7.2: Public Project Presentation and Updates

Philipp Koehn

Distribution: Public

EuroMatrix
Statistical and Hybrid Machine Translation
Between All European Languages
IST 034291 Deliverable 7.2

December 21, 2007

Project funded by the European Community
under the Sixth Framework Programme for
Research and Technological Development.

Information Society
Technologies

Project ref no.	IST-034291
Project acronym	EuroMatrix
Project full title	Statistical and Hybrid Machine Translation Between All European Languages
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 September 2006 / 30 Months

Distribution	Public
Contractual date of delivery	September 1, 2007
Actual date of delivery	September 13, 2007
Deliverable number	7.2
Deliverable title	Public Project Presentation and Updates
Type	Report, contractual
Status & version	Finished
Number of pages	19
Contributing WP(s)	WP7
WP / Task responsible	WP7 / 7.6
Other contributors	none
Author(s)	Philipp Koehn
EC project officer	Xavier Gros
Keywords	

The partners in EuroMatrix are:
Saarland University (USAAR)
University of Edinburgh (UEDIN)
Charles University (CUNI-MFF)
CELCT
GROUP Technologies
MorphoLogic

For copies of reports, updates on project activities and other EuroMatrix-related information, contact:
The EuroMatrix Project Co-ordinator
Prof. Hans Uszkoreit
Universität des Saarlandes, Computerlinguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
uszkoreit@coli.uni-sb.de
Phone +49 (681) 302-4115- Fax +49 (681) 302-4700

Copies of reports and other material can also be accessed via the project's homepage:
http://www.euromatrix.net/

EUROMATRIX: Statistical and hybrid machine translation between all European languages

A Project funded by the European Community under the Sixth Framework Programme (IST-5-034291-STP)

For the Europe of 23 languages:

the computer learns to translate

Information Society Technologies

The European Challenge

Existing MT systems for EU languages in 2005

	En	Ge	Fr	Sp	It	Po	Du	Gr	Sw	Fi	Po	Da	La	Cz	Hu	Sl	Ro	Bu	Sl	Ma	Li	Ir	Es
English		48	42	45	31	31	11	5	2	1	8	1	2	1	4	1	1	2	-	-	-	-	-
German	49		25	9	11	5	3	1	2	2	3	2	1	1	2	1	-	-	-	-	-	-	-
French	41	24		12	14	9	5	4	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-
Spanish	42	8	12		10	9	2	1	1	1	-	1	1	1	-	-	-	-	-	-	-	-	-
Italian	30	11	14	10		5	2	1	1	1	-	1	1	1	-	-	-	-	-	-	-	-	-
Portuguese	30	6	8	9	5		2	1	1	1	-	1	1	-	-	-	-	-	-	-	-	-	-
Dutch	11	3	5	2	2	2		1	1	1	-	1	1	-	-	-	-	-	-	-	-	-	-
Greek	4	1	4	1	1	1	1		1	1	-	1	-	-	-	-	-	-	-	-	-	-	-
Swedish	3	2	1	1	1	1	1	1		1	-	1	-	-	-	-	-	-	-	-	-	-	-
Finnish	3	2	1	1	1	1	1	1	1		-	1	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Danish	1	2	1	1	1	1	1	1	1	1	-		-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Statistical MT trained on EuroParl provides baseline coverage of 110 language pairs [Koehn 2005]

Other systems as enumerated in Compendium of Translation Software [Hutchins 2005]

Progress in machine translation

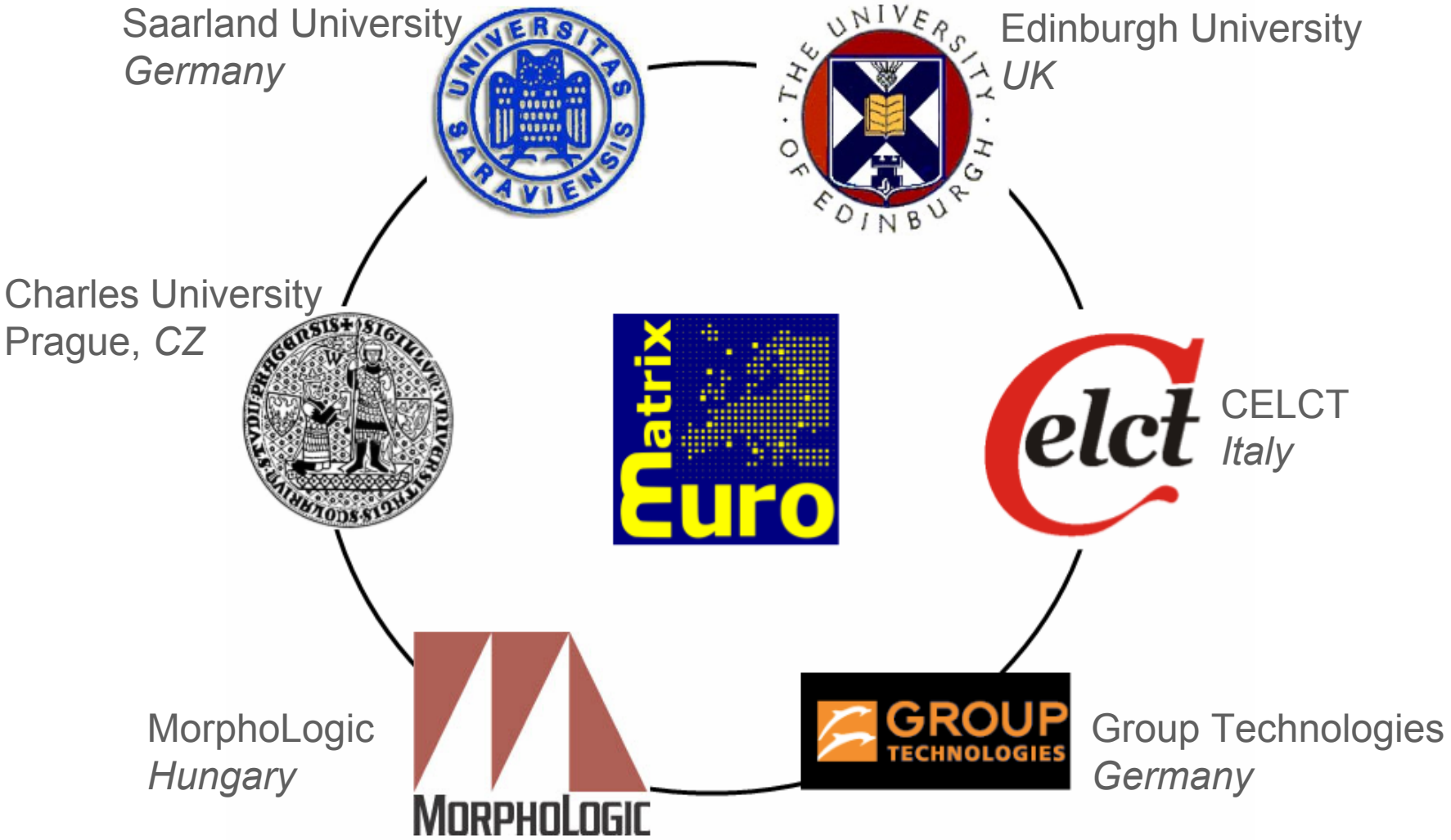
- **Data revolution**
  - **immense text resources** available in digital form (trillion of words)
  - large amounts of **translated text** become increasingly available (today 10-100s millions of words, maybe soon billions)
- **Statistical machine translation (SMT)**
  - development of **data-driven** statistical approach to MT
  - **competitive** with traditional approaches
- **Favorable research environment**
  - **US DARPA funding** for Arabic–English and Chinese–English SMT
  - **open** competitions, **sharing** of resources

however: limited academic research in European languages

Goals of the EuroMatrix Project

- Machine translation between **all EU language pairs**
  - 23 official EU languages, more to come →506..600 pairs
  - Baseline machine translation performance for all pairs
  - Starting point for national research efforts
  - More intensive effort on specific language pairs
- Creating an **open research** environment
  - Open source **tools** for baseline machine translation systems
  - Collection of open data **resources**
  - Open **evaluation campaigns** and **research workshops** ("marathons")
- Scientific **approaches**
  - **Statistical** phrase-based MT, extended by factored approach
  - **Hybrid** statistical/rule-based
  - Tree-transfer based on **tecto-grammatic** probabilistic models

EuroMatrix Consortium



EuroMatrix Objectives

- Translation systems for **all pairs of EU languages**, with a special focus on the languages of new and near-term prospective member states
- Efficient inclusion of **linguistic knowledge** into **statistical machine translation**
- The development and testing of **hybrid architectures** for the integration of rule-based and statistical approaches
- Organization, analysis and interpretation of a competitive annual **international evaluation of machine translation** with a strong focus on European economic and social needs
- The provision of **open source machine translation technology** including research tools, software and data
- A systematically compiled and constantly updated detailed **survey of the state of MT technology** for all EU language pairs

Research efforts in EuroMatrix

- **Statistical** phrase-based, extended by factored approach
  - builds on state-of-the art phrase-based approach
  - idea: add additional annotation at the word level (POS, morphology, ...)
  - effort centered at the University of Edinburgh
- **Hybrid** statistical/rule-based
  - integration of the Logos system with statistical methods
  - system combination / deep integration of components
  - effort centered at the University of Saarland, Saarbrücken
- Tree-transfer based on **tecto-grammatic** probabilistic models
  - based on long-term efforts, builds on parallel Czech–English treebank
  - transfer at the level of enriched dependency structures
  - effort centered at Charles University, Prague

Commitment to open resources

- A lot of **infrastructure** required to build a statistical MT system
  - parallel corpora
  - word alignment
  - language modeling
  - basic linguistic tools (tokenizers, taggers, morph. analyzers, parsers)
  - training statistical models
  - decoding

➔ MT systems become **too large** to be built completely by a small group

- **Sharing** of resources
  - avoids rebuilding the wheel everywhere
  - allows everybody to work on the state of the art
  - focus on novel solutions to current problems

Parallel Corpora

- **Europarl** corpus: proceedings of the European Parliament
  - Release of v3 in September 2007
  - 30-40 million word per language, all 11 official languages of the EU-15
- **News Commentary**: from http://www.project-syndicate.com/
  - Used in ACL WMT 2007 Shared Task
  - 1-2 million words in English, French, Spanish, German, Czech, Arabic, ...
- **Other** corpus projects
  - Acquis Communitaire: includes all 23 languages of EU-25
  - more data from European Union / European Commission

➔ good translation quality possible with this data

Moses: Open Source Toolkit

- **Open source** statistical machine translation system (developed from scratch 2006)
  - state-of-the-art **phrase-based** approach
  - novel methods: **factored translation models, confusion network decoding**
  - support for **very large models** through **memory-efficient** data structures

- Documentation, source code, binaries **available** at http://www.statmt.org/moses/
- Development also **supported by**
  - EC-funded **TC-STAR** project
  - **US** funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

Open evaluation

- Website hosted by the EuroMatrix project
- Machine translation for **all EU-25** languages
  - extending the matrix of MT systems to **23x22=506 language pairs**
  - information about **resources** for each language pair
  - **example output** to demonstrate translation performance
- **Ongoing evaluation** of translation quality
  - test sets from the last annual competition
  - **anybody can upload** their system's translations
  - **automatic** scoring with bleu, nist, meteor, and other metrics
  - facilitate **manual** evaluations (also a good **teaching tool**)
- Will go online early in 2008

EuroMatrix Events

- **Evaluation Campaign**
  - Dry-run March 2007, meeting at ACL 2007
  - First campaign March 2008, meeting at ACL 2008
  - Second campaign early 2009
- **Machine Translation Marathon**
  - April 2007 in Edinburgh: Summer school, research showcase
  - May 2008 in Berlin
    - Summer School
    - Open source conference
    - Research showcase
    - Early analysis of evaluation campaign results
- Conference "**Translingual Europe**", May 2008 in Berlin

  Inform invited representatives of industry, commerce, research and administration about recent progress in translation technology

Contact us:

http://www.euromatrix.net/

Hans USZKOREIT
Tel: +49-681-302-5282
Fax: +49-681-302-5338

euromatrix-coordinator@coli.uni-saarland.de