8.9: Annual Public Report

S. Busemann, O. Bojar, C. Callison-Burch, M. Federico, R. Garabik, J. van Genabith, P. Koehn, H. Schwenk, K. Simov, P. Wolf

Distribution: Public

EuroMatrixPlus
Bringing Machine Translation
for European Languages to the User

ICT 231720 Deliverable 8.9

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

Project ref no.	ICT-231720
Project acronym	EuroMatrixPlus
Project full title	Bringing Machine Translation for European Languages to the User
Instrument	STREP
Thematic Priority	ICT-2007.2.2 Cognitive systems, interaction, robotics
Start date / duration	01 March 2009 / 38 Months

Distribution	Public
Contractual date of delivery	November 15, 2010
Actual date of delivery	April 4, 2011
Date of last update	April 4, 2011
Deliverable number	8.9
Deliverable title	Annual Public Report
Type	Report
Status & version	Final
Number of pages	10
Contributing WP(s)	all
WP / Task responsible	DFKI
Other contributors	All Participants
Author(s)	S. Busemann, O. Bojar, C. Callison-Burch, M. Federico, R. Garabik, J. van Genabith, P. Koehn, H. Schwenk, K. Simov, P. Wolf
EC project officer	Michel Brochard
Keywords	

The partners in EuroMatrixPlus are:	DFKI GmbH, Saarbrücken (DFKI)
	University of Edinburgh (UEDIN)
	Charles University (CUNI-MFF)
	Johns Hopkins University (JHU)
	Fondazione Bruno Kessler (FBK)
	Université du Maine, Le Mans (LeMans)
	Dublin City University (DCU)
	Lucy Software and Services GmbH (Lucy)
	Central and Eastern European Translation, Prague (CEET)
	Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences (LSIL)
	Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS)

For copies of reports, updates on project activities and other EuroMatrixPlus-related information, contact:

The EuroMatrixPlus Project Co-ordinator
Prof. Dr. Hans Uszkoreit, DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
uszkoreit@dfki.de
Phone +49 (681) 85775-5282 - Fax +49 (681) 85775-5338

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

Chapter 1

General Remarks

This document describes the highlights of EuroMatrixPlus achieved during its second project year.

The open-source implementation Moses of factored statistical translation models (WP1) has continued to be used by many research teams and commercial users beyond the project, and the high volume of requests (and quick and helpful answers) on the Moses mailing list shows the popularity of this platform and its importance to the research community.

Several conferences and workshops co-organized by EuroMatrixPlus gave the project broad visibility within academia and industry. A major success was Translingual Europe 2010 co-organized with the META-NET project, where extensive interaction between scientists and industry took place. Slides are available from the programme at the conference homepage, and talks have been recorded and published on the Web.

The Second Joint EM+/CNGL Workshop ”Bringing MT to the User: Research on Integrating MT in the Translation Industry” was successful in bringing together MT researchers, developers, industrial users and translators to discuss issues that are most important in real world industrial settings involving MT, but currently not very popular in research circles.

The tradition to annually organize one-week MT marathons started in the predecessor project EuroMatrix was continued in 2010. The fifth MT marathon was held in August in Le Mans (France). It included lectures, talks and lab sessions, addressing a broad audience.

The project was amended to cover work on closely related languages (Czech and Slovak) and to investigate the usage of linguistic theories in combination with treebanks for MT. Two new work packages were added (WP9, and WP10), and two further partners, LSIL and IICT-BAS, joined the project.

The next chapter overviews the significant results achieved in the individual work packages during the reporting period. The output of more than 45 refereed publications demonstrates the impact of EuroMatrixPlus on machine translation.

Chapter 2

Project Progress

WP1: Rich Tree-Based Statistical Translation

Main achievements include the satisfactory progress in treebank annotation (structure completed, coreference completed in Czech, valency in progress), the improvements of the TectoMT platform[1], including the maximum-entropy dictionary in the deep-transfer MT system (Popel and Žabokrtský, 2010; Mareček et al., 2010) and the improvements in reordering of phrase-based and hierarchical models using source-side syntax (Bisazza and Federico, 2010; Bisazza et al., 2011).

We also believe our extended analysis of automatic and manual MT evaluation is of interest to the general MT community: explaining why BLEU fails in Bojar et al. (2010), interpreting post-edits and relating them to manual flagging of errors in Bojar (2011) and carrying out a complementary manual evaluation on the basis of question answering Berka et al. (2011).

WP2: Hybrid Machine Translation

How can stochastic methods be used to improve both the analysis and transfer phase of an existing RBMT system, in our case the Lucy RBMT system? An interface was developed by LUCY that allows access to so-called phrasal analyses which are constructed whenever the parser is not able to derive a full parse for an input sentence and DFKI implemented a mechanism to select the best analysis tree.

The results show that stochastic methods can be used to augment linguistic knowledge in an RBMT system, but they also stress the need for linguistically well-formed data. A study on bilingual transfer models has shown that an increase of translation quality can be achieved by crawling bilingual data for non-linguistic phrase-pairs, but this approach is outperformed by even shallow linguistic data structures such as terminology lists. Additionally, it presents a step toward easier domain adaptation for rule-based systems, as it is now possible to extract terminology lists from translation memories and other parallel data to improve the lexical coverage for limited application domains or usage scenarios.

WP3: Advanced Learning Methods for Machine Translation

The development and evaluation sets of the 2009 MT NIST evaluation campaign for the Arabic-to-English task have been translated into Italian and French by professional human translators and made available. Such resources allowed to measure the improvements on the SMT system for Arabic/Italian, an under resourced language pair. In particular, significant gains have been achieved by following two research directions, namely (i) pivot translation and (ii) exploitation of comparable corpora.

[1] http://ufal.mff.cuni.cz/tectomt/

Moreover, partners applied many of the results to systems developed for participating to open evaluation campaigns, like those organized by IWSLT10[2] and by the workshop on SMT of the ACL 2010[3].

Publications related to this WP are (Barrault, 2010; Lambert et al., 2010; Sanchis-Trilles and Cettolo, 2010; Shah et al., 2010), (Bisazza et al., 2010; Cettolo et al., 2010; Hardmeier and Federico, 2010; Rousseau et al., 2010; Zamora-Martínez et al., 2010; Schwenk, 2010; Bojar and Tamchyna, 2011; Abdul-Rauf and Schwenk, 2011)

WP4: Open Source Tools and Data

WP4 plays a strategic role in sharing research outcomes among the partners and disseminating them to the wider research community and the commercial world. The open source Moses toolkit is used by many research groups–in fact the paper describing it was the most highly cited paper in the ACL conferences in 2010[4].

We are also encouraged by the significant take-up by the commercial sector, as noted by the Translation Automation User Group (TAUS) in several articles on their web site[5].

WP5: "WikiTrans" Community-Based Translation Environments

Primary results for WP5 include

- A new training methodology for statistical machine translation, which includes new data structures and algorithms that allow for very rapid updating of the model, and which will facilitate quick inclusion of user suggestions of how to improve translations.
- Tools to allow monolingual users to interact with the translation models to produce translations that are much higher quality than through machine translation alone.

These results were published in the proceedings of peer-reviewed conferences and journals (Levenberg and Osborne, 2009; Levenberg et al., 2010; Koehn, 2010; Koehn, 2009a; Koehn and Haddow, 2009; Koehn, 2009b).

WP6: Integrated Localisation Workflow

- The Second Joint EuroMatrixPlus/CNGL Workshop (JEC 2010)[6] was a noted success, with the AMTA organisers offering to host the workshop again in future years. A workshop on industry-relevant MT related research questions with technical full size research papers (fully articulating research questions, data sets, experimental methods and evaluations) seems to have addressed a gap in the existing conference and workshop portfolio (either focusing on the NLP and MT science community or on business and governmental users with just short abstracts required) and was effective in "Bringing MT to the User".
- We developed a system integrating Machine Translation and Translation Memory techniques in order to provide complete high-quality translation for post-editing (Zhechev and van Genabith, 2010a; Zhechev and van Genabith, 2010b; Zhechev, 2010). This work has been timely and has inspired work reported in a number of other papers including (Koehn and Senellart, 2010) and (He et al., 2010).
- As far as we are aware our work on SPE for fully PB-SMT PE pipelines has produced the first positive result that this can indeed improve translation results.

[2] http://iwslt2010.fbk.eu/

[3] http://www.statmt.org/wmt10/

[4] http://linguification.wordpress.com/2010/12/29/paradigm-shift/

[5] http://www.tausdata.org/blog/2010/10/doing-business-with-moses-open-source-translation/
http://www.translationautomation.com/technology/will-there-be-a-thousand-moses-mt-systems.html

[6] http://web.me.com/emcnglworkshop/JEC2010/Home.html

WP7: Evaluation Campaign

In the first evaluation workshop, we performed a large-scale manual evaluation of 104 machine translation systems and 41 system combination entries. We used the ranking of these systems to measure how strongly automatic metrics correlate with human judgments of translation quality for 26 metrics. The proceedings of the workshop resulted in 16 publications on various scientific topics related to machine translation, along with 46 short papers describing the machine translation systems and automatic evaluation metrics that were submitted to the shared task.

The internal evaluation for the first year of the project was completed and successful. The internal evaluation for the second year is scheduled for the end of March 2011 to coincide with the public evaluation campaign. The results are summarized in the following table:

	Best System Score Year 0	Best System Score Year 1	Percent improvement
Czech → English	48%	63%	+15%
English → Czech	68%	75%	+7%
German → English	58%	67%	+9%
English → German	63%	62%	-1%
French → English	59%	67%	+8%
English → French	71%	79%	+8%

We satisfied criteria 1 except for the case of English → German, where we failed to see improvement (the drop in score is not statistically significant). All other language pairs showed considerable improvements, with cumulative average of 7.6% across all pairs.

WP8: Project Management and Dissemination

Dissemination activities were carried out by organizing

- Regular evaluation campaigns with the exchange of all the translation results and the results of human evaluations help to see and understand the differences and relative advantages of the different approaches and motivate all groups to engage in collaborations towards better, integrated approaches.
- Machine translation marathons, which form are a very effective instrument that helps to increase the mutual understanding of teams that address the MT problem from different backgrounds and work together on shared projects intensively for a week with subsequent collaborations for much longer time.

In the reporting period the project has co-organized several major events with participation from academia and industry:

- Translingual Europe 2010, June 2010, Berlin (Germany)[7]
- ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, July 2010, Uppsala (Sweden)[8]
- Fifth MT Marathon, September 2010, Le Mans (France)[9]
- Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry, November 2010, Denver (CO)[10]

[7] http://www.translingual-europe.eu/

[8] http://www.statmt.org/wmt10/

[9] http://lium3.univ-lemans.fr/mtmarathon2010/

[10] http://web.me.com/emcnglworkshop/JEC2010/Home.html

The Sixth Workshop on Statistical Machine Translation is scheduled for July 2011 Edinburgh (UK).

The project has generated 46 scientific publications in refereed conference proceedings, journals or books, some of which are cited at the end of this report.

WP9: Integrating Slovak Language Resources into the EUROMATRIXPLUS Framework

The main achievement in WP9 consists of the finalization of the Slovak-Czech parallel corpus, with about 720 000 sentence pairs (15 000 manually aligned). The Slovak-English parallel corpus contains significantly more texts than originally planned.

LSIL is organising the SLOVKO 2011 conference (Oct 20.22), which we will use to raise awareness of the project among the linguistic community in Slovakia.

WP10: HPSG-based Statistical Translation

We have a first version of Bulgarian Resource Grammar called BURGER[11]. The coverage of the grammar is being extended on the basis of the analyses within BulTreeBank[12], via transfer of the manual annotations of BulTreeBank for disambiguation of the parses within BURGER. The lexicon of the grammar is extended on the basis of the morphological lexicon available to us and the creation of the valency lexicon which we are constructing on the basis of BulTreeBank.

The format and a procedure for transfer of linguistic knowledge encoded in BulTreeBank to the BURGER format is reported in (Osenova and Simov, 2010).

A first version of a web service for the Bulgarian language pipeline is implemented.

[11] Published on `http://www.delph-in.net/`

[12] `http://www.bultreebank.org/`

References

Sadaf Abdul-Rauf and Holger Schwenk. 2011. Parallel Sentence Generation from Comparable Corpora for improved SMT. *Machine Translation.* Accepted for publication.

Loïc Barrault. 2010. Many: Open source mt system combination at wmt10. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 271–275.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95, March.

Arianna Bisazza and Marcello Federico. 2010. Chunk-Based Verb Reordering in VSO Sentences for Arabic-English Statistical Machine Translation. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 241–249, Uppsala, Sweden.

Arianna Bisazza, Ioannis Klasinas, Mauro Cettolo, and Marcello Federico. 2010. FBK @ IWSLT 2010. In *IWSLT*, pages 53–58, Paris, France.

Arianna Bisazza, Daniele Pighin, and Marcello Federico. 2011. Chunk-Lattices for Verb Reordering in Arabic-English SMT. *Machine Translation: Special Issue on MT for Arabic.* under minor revision for publication.

Ondřej Bojar and Aleš Tamchyna. 2011. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar. 2011. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95, March.

Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. 2010. Mining Parallel Fragments from Comparable Texts. In *IWSLT*, pages 227–234, Paris, France.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *IWSLT*, pages 283–289, Paris, France.

Yifan He, Yanjun Ma, Andy Way, and Josef van Genabith. 2010. Integrating n-best smt outputs into a tm system. In *Proceedings of the 23rd International Conference on Computational Linguistics (CoLing10): Poster Volume*, pages 374–382. Beijing, China.

Philipp Koehn and Barry Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *MT Summit XII.*

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In Ventsislav Zhechev, editor, *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC10)*, pages 21–31. Denver, CO.

Philipp Koehn. 2009a. A process study of computed aided translation. *Machine Translation*, 23(4):241–263.

Philipp Koehn. 2009b. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore, August. Association for Computational Linguistics.

Philipp Koehn. 2010. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the*

Association for Computational Linguistics, pages 537–545, Los Angeles, California, June. Association for Computational Linguistics.

Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. 2010. Lium smt machine translation system for wmt 2010. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 127–132.

Abby Levenberg and Miles Osborne. 2009. Stream-based randomised language models for SMT. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 756–764, Singapore, August. Association for Computational Linguistics.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 394–402, Los Angeles, California, June. Association for Computational Linguistics.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–201, Uppsala, Sweden. Association for Computational Linguistics.

Petya Osenova and Kiril Simov. 2010. Using the linguistic knowledge in bultreebank for the selection of the correct parses. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, pages 163–174, Tartu, Estonia, December.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, pages 293–304.

Anthony Rousseau, Loïc Barrault, Paul Delglise, and Yannick Estève. 2010. LIUM's statistical machine translation systems for IWSLT 2010. In *IWSLT*, pages 113–117.

German Sanchis-Trilles and Mauro Cettolo. 2010. Online Language Model adaptation via N-gram Mixtures for Statistical Machine Translation. In *Conference of the European Association for Machine Translation (EAMT)*, Saint-Raphal, France.

Holger Schwenk. 2010. Adaptation d'un systme de traduction automatique statistique avec des ressources monolingues. In *TALN*, page in press.

Kashif Shah, Loïc Barrault, and Holger Schwenk. 2010. Translation model adaptation by resampling. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 392–399.

Francisco Zamora-Martínez, María José Castro-Bleda, and Holger Schwenk. 2010. N-gram-based machine translation enhanced with neural networks for the French-English BTEC-IWSLT'10 task. In *IWSLT*, pages 45–52.

Ventsislav Zhechev and Josef van Genabith. 2010a. Maximising tm performance through subtree alignment and smt. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA-10)*. Denver, CO.

Ventsislav Zhechev and Josef van Genabith. 2010b. Seeding statistical machine translation with translation memory output through tree-based structural alignment. In Dekai Wu, editor, *Proceedings of the Fourth Workshop on Syntax and Structure in Statistical Translation (SSST-4)*, pages 43–51. Beijing, China.

Ventsislav Zhechev. 2010. Highlighting matched and mismatched segments in translation memory output through sub-tree alignment. In *Proceedings of the Translating and the Computer Conference 2010 (T&C-10)*. London, UK.