ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D6.10
Workshop report

Version No. 1.0
29/06/2012

Document Information

Deliverable number:	D6.10
Deliverable title:	Workshop report
Due date of deliverable:	30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Marko Tadić
Participants:	Božo Bekavac, Željko Agić, Nikola Ljubešić
Internal reviewer:	Tilde
Workpackage:	WP6
Workpackage title:	Dissemination
Workpackage leader:	FFZG
Dissemination Level:	PU
Version:	V1.0
Keywords:	dissemination activities, workshop

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.5	05/06/2012	Draft	FFZG	Marko Tadić	First draft
V0.6	25/06/2012	Draft	FFZG	Marko Tadić, Božo Bekavac, Željko Agić, Nikola Ljubešić	Reordering of some paragraphs
V0.7	29/06/2012	Draft	FFZG	Marko Tadić	Corrected minor typing errors

EXECUTIVE SUMMARY

In this deliverable the report on organisation and rolling out of two ACCURAT workshops is presented. The first one was targeting translation and localisation industry and it was organised within GALA2012 conference in Monte Carlo, on 2012-03-25. The second workshop was organised as one of LREC2012 post-conference workshops in Istanbul, on 2012-05-26. With these workshops ACCURAT specifically targeted for dissemination two different communities, industry and research respectively.

This deliverable encompasses Deliverable D6.5 which was postponed from M15.

Table of Contents

1 Introduction

The overall goal of WP6 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 worldwide impact and long-term sustainability. Also, one of the goals of WP6 is to provide insight into exploitation plans of the tools developed within the ACCURAT project.

This deliverable D6.10 is the report on two ACCURAT workshops that were organised in order to bring closer the targeted audience and disseminate information about the ACCURAT project during the last four months at the most important events in the field.

2 Purpose of ACCURAT workshops

According to the implementation of dissemination activities described in the Dissemination and Exploitation Plan (D6.1) two workshops had to be organised by the ACCURAT consortium in order to come closer to targeted audiences.

Following the stratification of targeted audience from D6.1, the first workshop was aiming to spread the information about the project to the localisation and translation industry, while the second was aiming at scientific community where the research results of the project would be presented, discussed and communicated in the manner that would allow two way communication, i.e., much needed feedback in the form of questions and answers, arguing about problematic spots etc.

We believe that the dissemination instruments described in this deliverable have significantly raised the profile and visibility of the project, as a complement to other dissemination channels used.

3 GALA2012 workshop

3.1 *Target audience*

Targeted audience for this workshop was localisation and translation industry, i.e. companies and individuals that might become potential users of ACCURAT results, and particularly the ACCURAT Toolkit for comparable corpora. To maximise the outreach to this kind of audience and after careful consideration at which event our dissemination effort would yield the most successful effect, we decided to contact the organisers of the GALA2012 conference that was held in Monte Carlo, from 26 to 28 March 2012.

GALA conferences are considered to be the top-most global events in this industry and in the end it turned right to organise a workshop in that context and with that kind of audience.

GALA conferences usually do not have satellite events, but this time our negotiations were successful, so the organisers allowed us to organise a half-day workshop under their name and in the same premises. Even more, they also supported our workshop with their logistics, so the whole organisation resulted with a successful event.

3.2 *Co-organisation*

The workshop was organised jointly with the LetsMT! ICT-PSP project (http://www.letsmt.eu). The general aim of the workshop was to demonstrate, among other things, the ACCURAT Toolkit for extracting parallel data from comparable corpora to enhance both, SMT and RBMT systems. Also, several use cases were presented how to use machine translation in different environments, such as translation, authoring and localisation. In order to accommodate both project needs, the workshop was given a title *Customized Machine Translation: Platform, Tools and Application LetsMT! cloud platform and ACCURAT tools*.

The workshop took place on 25 March 2012 with more than 30 registered participants among which were members of LT development teams of important players in the files such as PayPal and Adobe.

3.3 *Workshop programme*

The agenda of the workshop was the following:

09:00 Andrejs Vasiļjevs (Tilde): *The Quest for Better MT*

09:10 Achim Ruopp (Digital Silk Road): *Modern Ubiquitous Machine Translation – Threat or Opportunity?*

09:35 Indra Sāmīte (Tilde): *LetsMT! and ACCURAT at Your Service*

09:50 Marko Tadić (Univ. of Zagreb, FFZG): *ACCURAT Toolkit = More Data*

10:10 Raivis Skadiņš (Tilde) *LetsMT! Do-it-Yourself Demo*

10:30 coffee break

10:50 Use cases

- Gregor Thurmair (Linguatec): *Creating Lexicon Entries for Narrow Domains from Comparable Corpora*
- Mateja Verlič (Zemanta): *Using SMT in the Blogging Environment*
- Andrejs Vasiļjevs (Tilde): *Real World Evaluation of SMT in Localization*

11:40 Panel discussion *Customized MT: Truth or Myth?*

- Marko Tadić (moderator)
- Achim Ruopp
- Andrejs Vasiļjevs
- Indra Sāmīte
- Gregor Thurmair

12:30 End of the workshop

The panel discussion involved members of the audience and it turned into an interesting discussion about the possibility to use MT in everyday tasks, particularly for under-resourced languages and narrow domains. It looked like these topics provoked a lot of interest as it seems this is the next direction the translation and localisation industry is going to take, after achieving certain level of industrialisation for major languages.

The original list of speakers featured also a translation expert from EC DG Translation, but this person was hard to reach in desired time, so we remained with Dr. Achim Ruopp as the only invited speaker.

Figure 1 GALA2012 ACCURAT workshop panel discussion

4 The 5th BUCC workshop

4.1 *Target audience*

The target audience for the second workshop was research communities in the fields of computational linguistics, corpus linguistics, language technologies and natural language processing. The LREC2012 was recognised as a conference that can accommodate all our needs: European and global impact, considerable size, possibility to organise a satellite workshops, well positioned in the community, good timing in the respect to the project duration, etc. The decision to organise the whole day workshop at LREC2012 was made in early 2011 and preparations for it were going since mid-2011.

4.2 *Co-organising*

ACCURAT project responded to the call for workshop proposals and sent the first proposal. However, due to extremely large number of proposals sent to the LREC2012 organisers, it was conditionally accepted providing that it gets merged with the proposal for 5th Workshop on Building and Using Comparable Corpora. From the point of view of the organisers and having in mind the closeness of topics that both proposals covered, we had to agree in order to be present at the most important event of the community. The first proposal, the merged proposal and the final Call for Papers are attached as Annexes to this deliverable.

The joint proposal was accepted and the workshop was given a title *The Fifth Workshop on Building and Using Comparable Corpora (5th BUCC) with special topic Language Resources for Machine Translation in Less-Resourced Languages and Domains*, to accommodate the needs of the ACCURAT and other projects involved in organisation. The joint proposal catered for the situation that there are several EU-funded project dealing with the usage of parallel and comparable corpora running at the moment, and that for all of them such a workshop could be useful. Therefore, this joint proposal involved a number of projects as co-organising entities: ACCURAT, HyghTra, LetsMT!, TTC, while other similar project were invited to present their results that were relevant for the general topic of the workshop.

Two ACCURAT partners were selected in the organising committee of the workshop: Marko Tadić (FFZG) and Andrejs Vasiļjevs (Tilde), while Marko Tadić was also editor in chief of the workshop Proceedings:

(http://www.lrec-conf.org/proceedings/lrec2012/workshops/16.BUCC2012%20Proceedings.pdf).

The usual process of collecting contributions to the workshop started with a call for papers that was issued by organising committee and followed by blind reviewing process with at least two reviewers accepting the contribution. The program and proceedings were completed in time for publishing with the rest of LREC2012 materials.

4.3 *Workshop programme*

The 5th BUCC workshop was organised as the LREC2012 whole day post-conference workshop:

09:00 – 09:10 **Opening**

Oral Presentations 1: Multilinguality (Chair: Pierre Zweigenbaum)

09:10 – 09:30 Philipp Petrenz, Bonnie Webber: *Robust Cross-Lingual Genre Classification through Comparable Corpora*
09:30 – 09:50 Qian Yu, François Yvon, Aurélien Max: *Revisiting sentence alignment algorithms for alignment visualization and evaluation*

Invited Projects Session (Chair: Serge Sharoff)
09:50 – 10:10 Inguna Skadiņa: *Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation* (ACCURAT, http://www.accurat-project.eu)
10:10 – 10:30 Andrejs Vasiļjevs: *LetsMT! – Platform to Drive Development and Application of Statistical Machine Translation* (LetsMT!, http://www.letsmt.eu)

10:30 – 11:00 **Coffee Break**

Invited Project Session (Contd.)
11:00 – 11:20 Núria Bel, Vassilis Papavasiliou, Prokopis Prokopidis, Antonio Toral, Victoria Arranz: *Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform* (PANACEA, http://panacea-lr.eu)
11:20 – 11:40 Adam Kilgarriff, George Tambouratzis: *The PRESEMT Project* (PRESEMT, http://www.presemt.eu)
11:40 – 12:00 Béatrice Daille: *Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite* (TTC, http://www.ttc-project.eu)

12:00 – 12:30 **Panel Discussion with Invited Speakers**

12:30 – 14:00 **Lunch Break**

Oral Presentations 2: Building Comparable Corpora (Chair: Reinhard Rapp)
14:00 – 14:20 Aimée Lahaussois, Séverine Guillaume: *A viewing and processing tool for the analysis of a*
comparable corpus of Kiranti mythology
14:20 – 14:40 Nancy Ide: *MultiMASC: An Open Linguistic Infrastructure for Language Research*

Booster Session for Posters (Chair: Marko Tadić)
14:40 – 14:45 Elena Irimia: *Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair*
14:45 – 14:50 Iustina Ilisei, Diana Inkpen, Gloria Corpas, Ruslan Mitkov: *Romanian Translational Corpora: Building Comparable Corpora for Translation Studies*
14:50 – 14:55 Angelina Ivanova: *Evaluation of a Bilingual Dictionary Extracted from Wikipedia*
14:55 – 15:00 Quoc Hung-Ngo, Werner Winiwarter: *A Visualizing Annotation Tool for Semi-Automatical Building a Bilingual Corpus*
15:00 – 15:05 Lene Offersgaard, Dorte Haltrup Hansen: *SMT systems for less-resourced languages based on domain-specific data*
15:05 – 15:10 Magdalena Plamada, Martin Volk: *Towards a Wikipedia-extracted Alpine Corpus*
15:10 – 15:15 Sanja Štajner, Ruslan Mitkov: *Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness*
15:15 – 15:20 Dan Ştefănescu: *Mining for Term Translations in Comparable Corpora*
15:20 – 15:25 George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, Marina Vassiliou: *Accurate phrase alignment in a bilingual corpus for EBMT systems*
15:25 – 15:30 Kateřina Veselovská, Ngãy Giang Linh, Michal Novák: *Using Czech-English Parallel Corpora in Automatic Identification of* It
15:30 – 15:35 Manuela Yapomo, Gloria Corpas, Ruslan Mitkov: *CLIR- and Ontology-Based Approach for Bilingual Extraction of Comparable Documents*

15:35 – 16:30 **Poster Session and Coffee Break** (coffee from 16:00 – 16:30)

Oral Presentations 3: Lexicon Extraction and Corpus Analysis (Chair: Andrejs Vasiļjevs)
16:30 – 16:50 Amir Hazem, Emmanuel Morin: *ICA for Bilingual Lexicon Extraction from Comparable Corpora*
16:50 – 17:10 Hiroyuki Kaji, Takashi Tsunakawa, Yoshihoro Komatsubara: *Improving Compositional Translation with Comparable Corpora*
17:10 – 17:30 Nikola Ljubešić, Špela Vintar, Darja Fišer: *Multi-word term extraction from comparable corpora by combining contextual and constituent clues*
17:30 – 17:50 Robert Remus, Mathias Bank: *Textual Characteristics of Different-sized Corpora*
17:50 – 18:00 **Wrapup discussion and end of the workshop**

4.4 ACCURAT presentations

Apart from the invited talk in which the ACCURAT project was generally presented, and which was given by Inguna Skadiņa under title *Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation*, ACCURAT partners were present with three more additional presentations (one oral presentation and two posters):

- Nikola Ljubešić, Špela Vintar, Darja Fišer: *Multi-word term extraction from comparable corpora by combining contextual and constituent clues* (oral presentation)
- Elena Irimia: *Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair* (poster)
- Dan Ştefănescu: *Mining for Term Translations in Comparable Corpora* (poster)

It is general opinion that the ACCURAT project partners presented new, original and valuable results and that they contributed in spreading project results in clear and acceptable manner, thus raising the project visibility and notability.

Named Entity Mapping

- After the NEs are identified in the source and target texts, a mapping procedure outputs a list of entity pairs that are mutual translations
- ACCURAT developed a generic NE mapper that uses string similarity measures to determine which NEs map:
  - cosine similarity, LCS, DICE, Levenshtein distance, etc.
- The final score is a linear interpolation of these measures

Figure 2 Inguna Skadiņa presenting ACCURAT project

5 Conclusion

In this deliverable a detailed description of two workshops that ACCURAT project organised was given, the role of ACCURAT partners explained and dissemination effect detailed. As a conclusion it can be said that these dissemination instruments have conveyed the information about the project, contributed to rising of the awareness of its achievements and to present its results to the targeted audience: localisation and translation industry and research community.

Appendix A: GALA2012 workshop announcement

Customized Machine Translation: Platform, Tools and Application
LetsMT! cloud platform and ACCURAT tools
Sunday, 25 March 09:00 - 12:00
Le Méridien Beach Plaza, Zephyr Ballroom, 4th Floor

Presenters: Andrejs Vasiļjevs (Tilde), Indra Sāmīte (Tilde), Marko Tadić (University of Zagreb), Raivis Skadiņš (Tilde), Gregor Thurmair (Linguatec), Mateja Verlič (Zemanta), representative of DG Translation, European Commission.

This free workshop is for localization industry practitioners who want to stay on top of developments in the machine translation (MT) field. Learn how to make MT work for you and increase your translation productivity, provide faster service to your customers, and give your company a competitive edge.

You will see innovative technologies focused on empowering MT users and enabling MT for less-resourced languages. These are the results of two large-scale pan-European university and industry collaboration projects LetsMT! and ACCURAT, supported by EU R&D FP7 and ICT-PSP Programs.

Come and learn how to benefit from the self-service “MT factory” in the cloud at letsmt.com where you can build customized MT systems. ACCURAT tools will help you to obtain much-needed parallel data by collecting it from the multilingual web. We will also present specific applications for custom MT, such as sentiment analysis of multilingual newsfeeds, use of SMT systems for authoring tools, enhancing rule-based MT systems with data from comparable corpora, and use of MT in real world localization processes, with a particular focus on under-resourced languages.

You will have the opportunity to meet and learn from top European MT researchers, as well as business practitioners who use MT in translation on a daily basis. More info at the workshop webpage.

Special offer: Register, come and get an exclusive free 2 month trial to make your custom MT system.

You can register for this event when you register for GALA 2012. If you have already registered for the conference, you can simply return to the online store and add this free session to your registration. If you need assistance, please contact Allison Ferch.

Appendix B Text of the first LREC2012 workshop proposal

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 December, 2011
Submission of papers: 15 February, 2012

Notification of acceptance: 10 March, 2012
Workshop: May 26, 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 Bērziņš
Andreas Eisele
Rob Gaizauskas
Nikos Glaros
Lene Offersgaard
Serge Sharoff
Inguna Skadiņa
Raivis Skadiņš
Marko Tadić
Jörg Tiedemann
Dan Tufiş
Andrejs Vasiljevs
Oliver Wilson

Appendix C: Text of the LREC2012 workshop proposal merged with the proposal for the 5th BUCC workshop

LREC 2012 Workshop Proposal Submission Form

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

5th WORKSHOP ON BUILDING AND USING COMPARABLE CORPORA

Special Theme: “Language Resources for Machine Translation in Less-Resourced Languages and Domains”

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

Given the positive reception of the four previous editions of the "Workshop on Building and Using Comparable Corpora" which took place at LREC 2008, ACL-IJCNLP 2009, LREC 2010, and ACL-HLT 2011 we would like to propose to continue the series at LREC 2012. As the focus of the workshop is on a particular type of resource, co-locating it with the leading international conference on language resources appears to be a perfect fit.

The workshop aims to bring together language engineers as well as linguists interested in the constitution and use of comparable corpora. Research on comparable corpora is active but used to be scattered among many workshops and conferences. Hence this workshop series bundling this research and giving it a better platform.

As in many cases people feel a need to go beyond easily reachable parallel corpora and as in the past few years considerable progress has been made in the field, many international projects emphasizing work on comparable corpora and usage of “hidden” parallel data have been recently conducted or are still going on, among them ACCURAT, LetsMT!, Panacea, PreseMT, TTC, Kelly, HyghTra, and Monotrans.

At the previous editions of the workshop invited speakers were Kevin Knight, Ken Church, Adam Kilgarriff, and Serge Sharoff, and panellists included Ken Church, Andreas Eisele, Pascale Fung, Hitoshi Isahara, Kyo Kageura, Adam Kilgarriff, Kevin Knight, Franz Och, Uwe Quasthoff, Richard Sproat, and Benjamin Tsou, all confirming the interest in this field at their home institutions and encouraging us to continue with the workshop series. The workshop has been endorsed by ACL SIGWAC and FlareNet, and is intended to be endorsed by META-NET.

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

In the language engineering and the linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Interest in non-parallel forms of comparable corpora in language engineering primarily ensued from the scarcity of parallel corpora. This has motivated research concerning the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Non-parallel yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a non-parallel corpus requires control over the selection of source texts in both languages.

Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English-French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. To exemplify such issues in a practical setting, this year's special focus will be on "Language Resources for Machine Translation in Less-Resourced Languages and Domains" with the aim of overcoming the shortage of parallel resources when building MT systems for less-resourced languages and domains, particularly by usage of comparable corpora for finding parallel data within and by reaching out for "hidden" parallel data. Lack of sufficient language resources for many language pairs and domains is currently one of the major obstacles in further advancement of machine translation.

We plan an invited panel where the coordinators of five major projects will present their work.

We solicit contributions including but not limited to the following topics:

Topics related to the special theme:

* 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

Building Comparable Corpora

* Human translations
* Automatic and semi-automatic methods
* Methods to mine parallel and non-parallel corpora from the Web
* Tools and criteria to evaluate the comparability of corpora
* Parallel versus non-parallel corpora, monolingual corpora
* Rare and minority languages
* Across language families

* Multi-media/multi-modal comparable corpora

Applications of Comparable Corpora

* Human translations
* Language learning
* Cross-language information retrieval & document categorization
* Bilingual projections
* Machine translation
* Writing assistance

Mining from Comparable Corpora

* Extraction of parallel segments or paraphrases from comparable corpora
* Extraction of bilingual and multilingual translations of single words and multi-word expressions; proper names, named entities, etc.

Summary of the Call :
(See the final version of Call for Papers in the next section)

Estimated Audience *[Required]* **:**
Participation has been solid with more than 50 registered participants e.g. at LREC 2010 and at ACL-HLT 2011. As this year we expect additional participation from the members of the invited projects, our estimate is that registration numbers will further increase.

Duration of the workshop *[Required]* **:**
1 day (full day)

Tentative Schedule :
09:00 Session 1: Building and using comparable corpora. [5 presentations]
10:40 Coffee break
11:00 Session 2: Comparable Corpora for MT in less-resourced languages and domains [5 presentations]
13:00 Lunch break
14:00 Session 3: Identifying word translations in comparable corpora [5 presentations]
15:10 Coffee break
15:30 Session 4: Collaborative platforms for data sharing and MT system building [5 presentations]
17:30 Session 5: Evaluation of comparable corpora [5 presentations]
18:30 End 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

Depending on the room situation, a clipable microphone for the speakers and two handheld microphones for the audience (during discussion) and the panel would be helpful.

Contact Person *[Required]* **:**

Reinhard Rapp (Universities of Leeds and Mainz)

Marko Tadić (University of Zagreb)

Email Address of the Contact Person *[Required]* **:**

reinhardrapp@gmx.de

marko.tadic@ffzg.hr

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

Reinhard Rapp

Marko Tadić

Serge Sharoff

Andrejs Vasiljevs

Pierre Zweigenbaum

Programme Committee *[If available]* **:**

The core of our reviewers has been rather stable over the years, so typically most of the past PC members are willing to serve again. In the following list members who already confirmed for 2012 are marked with an asterisk, other confirmations are pending.

* Srinivas Bangalore (AT&T Labs, USA)

* Caroline Barrière (National Research Council Canada)

* Chris Biemann (Microsoft / Powerset, San Francisco, USA)

* Lynne Bowker (University of Ottawa, Canada)

* Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)

* Kurt Eberle (Lingenio, Heidelberg, Germany)

* Andreas Eisele (DFKI, Saarbrücken, Germany)

* Rob Gaizauskas (University of Sheffield, UK)

* Éric Gaussier (Université Joseph Fourier, Grenoble, France)

* Nikos Glaros (ILSP, Athens, Greece)

* Gregory Grefenstette (Exalead/Dassault Systemes, Paris, France)

* Silvia Hansen-Schirra (University of Mainz, Germany)

* Hitoshi Isahara (NICT, Tokyo, Japan)

* Kyo Kageura (University of Tokyo, Japan)
* Adam Kilgarriff (Lexical Computing Ltd, UK)
* Natalie Kübler (Université Paris Diderot, France)
* Philippe Langlais (Université de Montréal, Canada)
* Tony McEnery (Lancaster University, UK)
* Emmanuel Morin (Université de Nantes, France)
* Dragos Stefan Munteanu (Language Weaver Inc., USA)
* Lene Offersgaard (University of Copenhagen, Denmark)
* Reinhard Rapp (University of Tarragona, Spain)
* Sujith Ravi (Yahoo! Research, USA)
* Serge Sharoff (University of Leeds, UK)
* Michel Simard (National Research Council Canada)
* Inguna Skadiņa (Tilde, Riga, Latvia)
* Jörg Tiedemann (University of Uppsala, Sweden)
* Dan Tufis (Romanian Academy, Bucharest, Romania)
* Oliver Wilson (University of Edinburgh, UK)
* Michael Zock (LIF, CNRS Marseille, France)
* Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

Appendix D: The final Call for Papers of the merged LREC2012 workshop

The final CfP was issued on 2011-12-23 to the following mailing lists:

- corpora@uib.no
- bionlp@lists.ccs.neu.edu
- lr_egroup@mail.iiit.ac.in
- nodali@helsinki.fi
- acl@aclweb.org
- sentproc@lists.qc.cuny.edu
- sigsemitic@cs.um.edu.mt
- sigann@cs.vassar.edu
- ln-request@cines.fr
- researchers@pascal-network.org
- slavicorp@chopin.ipipan.waw.pl
- slavicling@utlists.utexas.edu

The text of the Cfp follows:

Apologies for multiple postings
Please distribute to colleagues

==

5th WORKSHOP ON BUILDING AND USING COMPARABLE CORPORA

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

Co-located with LREC 2012
Lütfi Kirdar Istanbul Exhibition and Congress Centre
Saturday, 26 May 2012

DEADLINE FOR PAPERS: 15 February 2012

http://hnk.ffzg.hr/5bucc2012

Endorsed by
* ACL SIGWAC (Special Interest Group on Web as Corpus)
* FLaReNet (Fostering Language Resources Network)

==

MOTIVATION

In the language engineering and the linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, it is chiefly motivated by the need to use comparable corpora as training data for statistical

NLP applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

The scarcity of parallel corpora has motivated research concerning the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Non-parallel yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a non-parallel corpus requires control over the selection of source texts in both languages.

Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English-French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. To exemplify such issues in a practical setting, this year's special focus will be on

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

with the aim of overcoming the shortage of parallel resources when building MT systems for less-resourced languages and domains, particularly by usage of comparable corpora for finding parallel data within and by reaching out for "hidden" parallel data. Lack of sufficient language resources for many language pairs and domains is currently one of the major obstacles in further advancement of machine translation.

TOPICS

We solicit contributions including but not limited to the following topics:

Topics related to the special theme:

* comparable corpora use in MT
* comparable corpora processing tools/kits for MT
* 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

Building Comparable Corpora:

* Human translations
* Automatic and semi-automatic methods
* Methods to mine parallel and non-parallel corpora from the Web
* Tools and criteria to evaluate the comparability of corpora
* Parallel vs non-parallel corpora, monolingual corpora
* Rare and minority languages
* Across language families
* Multi-media/multi-modal comparable corpora

Applications of comparable corpora:

* Human translations
* Language learning
* Cross-language information retrieval & document categorization
* Bilingual projections
* Machine translation
* Writing assistance

Mining from Comparable Corpora:

* Extraction of parallel segments or paraphrases from comparable corpora
* Extraction of bilingual and multilingual translations of single words and multi-word expressions; proper names, named entities, etc.

IMPORTANT DATES (TENTATIVE)

15 February 2012	Deadline for submission of full papers
10 March 2012	Notification of acceptance
20 March 2012	Camera-ready papers due
26 May 2012	Workshop date

SUBMISSION INFORMATION

Papers should follow the LREC main conference formatting details (to be announced on the conference website http://www.lrec-conf.org/lrec2012/) and should be submitted as a PDF-file of no more than ten pages via the START workshop manager: https://www.softconf.com/lrec2012/BUCC2012/ Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings.

Double submission policy: Parallel submission to other meetings or publications are possible but must be immediately notified to the workshop organizers.

When submitting a paper through the START page, authors will be asked to provide information about the resources that have been used for the work described in their paper or are an outcome of their research. For details on this initiative, please refer to http://www.lrec-conf.org/lrec2012/?LRE-Map-2012.
Authors will also be asked to contribute to the Language Library, the new initiative of LREC 2012.

For further information, please contact
Reinhard Rapp reinhardrapp (at) gmx (dot) de
or Marko Tadic marko.tadic (at) ffzg (dot) hr

ORGANISERS

Reinhard Rapp, Universities of Mainz (Germany)and Leeds (UK)
Marko Tadic, University of Zagreb (Croatia)
Serge Sharoff, University of Leeds (UK)
Andrejs Vasiljevs, Tilde SIA, Riga, Latvia
Pierre Zweigenbaum, LIMSI, CNRS, Orsay, and ERTIM, INALCO, Paris (France)

SCIENTIFIC COMMITTEE

* Srinivas Bangalore (AT&T Labs, USA)
* Caroline Barrière (National Research Council Canada)
* Chris Biemann (Microsoft / Powerset, San Francisco, USA)
* Lynne Bowker (University of Ottawa, Canada)
* Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
* Andreas Eisele (DFKI, Saarbrücken, Germany)
* Rob Gaizauskas (University of Sheffield, UK)
* Éric Gaussier (Université Joseph Fourier, Grenoble, France)
* Nikos Glaros (ILSP, Athens, Greece)
* Gregory Grefenstette (Exalead/Dassault Systemes, Paris, France)
* Silvia Hansen-Schirra (University of Mainz, Germany)
* Kyo Kageura (University of Tokyo, Japan)
* Adam Kilgarriff (Lexical Computing Ltd, UK)
* Natalie Kübler (Université Paris Diderot, France)
* Philippe Langlais (Université de Montréal, Canada)
* Tony McEnery (Lancaster University, UK)
* Emmanuel Morin (Université de Nantes, France)
* Dragos Stefan Munteanu (Language Weaver Inc., USA)
* Lene Offersgaard (University of Copenhagen, Denmark)
* Reinhard Rapp (Universities of Mainz, Germany, and Leeds, UK)
* Sujith Ravi (Yahoo! Research, Santa Clara, CA, USA)
* Serge Sharoff (University of Leeds, UK)
* Michel Simard (National Research Council Canada)
* Inguna Skadina (Tilde, Riga, Latvia)
* Monique Slodzian (INALCO, Paris, France)
* Benjamin Tsou (The Hong Kong Institute of Education, China)
* Dan Tufis (Romanian Academy, Bucharest, Romania)
* Justin Washtell (University of Leeds, UK)
* Michael Zock (LIF, CNRS Marseille, France)
* Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)