LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored MT

www.letsmt.eu

Project no. 250456

Deliverable D7.5.2
LetsMT! seminars and workshops

Version No. 0.8
28/10/2011

Document Information

Deliverable number:	D7.5.2
Deliverable title:	LetsMT! seminars and workshops
Due date of deliverable according to DoW:	31/10/2011
Actual submission date of deliverable:	31/10/2011
Main Author(s):	Marko Tadić
Participants:	Andris Karpovs
Reviewer	
Workpackage:	WP7
Workpackage title:	Dissemination and sustainability
Workpackage leader:	FFZG
Dissemination Level:	PU
Version:	V0.8
Keywords:	seminars and workshops

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.5	20/09/2011	draft	FFZG	Marko Tadić	
V0.8	28/10/2011	completed for review	FFZG	Marko Tadić	
V0.9	31/10/2011	updated	Tilde	Inguna Skadiņa, Andris Karpovs	
V1.0	31/10/2011	Final	FFZG	Marko Tadić	Final

EXECUTIVE SUMMARY

With this deliverable a list of conferences and workshops that have been co-organized in collaboration with LetsMT! project is given as the second part of deliverable D7.5.

Table of Contents

1 Introduction

Dissemination of information about the projects is one of activities that rises the awareness about the project, its goals, achievements, partners involved as well as funding part(ies). Although it does not produce tangible results that are produced by the main research activities of the project, the Dissemination and sustainability work package (WP7) with its results that convey information about the project is considered as important as the other WPs.

One of goals of WP7 is to **disseminate** project results and to transfer the project knowledge, technologies, lessons learned and best practices to interested communities and thus to ensure their impact and long-term sustainability.

In a research and professional communities, conferences, workshops and seminars represent the most straightforward way to personally spread information. It includes both, unidirectional and bidirectional, i.e. interactive communication when it is needed. Apart from participating at such events organised by other parties, LetsMT! project is about to organise at least two international workshops and a seminar with relevant target groups, usually in association with major international conferences.

This deliverable D7.5.2 follows the previous D7.5.1 and gives a list of conferences and workshops that have been co-organised by LetsMT! since.

2 Conferences coorganised in cooperation with LetsMT!

2.1 FLaReNet Forum 2011

By creating a consensus among major players in the field, the mission of FLaReNet is to identify priorities as well as short, medium, and long-term strategic objectives and provide consensual recommendations in the form of a plan of action for EC, national organisations and industry. Through the exploitation of new collaborative modalities as well as workshops and meetings, FLaReNet will sustain international cooperation and (re)create a wide Language community. In this respect the FLaReNet Forum 2011 was organised in Venice (26-27 May 2011) and LetsMT! was between coorganising projects. This also gave the opportunity to Andrejs Vasiļjevs to present the project goals and achievements at the conference under the title *How to get more data for under-resourced languages and domains?* (see more at http://project.letsmt.eu/index.php?p=video-lectures where the whole presentation is accessible as video lecture). The conference URL is http://www.flarenet.eu/?q=FLaReNet_Forum_2011.

2.2 NooJ2011 conference

The main aim of this conference is to provide a forum for the sharing of new ideas and recent advances in human language processing by concentrating on the NooJ development environment. The conference intends to:

- give NooJ users and researchers in linguistics and in computational linguistics the opportunity to meet and share their experience as developers, researchers and teachers;

- present to NooJ users the latest linguistic resources and NLP applications developed for/with NooJ, its latest functionalities, as well as its future developments;

- offer researchers and graduate students two tutorials (one basic and one advanced) to help them parse corpora and build NLP applications using NooJ;

- provide the occasion to present and discover the recent developments of NooJ itself.

Since the NooJ can also offer environment to develop machine translation, it was interesting to investigate how this combines with LetsMT! pure SMT and more applicative approach. The LetsMT! project was presented with a poster.

This year the conference was organised in Dubrovnik (13-15 June 2011). The conference is the successor of a series of successful conferences held in Barcelona (2007), Budapest (2008), Tozeur (2009) and Komotini (2010). The conference URL is http://lt.ffzg.hr/nooj2011/.

2.3 SlaviCorp2011

On 27 August 2009 in Stara Lesna, the International Committee of Slavists established the Corpus Linguistics Committee, following a motion tabled by representatives of a number of Slavic national corpora. The proponents of the motion included members of major corpus linguistics teams from Bulgaria, Croatia, Czech Republic, Poland, Russia, Slovakia, Slovenia and Ukraine. The topics of the conference include, but are not limited to the following:

- recent developments in national Slavic corpora;

- new Slavic corpus projects;

- parallel Slavic corpora;

- Slavic corpora and corpus based lexica;

- diachronic Slavic corpora and digital text collections;

- annotation in Slavic corpora;

- applications of Slavic corpora: research, teaching, LT, M(A)T etc.

This year conference also included participation from Baltic languages, so Lithuanian was present also, opening the space for other Baltic languages to joint at the future conferences. Since better part of Slavic languages could be considered under-resourced in the respect of existing MT systems, LetsMT! project, presented by a poster, appeared a platform where they could find the opportunity to reduce their under-resourcedness, or could find an appropriate usage of parallel corpora where some Slavic language take part in language pair.

The conference URL is http://hnk.ffzg.hr/slavicorp2011.

3 LetsMT! Presentations at the Conferences and Workshops

3.1 Multilingual Web Pisa Workshop

Multilingual Web Pisa Workshop is one of the four workshops that survey and share information about currently available best practices and standards that can help content creators and localizers address the needs of the multilingual Web, including the Semantic Web. They also provide an important opportunity to identify gaps that need to be addressed. The workshop is also designed as an opportunity for participants to network and share information between and across the various different communities involved in enabling the multilingual Web.

Andrejs Vasiļjevs gave a presentation titled Bridging technological gap between smaller and larger languages describing motivation behind LetsMT! project and its key results.

3.2 EAMT 2011

The European Association for Machine Translation (EAMT) organises its yearly conferences regularly in May. This year the venue was Faculty of Arts, Katholieke Universitet Leuven in Belgium. This two day conference started on 2011-05-30 and on 2011-05-31. At the conference Raivis Skadiņš and Andrejs Vasiļjevs presented work done in LetsMT! project on evaluation of SMT impact on localization work under title *Evaluation of SMT in localization to under-resourced inflected language*. In the conference there was also an exhibition of European projects related to machine translation. Since this audience is considered to be a natural lieu for LetsMT! services, our poster and flyers were presented there by Raivis Skadiņš (Tilde).

Paper in proceedings:

Skadiņš R., Puriņš M., Skadiņa I., Vasiļjevs A., Evaluation of SMT in localization to under-resourced inflected language, in Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, p. 35-40, May 30-31, 2011, Leuven, Belgium

3.3 Localization World Barcelona 2011

The Localization Word 2011 Conference took place in Barcelona, Spain, on 14-16 June 2011. It is a conference and networking organization dedicated to the language and localization industries. This year, it gathered more than 550 industry professionals from all over the world. LetsMT! was presented with its poster and leaflets and attracted the conference participants' interest. One of the main focus points at the Conference was also the recent development in the MT field as well as MT experiences at the leading IT companies.

3.4 META-FORUM 2011

The central disseminating event of the META-NET community in 2011, META-FORUM 2011 took place in Budapest on 2011-06-27 and 2011-06-28. Within this two days of conference densly packed with a number of parallel actions -- oral presentations, poster presentations, software demonstrations, an exhibition of European projects was organised. The LetsMT! project poster (and leaflets) was presented and attracted a lot of interest from researchers and companies.

3.5 MT Summit XIII

The LetsMT! Platform was demonstrated at MT Summit XIII, the largest event MT community. MT Summit XIII was organized by the International Association for Machine Translation (IAMT) and hosted by the Asia-Pacific Association for Machine Translation (AAMT) and Chinese Information Processing Society of China (CIPS). It was held in Xiamen, China, from 19 to 23 September 2011. The MT Summit brings together all members of the MT community: users, developers, and researchers - from around the world. The goal of the Summit conferences is to enable the members of the MT community to learn about advances in research and product development, as well as to hear about the ways in which MT is being used successfully in various environments. Participants share stories of progress they have fostered, and can assess the status of technology and deployment internationally. LetsMT! Platform was presented in system presentation session by Raivis Skadiņš under title *LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines*. The *LetsMT!* platform has received great interest from conference participants, especially practitioners in this field.

Paper in proceedings:

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2011). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In Proceedings of the 13th Machine Translation Summit (pp. 507-511). Xiamen, China

3.6 TAUS user conference 2011

The TAUS User Conference is the once-a-year, non-sponsored event focused on translation automation, localization business innovation and industry collaboration. At this conference leading practitioners are tackling the most pressing issues and opportunities facing the global industry and sharing the best practice in translation technologies and strategies.

Andrejs Vasiļjevs gave presentation titled "Technologies for smaller languages" at the Call for Ideas session where he demonstrated live how LetsMT can train MT system from user provided data.

3.7 Localization World Silicon Valley 2011

Localization World gathers representatives from companies interested in production or sales in international markets as well as experts in software, social media, advertising, marketing and publishing, specialists in localization, international web development and sales. Localization World Silicon Valley delivers comprehensive education and networking opportunities for both beginners and experienced professionals involved in multinational growth.

Indra Samite presented LetsMT! at the system demonstration session of the conference. Andrejs Vasiļjevs presented work on LetsMT! at the panel session "MT: Free ride is over". Tilde provided continuous demonstrations of the LetsMT! Platform at the conference exhibition booth.

4   Workshops (co)organised in cooperation with LetsMT

4.1 Joint workshop with project ACCURAT at LREC2012 conference

On 2011-10-24 a proposal for a joint workshop with the FP7 project ACCURAT was submitted at the major conference in the field – LREC2012. The general aim of the workshop is to bring together the experts in language resources and machine translation in order to tackle the topic of lack of parallel resources in building the MT systems when it comes to less-resourced languages (LRL) and/or domains (LRD).

The proposal for the workshop is in Appendix A of this deliverable. We expect the acceptance of this workshop proposal and successful presentation of LetsMT! to the targeted audience that represents the research community consisting of corpus and computational linguists primarily.

4.2 Joint workshop with project ACCURAT at GALA conference

In the time of writing this deliverable, a contact has been established with the GALA (Globalization and Localization Association) and organisers of their annual conference GALA2012 that will be held in Monaco (26-28 March 2012). They offered us one of their predefined formats for presentations (90 minutes slot, round table, panel, demonstration etc.) or a separate organised event in the larger form of a workshop. At this moment we are in the process of negotiations and most probably we will opt for a separate workshop format.

The conference URL is http://www.gala-global.org/conference.

5 Conclusions

With this deliverable a list of conferences and workshops that have been or are planned to be co-organized in collaboration with LetsMT! project is given as the second part of deliverable D7.5.

Appendix A

LREC 2012 Workshop Proposal Submission Form

Workshop Title *[Required]* **:**

Language Resources for Machine Translation in Less-Resourced Languages and Domains

Workshop Description *[Required]* **:**

The general aim of the workshop is to bring together the experts in language resources and machine translation in order to tackle the topic of lack of parallel resources in building the MT systems when it comes to less-resourced languages (LRL) and/or domains (LRD).

Workshop will be jointly organised and supported by two projects: FP7 project ACCURAT and ICT-PSP project LetsMT!, both dealing with this issue from two different angles.

ACCURAT is investigating how to use data from comparable corpora when there are not enough parallel data for LRL or LRD. The results of the project achieved so far will be presented to the wide audience and thoroughly discussed.

LetsMT! is building a platform for building tailor-made SMT systems based on your own parallel data from either general or specialised domains. This platform will also feature open access to SMT systems that are being build on public data. The scientific and technological issues will be presented and discussed.

At the workshop two to three invited speakers are expected, while the rest of the programme will consist of contributions collected through the CFP.

Motivation and Topics of interest *[Required]* **:**

Lack of sufficient language resources for many languages and domains is currently one of the major obstacles in further advancement of machine translation. The main goal of this workshop is to present what has been investigated in this area in two main directions: 1) usage of comparable corpora where there is not enough parallel data (like in less-resourced languages and domains); 2) reaching out for hidden parallel data. These two directions are largely covered by two projects, ACCURAT and LetsMT! that are supported by FP7 and ICT-PSP programmes respectively.

The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of parallel language resources to improve MT quality significantly for under-resourced languages and narrow domains, such as Latvian, Croatian, Greek and renewable energy domain, medical domain, ICT domain. This novel methodology and fully functional model for exploiting comparable corpora will contribute not only to MT, but also to other tasks in LT such as alignment (sentence, phrase or word level), bi- or multilingual lexicon extraction, multilingual NE detection etc.

The main challenge of LetsMT! project lie in the fact that SMT systems largely depend on the size of training data. Since the majority of parallel data is in more-resourced languages, SMT systems for these languages are of much better quality compared to systems for less-resourced languages. Current systems are built on data accessible on the web, but it is just a fraction of all parallel texts. Most of them still reside in the local systems of different corporations, public and private institutions, and desktops of individual users. The cost and the know-how required for building custom MT solutions deter many small-to-medium companies from utilizing the power of MT technologies. LetsMT! project is building an innovative online collaborative platform for data sharing and MT building. This platform will support upload of public as well as proprietary MT training data and building of multiple MT systems, public or proprietary, by combining and prioritizing this data. How it could be done is a challenge of its own regarding the processing of parallel corpora and their further usage.

Topics of interest:
- comparable corpora usage
- comparable corpora processing tools/kits

- parallel corpora usage
- parallel corpora processing tools/platforms
- MT for less-resourced languages
- MT for less-resourced domains
- open source SMT systems (Moses, etc.)
- publicly available SMT
- public MT systems related IPR issues

Summary of the Call :

LREC2012 Workshop: Language Resources for Machine Translation in Less-resourced Languages and Domains

The first Call for Papers

Lack of sufficient language resources for many languages and domains currently is one of the major obstacles in further advancement of machine translation. The main goal of this workshop is to present what has been investigated in this area in two main directions: 1) usage of comparable corpora where there is not enough parallel data (like in less-resourced languages and domains); 2) reaching out for hidden parallel data.

We are inviting new, unpublished contributions that will present the findings, analysis and evaluation of novel methods how comparable corpora can compensate for this shortage of parallel language resources or how a larger quantity and richer parallel data that are not available over web could be collected from hidden sources. The primary aim of usage of these two approaches is to improve MT quality significantly for under-resourced languages and narrow domains, but resources and techniques developed for this purpose could also serve other LT tasks such as multilingual dictionary extraction, multilevel alignment, multilingual NE detection, multilingual information extraction etc.

Topics of interest of the workshop are, but not limited to:
- comparable corpora usage
- comparable corpora processing tools/kits
- parallel corpora usage
- parallel corpora processing tools/platforms
- MT for less-resourced languages
- MT for less-resourced domains
- open source SMT systems (Moses, etc.)
- publicly available SMT
- public MT systems related IPR issues

The papers will be peer reviewed and accepted ones will be included in the LREC2012 proceedings.

Important dates:

Call	for	papers:	20	November,	2011
Submission	of	papers:	10	January,	2012
Notification	of	acceptance:	10	February,	2012

Workshop: May 25, 2012

Estimated Audience *[Required]* : 30-40 participants
Duration of the workshop *[Required]* : 1 day
Tentative Schedule :

9:00-9:30 Opening, introduction

Session 1
9:30-10:30   keynote speaker 1

10:30-11:00	coffee break

Session 2

11:00-11:20	speaker 1
11:20-11:40	speaker 2
11:40-12:00	speaker 3
12:00-12:20	speaker 4
12:20-12:40	speaker 5

12:40-14:30	lunch break

Session 3

14:30-15:30	keynote speaker 2

15:30-16:00	coffee break

Session 4

16:00-16:20	speaker 6
16:20-16:40	speaker 7
16:40-17:00	speaker 8
17:00-17:20	speaker 9
17:20-17:50	general discussion
17:50-18:00	closing of the workshop

In the case of more applications the contributions could be shortened to 15 min and/or working time could be adapted.

Technical Requirements :

Lecturing room for ca 40 participants

LCD projector

Internet access

Contact Person *[Required]* **:** Aivars Bērziņš

Email Address of the Contact Person *[Required]* **:** aivars.berzins@Tilde.lv

Organizing Committee *[Required]* **:**

Ahmet Aker

Aivars Berzinš

Andreas Eisele

Rob Gaizauskas

Nikos Glaros

Lene Offersgaard

Serge Sharoff

Inguna Skadina

Raivis Skadinš

Marko Tadić

Jörg Tiedemann

Dan Tufiş
Andrejs Vasiljevs
Oliver Wilson