

### 3.1 Summary



[www.mantra-project.eu](http://www.mantra-project.eu)

MANTRA (Multilingual Annotation of Named Entities and Terminological Resource Acquisition) is a STREP project that will provide multilingual terminologies and semantically annotated multilingual documents, e.g., patent texts, to improve the accessibility of scientific information from multilingual documents. Two SMEs are project partners and are exploiting the resources to improve the quality and functionality of their product offerings.

The MANTRA project capitalizes on parallel document corpora from which translational correspondences will be computed by the use of different alignment methods. Fortunately, the biomedical domain - the application domain for MANTRA - offers a rich variety of such parallel corpora. The multilingual document sets will be exploited to harvest terms and concept representations in different languages in order to augment currently available terminological resources such as the Medical Subject Headings (MeSH).

As part of the project work the project partners have prepared three corpora:

- Patent claims in en, de, fr (~120,000 units each)
- Medline abstract titles in en, de, fr, es, and nl where about 1.6 mio English titles are available in one of the other languages predominantly German (~719,000) or French (~570,000), and
- EMEA documents in en, de, fr, es and nl (~365,000 units each)

All units have been standardized using publicly available formats to make them available to the research community.

The available multilingual terminological resources (MeSH, Snomed-CT, MedDRA) have been processed and analysed for their suitability for the Mantra challenge. A subset of the public distribution has been made accessible through a special access point.

The project partners themselves have used different annotation solutions to prepare the first silver standards and to test the first evaluation methods for the Mantra challenge. In particular, two evaluation measures have been identified:

- (A) Measuring the identified non-English mentions from the challenge contributions against a mono-lingual SSC based on the contributions from the challenge participants
- (B) Measuring the identified CUIs against the English SSC that has been produced from the project partners

The Mantra corpora were available from January 2013 to May 2013. The challenge participants could deliver their annotations till May 31<sup>st</sup>, when the challenge was closed. In total, seven teams contributed to the challenge and delivered in total 66 annotated files in all languages. The project partners that were amongst the challenge participants delivered solutions that did not exist before the challenge, i.e. all the contributions to the challenge were based on novel solutions.

Spanish was the most popular non-English language followed by French and German. The submission numbers for es, de and fr were sufficiently large to generate the SSCs for the evaluation, whereas the submission numbers for nl do not allow an evaluation that is unbiased by individual systems due to the lower number of submissions. The evaluation has been performed in July to September 2013 and the results have been presented at the CLEF-ER workshop in Valencia (22-26 September 2013). The results do not lead to a single solution that serves as a clear front-runner solution, but to different systems that showed advantages in the different evaluation forms.

One result is that for the mention identification (evaluation A) those systems performed better that either translated the English terms into non-English representations or used multi-lingual resources of different kinds to suggest and then apply term candidates against the non-English corpus. For the CUI assignment in non-English corpora (evaluation B) the machine-translation solutions showed better performance.

In addition to the project work linked to the Mantra challenge and the preparation of the resources, the SME project partners are preparing their proprietary IT solutions and products for the use of multilingual resources. The prototype development is on-going, and the SME project partners have presented preliminary solutions in the Mantra review meeting.

The project partners have also established collaborative links with key researchers in the domain of multilingual medical terminological resources:

- Dr. Oliver Bodenreider, National Library of Medicine, Washington DC, USA (for English)
- Prof. Pierre Zweigenbaum, LIMSI, Paris, France (for French)
- Prof. Stefan Schulz, Universität Graz, Austria (for German)
- Dr. Ronald Cornet, University of Amsterdam, The Netherlands (for Dutch)
- Dr. Laura Furlong, Lluís Padró (for Spanish)

The project partners have aligned their initiatives with the CLEF framework leading to the result that the Mantra workshop has been organized as part of the CLEF conference in Valencia (22-26 September 2013). The Mantra challenge has been labelled as CLEF-ER challenge and CLEF-ER workshop, has been advertised as part of the CLEF conference, the requirements of the CLEF conference with regards to publication submissions and peer review have been fulfilled, the language chairs have been invited to the CLEF workshop, and further collaborative benefits have been worked out during the workshop with the CLEF-eHealth initiative.