 languages, 111 sentences
  - Results reported in D11.3

	ACE-in-GF					Google Translate				
	BLEU	**NIST**	**TER**	**WER**	**PER**	**BLEU**	**NIST**	**TER**	**WER**	**PER**
CAT	0,809	8,803	0,101	0,231	0,223	0,716	7,993	0,151	0,265	0,232
DAN	0,716	8,233	0,142	0,263	0,208	0,623	7,452	0,186	0,324	0,244
DUT	0,899	9,335	0,056	0,223	0,158	0,735	8,371	0,133	0,275	**0,170**
FIN	**0,948**	**9,336**	**0,026**	**0,147**	**0,132**	0,446	6,053	0,321	0,401	0,365
FRE	0,873	8,998	0,073	0,221	0,179	0,784	8,284	0,128	0,258	0,217
GER	0,850	9,027	0,060	0,162	0,152	0,660	7,943	0,166	0,289	0,187
ITA	0,822	8,626	0,090	0,191	0,173	0,793	8,186	0,116	**0,204**	0,181
NOR	0,718	8,142	0,116	0,248	0,187	0,687	7,795	0,152	0,240	0,199
SPA	0,788	8,835	0,095	0,224	0,198	0,708	7,994	0,167	0,281	0,212
SWE	0,889	9,303	0,056	0,300	0,226	**0,794**	**8,723**	**0,093**	0,260	0,194
avg	**0,831**	**8,864**	**0,081**	**0,221**	**0,184**	**0,695**	**7,879**	**0,161**	**0,280**	**0,220**

	Accepted %	Post-edited %	Rejected %
Catalan	56 %	26 %	18 %
Danish	43 %	32 %	25 %
Dutch	74 %	11 %	15 %
Finnish	87 %	10 %	3 %
French	69 %	8 %	23 %
German	68 %	16 %	16 %
Italian	55 %	7 %	38 %
Norwegian	47 %	28 %	25 %
Spanish	47 %	48 %	5 %
Swedish	72 %	15 %	13 %
Average	61,8 %	20,1 %	18,1 %

Notes on ACE-in-GF

More variation than in the phrasebook:

what doesn't John hate?

E1: Qu'est-ce que John ne déteste pas?

E1: John ne hait pas quoi?

GF: Que n'hait pas John?

Google: Ce n'est pas John haine?

Patents

- Automatic metrics evaluated in D5.3
- Pure SMT, a GF hybrid (R4), Google, Systran and Pluto translations (DE, FR) ranked on a TAUS scale of 1-4:

4 = Complete. All of the information in the source was available from the target; reading the source did not add to information or understanding.

3 = Useful: The information in the target was correct and clear, but reading the source added some additional information or understanding.

2 = Marginal. The information in the target was correct, but reading the source provided significant additions of clarifications.

1 = Poor. The information in the target was unclear and/or incorrect; reading the source would be necessary for understanding.

German

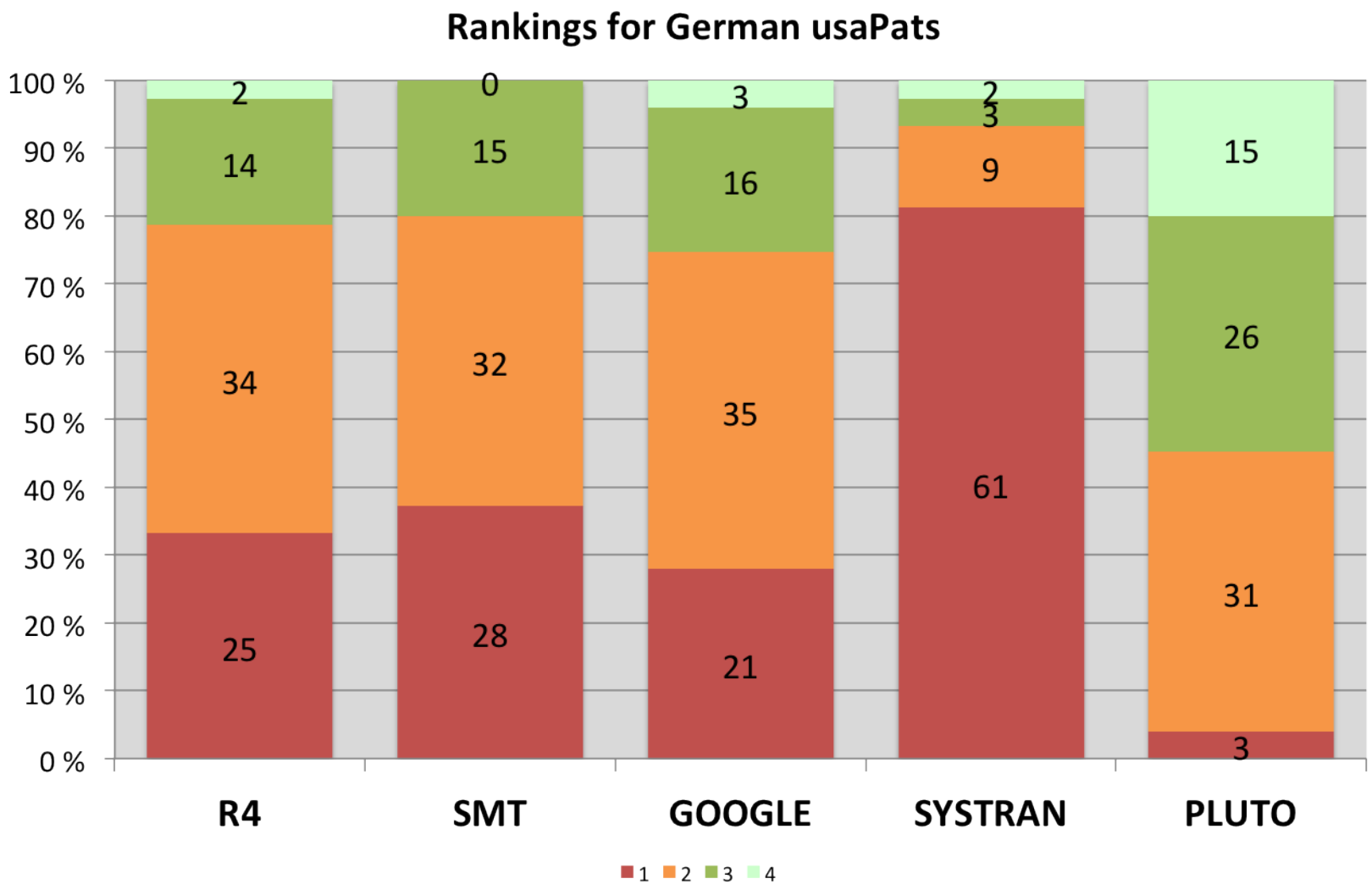
patsA61D (149 sent.)	**Hybrid R4**	**SMT**	**Google**	**Systran**
mean	3,074	3,087	**3,221**	2,450
median	3,000	3,000	3,000	2,000
epoA61D (200 sent.)				
mean	2,650	2,800	**2,900**	2,085
median	3,000	3,000	3,000	2,000
usaPats (75 sent.)				
mean	1,744	1,674	**1,846**	1,178
median	2,000	2,000	2,000	1,000
All				
mean	2,667	2,729	**2,856**	2,071
median	3,000	3,000	3,000	2,000

French

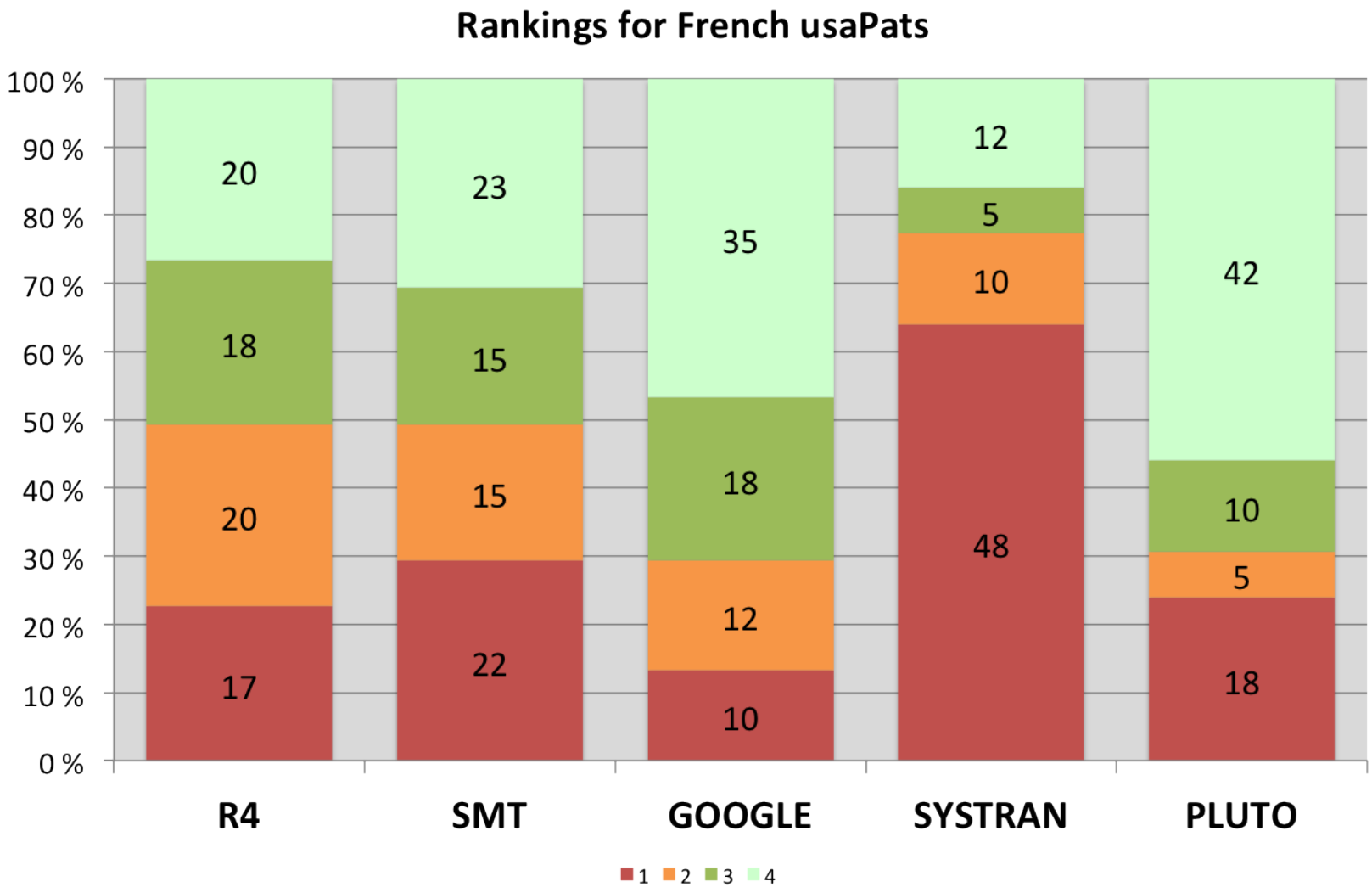
patsA61D	**Hybrid R4**	**SMT**	**Google**	**Systran**	**Pluto**
mean	3,095	3,088	**3,284**	2,554	n/a
median	4	3	4	3	n/a
epoA61D					
mean	2,969	2,990	**3,136**	2,068	n/a
median	3,000	3,000	4,000	2,000	n/a
usaPats					
mean	2,547	2,520	**3,040**	1,747	3,013
median	3	3	3	1	**4**
all					
mean	2,932	2,935	**3,167**	2,179	n/a
median	3	3	4	2	n/a

usaPats

- The usaPats batch had no reference translations, but it was compared against PLuTO (http://www.pluto-patenttranslation.eu/) on the 1-4 scale



	r4	smt	google	systran	pluto
Median	2	2	2	1	3
Average	1,744	1,674	1,846	1,178	2,566



	r4	smt	google	systran	pluto
Median	3	3	3	1	4
Average	2,533	2,507	3,027	1,733	3,013

Notes on patents

- Google or Pluto gets better scores, Pluto significantly so
- There's absolutely no difference between the GF hybrid and SMT rankings
- Even exact matches with the reference translation got sometimes ranked as 1 or 2: Even the published human translations are evaluated as unpublishable and very low BLEU scoring sentences got a 4.

Maths

- Mathematical clauses were evaluated in DE, SW, FI, FR by a maths expert.
- DE, SW, FR were evaluated as "good" with some issues with some preposition and terminology issues. Some ambiguous sentences were also spotted.
- Finnish had a lot of issues that are being fixed

Cultural heritage

- Museum data structurally repetitive
  - Not meaningful task for evaluators
  - Subjective terminological preferences
- Error analysis by construction

Evaluation of effort of grammar development

- GF best practices and new tools
  - not yet in place at outset
  - bug tracking with TRAC
- Estimating effort expended
  - Adding a new language for a domain (WP 10)
  - Writing grammars for a new domain (WP 8)
  - Fixing grammar errors (WP 3)

Adding a new language

- Phrasebook case

Language	Language skills	GF [4] skills	Informed development	Informed testing	Impact of external tools	RGL Changes	Overall effort
Bulgarian	###	###	-	-	?	#	##
Catalan	###	###	-	-	?	#	#
Danish	-	###	+	+	##	#	##
Dutch	-	###	+	+	##	#	##
English	##	###	-	+	-	-	#
Finnish	###	###	-	-	?	#	##
French	##	###	-	+	?	#	#
German	#	###	+	+	##	##	###
Italian	###	#	-	-	?	##	##
Norwegian	#	###	+	-	##	#	##
Polish	###	###	+	+	#	#	##
Romanian	###	###	-	-	#	###	###
Spanish	##	#	-	-	?	-	##
Swedish	##	###	-	+	?	-	##

Adding a new domain and language

- Museum case
  - Abstract grammar:
    - Input: museum data, model verbalizations
    - Effort: medium term (incl. ontology dev.)
  - Adding languages:
    - Input: abstract grammar, RGL
    - Effort: 98% reduction from using RGL
  - GF skills:
    - high to medium (computer scientist + computational linguist)

Fixing bugs in grammar

- Lexicon errors
  - minutes
- Application grammar errors
  - tens of minutes
- Resource grammar errors
  - up to a day
- GF skills
  - medium (computational linguist)