



EuroMatrixPlus

Evaluation, Localisation, Open Source

Josef van Genabith

Centre for Next Generation Localisation CNGL

School of Computing

Dublin City University, Ireland



Overview



- EuroMatrix (2006-2009)
- EuroMatrixPlus (2009 -2012)
- Evaluation
- Localisation
- Open Source

EuroMatrix 2006-2009



Goals

- MT between all EU languages
- Open Research Environment
- Open Source



EuroMatrix 2006-2009



Partners

- University of Saarbrücken
- University of Edinburgh
- Charles University Prague
- CLECT
- Group Technologies
- Morphologic



EuroMatrix 2006-2009



Existing MT systems for EU languages

[from Hutchins, 2005]

	Cze	Dan	Dut	Eng	Est	Fin	Fre	Ger	Gre	Hun	Ita	Lat	Lit	Mal	Pol	Por	Slo	Slo	Spa	Swe		
Czech	–	.	.	1	.	.	1	1	.	.	1	4	
Danish	.	–	1	1	
Dutch	.	.	–	6	.	.	2	1	9	
English	2	.	6	–	.	.	42	48	3	3	29	1	.	.	7	30	2	.	48	1	222	
Estonian	–	0	
Finnish	.	.	.	2	.	–	.	1	3	
French	1	.	2	38	.	.	–	22	3	.	9	.	.	.	1	5	.	.	10	.	91	
German	1	1	1	49	.	1	23	–	.	1	8	.	.	.	4	3	2	.	8	1	103	
Greek	.	.	.	2	.	.	3	.	–	5	
Hungarian	.	.	.	1	.	.	.	1	.	–	2	
Italian	1	.	.	25	.	.	9	8	.	.	–	.	.	.	1	3	.	.	7	.	54	
Latvian	.	.	.	1	–	1	
Lithuanian	–	0	
Maltese	–	0	
Polish	.	.	.	6	.	.	1	3	.	.	1	.	.	.	–	2	.	.	1	.	14	
Portuguese	.	.	.	25	.	.	4	4	.	.	3	.	.	.	1	–	.	.	6	.	43	
Slovak	.	.	.	1	.	.	.	1	–	.	.	.	2	
Slovene	–	.	.	0	
Spanish	1	.	.	42	.	.	8	7	.	.	7	.	.	.	1	6	.	.	–	.	72	
Swedish	.	.	.	2	.	.	.	1	–	3
	6	1	9	201	0	1	93	99	6	4	58	1	0	0	15	49	4	0	80	2		

EuroMatrix 2006-2009



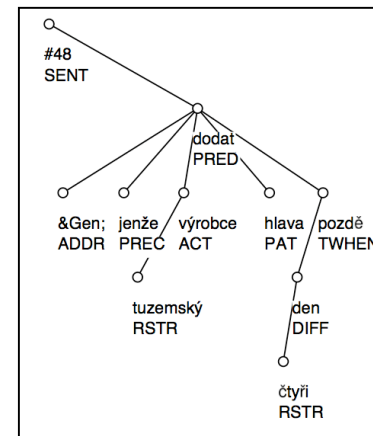
	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

using the Acquis corpus)

[from Koehn et al., 2009]

Approaches

- Statistical Phrase-Based SMT (+ factors)
- Hybrid: RBMT and SMT
- Linguistically-Rich SMT (Prague Dependency-Bank)



Achievements

- Moses PB-SMT
- Open source tools
- Training data
- Evaluation campaigns WMT
- MT Marathons
- ...



EuroMatrix 2006-2009



	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

using the Acquis corpus)

[from Koehn et al., 2009]



Lessons Learned:

- SMT struggles with
 - large divergence between languages (syntactic, word-order)
 - Rich morphology (target side)
- SMT performs well on in-domain data
- RBMT often better on out-of domain data

EuroMatrixPlus 2009-2012



Lessons Learned:





Objectives:

- Improving MT Quality
 - Hybrid statistical/rule-based
 - Tree-based (hierarchical, syntactic, tecto-grammatic)
 - Improved learning methods
- Open Research/Community
 - Open source tools
 - Evaluation campaign
 - MT Marathon



Objectives:

- Bringing Translation to the User
 - Professionals:
 - Localisation/Translation Industry
 - Individual translators
 - The Public:
 - Wiki translation

EuroMatrix 2006-2009



Partners

- University of Saarbrücken Germany
- University of Edinburgh UK
- Charles University Prague Czech Republic
- Johns Hopkins University USA
- Fondazione Bruno Kessler Italy
- Université du Maine, Le Mans France
- Dublin City University Ireland
- Lucy Software and Service Germany
- Central and Eastern European Translation Czech Republic



Evaluation WMT 2010:

- ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR
- Uppsala, Sweden, July 15th and 16th 2010
- Three tasks:
 - Translation: English, German, Spanish, French, Czech (into English and from English)
 - System Combination
 - MT Automatic Evaluation (**BLEU** ...)



Evaluation Results:

- Sneak Preview
- Not BLEU-scores
- Human Evaluation
- > 75,000 pair-wise comparisons (\Rightarrow ranking)
- \Rightarrow 153 MT systems

From English

- EN-CS 17

EM+: 1, 7, 8

- EN-DE 18

EM+: 3, 4, 9, ...

- EN-FR 19

EM+: 3, 7, ...

- EN-ES 16

EM+: 5, 6, ...

Into English

- ES-EN 14

EM+: 2

- FR-EN 24

EM+: 3

- CS-EN 12

EM+: 6, 7, 9

- DE-EN 25

EM+: 6, 8, 9, ...

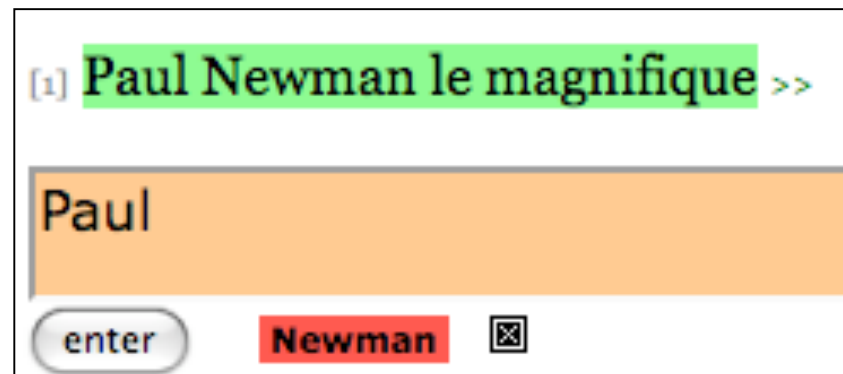


MT in the Localisation/Translation Industry:

- Integration of MT into Localisation Workflows
- MT/TM
- MT confidence scores \approx TM fuzzy match scores
- MT and mark-up
- Pricing MT
- Post-editing MT/TM output
- ...

Post-editing MT/TM output (I):

- Interactive/predictive MT



EuroMatrixPlus 2009-2012



Post-editing MT/TM output (II):

- Ranking word/phrase translations

Paul	Newman	le	magnifique
Paul	Newman	the	wonderful
Mr	Newman ,	the	magnificent
Mr Paul	Newman here	the	wonderful
as Paul	Committee		beautiful
another	Newman , who speaks		magnificent
with Paul		the	splendid
, Paul		the	excellent
of Paul		the	beautiful
work of Paul			it
the words of Paul			great

Post-editing MT/TM output (III):

- Tracking MT post-edits

<< [2] L'inoubliable interprète de "Butch Cassidy et le Kid" est mort des suites d'un cancer, à l'âge de 83 ans, dans sa maison du Connecticut. >>
The unforgettable ~~interpreter~~ actor of " Butch Cassidy and the Sundance Kid " died as a result of cancer 7 at the age of 83 ~~years~~ 7 in his house in Connecticut . (9 edits)

The unforgettable actor of "Butch Cassidy and the Sundance Kid" died as a result of cancer at the age of 83 in his house in Connecticut.

EuroMatrixPlus 2009-2012



Translation Tool translate - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://tool2.statmt.org/sentences/translate/563

Status Wiki Mail MgMail Edu News

Translation Tool pkoehn logout

Sentence 2 of 20 [1] [2] [4] [6] [8] [11] [13] [16] [19]

[1] Spitzen von Hamburger CDU und Grünen öffnen Weg zu Koalitionsverhandlungen. [2] Das erste schwarz-grüne Bündnis auf Landesebene rückt näher. Die Spitzen von CDU und Grünen in Hamburg halten ihre Differenzen für überwindbar. [3] In einer Sondiergrunde beschlossen sie, in den Parteigremien über den Start von Koalitionsverhandlungen zu beraten. [4] Hamburg - Sechs Stunden sprachen sie miteinander. [5] Dann verkündeten CDU-Chef Michael Freytag und Grünen-Chefin Anja Hajduk, das Trennende zwischen den Parteien sei überbrückbar.

[1] Leaders of the Hamburger CDU and Greens open path to coalition negotiations. [5] Then the CDU-leader Michael Freytag and Green party leader Anja Hajduk the division between the parties is bridgable.

<< [2] Das erste schwarz-grüne Bündnis auf Landesebene rückt näher: Die Spitzen von CDU und Grünen in Hamburg halten ihre Differenzen für überwindbar. >>

enter the first

das	erste	schwarz	@-@	grüne	Bündnis	auf	Landesebene	rückt	näher	:	die	Spitzen
the first	black	@-@	green	alliance	in favour of	is approaching	:	the leaders				
the first	black	@-@	green	the alliance	in favour	approaches	that the people at the top					
for the first	black		Green	Alliance	on	national	we are coming to	at the top				
this	in black and white	@-@	green	cooperation	in	Belarus approaches	the top					
the first of	the black	the Green	NATO	seek to	we	closer	the this					

EuroMatrix 2006-2009



Open Source

- Moses
<http://www.statmt.org/moses/>
- Joshua
<http://joshua.sourceforge.net/Joshua/Welcome.html>
- IRSTLM Language Modeling
<http://sourceforge.net/projects/irstlm/>
- Europarl
<http://www.statmt.org/europarl/>
- ...

EuroMatrixPlus 2009-2012



EM: <http://www.euromatrix.net/>

EM+: <http://www.euromatrixplus.net/>

EM++: <http://???>

Questions?

