### Project acronyms:
**MANTRA**

### Project full title:
Multilingual Annotation of Named Entities and Terminology Resources Acquisition

### Proposal/Contract no.:
FP7 – 296410 (FP7-ICT-2011-SME-DCL)

### Project Duration:
July 01, 2012 – June 30, 2014

### Website:
www.mantra-project.eu

---

### Deliverable D5.1

**“Challenge Outcome”**

---

### Project coordinator:
University of Zürich  
Dr. Dietrich Rebholz-Schuhmann  
Binzmühlestrasse 14  
8050 Zürich  
Tel: +41 44 63 54 30  
Mail: rebholz@ifi.uzh.ch

### Authors:
Udo Hahn, Johannes Hellrich, Dietrich Rebholz-Schuhmann

### Date of preparation:
15-Nov-2013 (Version 1.0)

### Dissemination level:
Protected
Table of Contents

1. Summary ........................................................................................................................................ 3
2. Introduction ..................................................................................................................................... 4
3. Challenge setup .............................................................................................................................. 5
4. Participation .................................................................................................................................... 6
   Participants ...................................................................................................................................... 6
   Overview on the used solutions ........................................................................................................ 7
5. Evaluation ......................................................................................................................................... 9
   Preparation of the non-English SSCs ............................................................................................... 9
   Performing the evaluation .................................................................................................................. 11
   Mono-lingual mention evaluation (Evaluation A) ............................................................................ 11
   Cross-lingual concept evaluation (Evaluation B) ............................................................................ 12
   Harmonised means of evaluation A+B (Evaluation AB) ................................................................. 12
   Mono-lingual de-annotated mention evaluation (evaluation A/-t) .................................................. 12
6. Results ............................................................................................................................................... 13
   Best overall systems according to the F-measure ............................................................................ 13
   Assessment of the FP, FN, TP result across all solutions ............................................................... 14
   F-measure across the different semantic groups ............................................................................ 18
7. Evaluation A ..................................................................................................................................... 21
   Different measures ........................................................................................................................... 21
   F-measures across semantic groups ............................................................................................... 23
8. Evaluation B ..................................................................................................................................... 26
   Different measures ........................................................................................................................... 26
   F-measures across semantic groups ............................................................................................... 27
9. Other evaluations ............................................................................................................................. 30
   Combining evaluation A and evaluation B ....................................................................................... 30
   Harmonised means of evaluation A+B (Evaluation AB) ................................................................. 31
   Evaluation A/-t ................................................................................................................................. 34
   Correlation of challenge contribution solutions .......................................................................... 35
10. Conclusion and Final Remarks ..................................................................................................... 37
11. References ....................................................................................................................................... 39
1. Summary

The “Cross-Language Evaluation Forum” (CLEF) has hosted the “CLEF entity recognition” workshop (CLEF-ER). This workshop has been setup to discuss the results from the CLEF-ER challenge that has been organized by the Mantra project partners to improve entity recognition (ER) in parallel multilingual document collections. It brings together researchers from the domains of entity recognition in the biomedical domain, normalization of entity mentions, and machine translation for the challenges linked to the identification of concepts in languages where only limited support through specialized terminologies is given.

A large set of documents in different languages, i.e. Medline titles, EMEA drug label documents and patent claims, have been prepared to enable ER in parallel documents. Each set of documents forms a corpus-language pair (CLP), for example the full set of Medline titles in German is the “EMEA/de” CLP, and the number of documents for each CLP vary from about 120,000 for patents up to 760,000 for Medline titles. The challenge participants have been asked to annotate entity mentions with concept unique identifiers (CUIs) in the documents of their preferred non-English language.

The main task is concerned with the attribution of CUIs to entities in the non-English corpora. The challenge participants could make use of a prepared terminological resource for entity normalization and of our English silver standard corpora as an input for concept candidates in the parallel non-English documents.

Several evaluation measures have been applied to determine the best performing solutions against the different CLPs. The F-measure over all annotations was used as the most basic means and was complemented with average F-measures across the semantic groups linked to the annotations. Furthermore, assessments were executed that determined the best ranked solutions across all CLPs and for the different corpora only. The evaluations were based on the entity recognition in the non-English languages (called “evaluation A”) and on the assignment of the correct concept unique identifiers (CUIs) in the non-English CLPs, where the CUIs have been compared against an English Silver Standard Corpus (SSC, called “evaluation B”).

In our assessment we came to the conclusion that on average the task A has been solved at a higher performance level than the task B. Furthermore, performance levels were lower for the patent CLPs in comparison to the Medline CLPs and the EMEA CLPs. We can also read from the results that lexicon-based or terminology-focused solutions perform better in the task A (solutions D1 and E1), where solutions based on statistical machine translation and machine learning perform better in task B (solutions F1 and G1, apart from D1).

It is necessary to take into consideration that task A makes use of the challenge participants contributions to generate the SSC and thus performances of the challenge participants tend to be higher, whereas the evaluation of task B is only based on the pre-annotations from the Mantra project partners and thus measures the challenge performances in a more neutral way. None of the solutions from the pre-annotations (i.e. the Mantra project partners) did contribute to the Mantra challenge to avoid any biases.
2. Introduction

Challenges form an important means in the scientific community to help improving innovative technologies in biomedical data analysis: e.g. different CLEF challenges such as CLEFeHealth and CLEF-IP [Catarci et al., 2012; Roda et al., 2010], the BioCreative sequel [Krallinger et al., 2008; Morgan et al., 2008], the BioNLP Shared Tasks [Cohen et al., 2009], and the CALBC challenge [Rebholz-Schuhmann et al., 2010; Rebholz-Schuhmann et al., 2011]. Most challenges propose a gold standard corpus (GSC) that is then used for the benchmarking of the proposed solutions. In addition, further challenges have been organized that have prepared and used a silver standard corpus (SSC) instead, i.e. the CALBC challenge I and II [Rebholz-Schuhmann et al., 2010].

The SSC approach is novel in the sense that it

1. makes use of large-scale corpora in contrast to corpora of limited size,
2. generates the annotations in the corpora with automatic means by harmonizing the pre-annotations from different tagging solutions,
3. may even use different harmonization schemes to install different characteristics into the SSC to enable alternative evaluation schemes.

The CLEF-ER challenge is unique in the sense that it combines different expectations and technologies, such as entity recognition in the biomedical domain with multilingual approaches and machine translation. Furthermore, the CLEF-ER challenge anticipates the processing and management of large resources and will exploit the delivered results for the development of augmented terminological resources.

The following publications are already available for background information:

- Overview paper on the execution of the CLEF-ER challenge [Rebholz-Schuhmann et al., 2013a].
- A publication that reports on the MANTRA resources used in the CLEF-ER challenge [Rebholz-Schuhmann et al., 2013b].
- Methodological papers on the calculation of the centroids and the modifications of the centroids for the CLEF-ER challenge [Lewin et al., 2012; Lewin et al., 2013]
- A publication on preliminary results in the generation of a gold standard corpus [Kors et al., 2013].
- A publication on the preliminary results of the challenge participants against the SSC. [Rebholz-Schuhmann et al., 2013c].
3. Challenge setup

The challenge material has been prepared in Q4’12 and in the beginning of 2013. After the English SSC was available and the non-English corpora have been standardized and aligned – using the Mantra Annotate (MAN) and Mantra Align (MAL) files (see deliverable D3.1) – the challenge could be opened. The following timelines have been pursued for the challenge rollout (deadlines):

- Subselection of the UMLS terms / ontologies  
  [21 Jan 2013]

- In total 14 annotated corpora have been prepared: three corpora in three languages (en, fr, de, 9 CPLs – for Corpus Language Pairs) + two corpora in Spanish and Dutch (4 CPLs)
  In total: 13 annotated document sets  
  [21 Jan 2013]

- Closing of the challenge  
  [31-May-2013]

The English SSC has been generated from the six annotated corpora that have been produced with automatic means by the six project partners. The methods for the generation of the SSC are described in [Lewin et al., 2012; Lewin et al., 2013]. The voting threshold has been set to an agreement of 3 solutions on the term location and on a boundary threshold of at least 2 for the left and for the right boundary.
4. Participation

Seven groups participated in the CLEF-ER challenge (see table 3.1 and 4.1) and contributed annotated corpora for the evaluation.

Participants

The participants mainly are from central Europe (Germany, France, Italy, Netherlands, Spain, UK, Switzerland) and the USA. In total, the participants are linked to 10 different sites. All participants are listed in the following overview:

- Giuseppe Attardi, Andrea Buzzelli, Daniele Sartiano
  Dipartimento di Informatica, Università di Pisa Italy
  {attardi, buzzelli, sartiano}@di.unipi.it

- Rafael Berlanga (a), Antonio Jimeno-Yepes (b), María Pérez-Catalán (a)
  (a) Department of Languages and Computer Systems Universitat Jaume I, Castelló, Spain
  (b) National ICT Australia
  {berlanga,maria.perezg}@uji.es, antonio.jimeno@gmail.com

- Andreea Bodnari (a,b), Aurélie Névéol (b), Ozlem Uzuner (a,c), Pierre Zweigenbaum
  (b), Peter Szolovits (a):
  (a) MIT, CSAIL, Cambridge, Massachusetts, USA;
  (b) LIMSI-CNRS, rue John von Neumann, F-91400 Orsay, France;
  (c) Department of Information Studies, University at Albany, SUNY, Albany, New York, USA

- Simon Clematide, Martin Davtyan, Fabio Rinaldi, and Dietrich Rebholz-Schuhmann
  University of Zurich, Institute of Computational Linguistics, Binzmuehlestr. 14, 8050
  Zurich, Switzerland
  {siclemat,davtyan,rinaldi,rebholz}@cl.uzh.ch

- Johannes Hellrich and Udo Hahn
  Jena University Language & Information Engineering (JULIE) Lab, Friedrich-
  Schiller-Universit¨at Jena, Jena, Germany
  johannes.hellrich@uni-jena.de, udo.hahn@uni-jena.de

- Georgios Kontonatsios, Ioannis Korkontzelos, and Sophia Ananiadou
  National Centre for Text Mining, School of Computer Science, The University of
  Manchester
  (gkontonatsios@cs.man.ac.uk,
  {fioannis.korkontzelos, sophia.ananiadou}@manchester.ac.uk)

- Erik M. van Mulligen, Quoc-Chinh Bui, Jan A. Kors
  Dept of Medical Informatics, Erasmus University Medical Center Rotterdam, The
  Netherlands
  {j.kors,e.vanmulligen}@erasusmc.nl, bqchinh@gmail.com
Overview on the used solutions

For all annotation solutions the challenge participants gave an overview on the components of their approach (see table 4.1). This information can be exploited to determine, how the components of the annotation solution influenced the challenge outcome. Apart from the components that are part of the annotation solutions, the table also indicates to which languages the annotation solutions have been applied.

56 submissions have been received from 7 participating teams, where 3 teams are members of the project consortium: three project partner teams with 39 submissions and four participating teams with 27 submissions. Overall, the total submission numbers show a reasonable size and also the number of participants is significant, since contributions came from seven teams, from ten different sites involving 23 team members.

Other challenges that are in a similar state of maturity, i.e. they have been newly established could motivate about the same number of participating teams or significantly less (like BioASQ, CLEF QA4MRE). Furthermore, although only four external teams participated into the challenge, they did still contribute large number of contributions, i.e. 27 submissions, which is crucial for the successful execution of the Mantra project.

Finally, the following considerations have to be kept in mind to explain that the challenge addressed mainly special groups that work in European languages, are processing large volumes of documents and terminologies, and find interest in biomedical domain knowledge:

1) The CLEF-ER challenge has been proposed for the first time and therefore the project partners had to raise enough awareness throughout Europe and the world to reach the research community.

2) It exposes a high complexity, i.e. the challenge participants had to do their annotations in one of the selected non-English languages.

3) The high complexity of the approach is also reflected in the extensive use of resources and the high integration effort (see table 3).

4) The participants had to also use the provided semantic biomedical terminological resources for their annotations, which is a burden to teams who are not involved in biomedical semantics solutions and requires special background knowledge.

5) The use of a silver standard corpus in contrast to the gold standard is yet less common and less well known and therefore the CLEF-ER challenge had to overcome a discomfort in the research community to participate.
<table>
<thead>
<tr>
<th>Use of Mantra TR</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of Mantra SSC (in English)</td>
<td>yes</td>
<td>yes</td>
<td>yes(?)</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>Yes</td>
</tr>
<tr>
<td>Statistical Machine Translation</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>No</td>
</tr>
<tr>
<td>Own Dictionary from SSC</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
</tr>
<tr>
<td>Phrasal Alignment / SMT</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>No</td>
</tr>
<tr>
<td>Word Alignment / SMT</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>No</td>
</tr>
<tr>
<td>Indexing (corpora), lexical lookup</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>Yes</td>
</tr>
<tr>
<td>NP identification / Chunking</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
</tr>
<tr>
<td>Multiple assignment of CUIs</td>
<td>Yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>Yes(?)</td>
</tr>
<tr>
<td>Use of Entity disambiguation</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>No</td>
</tr>
<tr>
<td>Evaluation</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>No</td>
</tr>
<tr>
<td>Languages</td>
<td>en, es</td>
<td>en, es</td>
<td>fr, es</td>
<td>en, de, nl, fr, es</td>
<td>de, fr</td>
<td>en, de, es, fr, nl</td>
<td>en, de, es, fr</td>
</tr>
<tr>
<td>New resources</td>
<td>Translated corpus</td>
<td>-</td>
<td>NP taggers in 3 languages</td>
<td>Translated terminological resource</td>
<td>Enriched terminological resource</td>
<td>Enriched terminological resource</td>
<td>-</td>
</tr>
<tr>
<td>Other resources</td>
<td>-</td>
<td>UMLS</td>
<td>UMLS, Wikipedia</td>
<td>Stanford parser, Malt parser, MetaMap, Giza++</td>
<td>MeSH, MedDRA, Snomed-CT</td>
<td>BabelNet (WordNet, Wikipedia)</td>
<td>Lingpipe gazetteer, JCoRe NER engine</td>
</tr>
<tr>
<td>Other tools</td>
<td>Tanl Tagger for ER (MEMM based)</td>
<td>UMLS</td>
<td>Stanford parser, Malt parser, MetaMap, Giza++</td>
<td>Google Translate</td>
<td>GERTWOL, OntoGene term matcher</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Synopsis</td>
<td>ER in a translated corpus</td>
<td>Indexing of the terminology, documents as queries</td>
<td>Synopsis-MC co-training approach on pairs of languages</td>
<td>Translation of the terms via Google, indexing of corpora</td>
<td>Translation of terms via BabelNet, lexical lookup in corpora</td>
<td>Phrase-based SMT &amp; NER</td>
<td>ML approach to identify pairs of terms in 2 languages</td>
</tr>
</tbody>
</table>

Table 4.1 (Annotation solutions): All challenge participants have been asked to describe their annotation solutions. The table gives an overview on the result. It becomes clear that the different teams used different combinations of solutions including exploiting the lexical resources and the integration of statistical machine translation.

Note: Participant G contributed to the challenge, but refrained to be named officially on the challenge due to technical difficulties.
5. Evaluation

The challenge participants have received the non-English corpora and an English SSC, which was based on the annotations from the project partners. The challenge participants had to annotate the non-English corpora and send these corpora back to the submission site. The participants have been asked to annotate the mention of a biomedical entity with a CUI.

For the evaluation the following approaches have been identified:

- **Evaluation B**: The CUI annotation from the non-English annotated corpus can be evaluated against the CUI annotations in the provided English SSC.

- **Evaluation A**: The so-called “mention” annotation in the non-English annotated corpus can be compared against a non-English SSC that will be generated from the challenge contributions, i.e. all submitted annotated contributions are used for the generation of an SSC and this SSC is then used to measure the performance of the challenge contributions against the SSC (called the “Silver Standard Approach”).

- **Evaluation A/-t**: The Mantra challenge aims at determining novel terms from the parallel corpora. Therefore, the project partners have introduced this additional variant of evaluation A, where in the evaluation against the non-English SSC (= evaluation A) the known non-English terms have been considered. As a result, this evaluation gives a measure to which degree the challenge contribution predicted a non-English term that has been introduced to the harmonized set in the generation of the SSC. This approach measures to which degree a solution produces novel non-English terms similar to the contribution of the other challenge solutions.

**Preparation of the non-English SSCs**

The SSC is generated from the annotated corpora by automatic means. The methods for the generation of the SSC are described in [Lewin et al., 2012; Lewin et al., 2013].

After the challenge has been closed, all submissions have been processed. The contributions have been tested on their compliance with the required formats. Where possible any inconsistencies have been resolved. In the communication with the challenge participants, causes of data inconsistencies have been identified and removed. The challenge participants could suggest those submissions that will be included into the SSC generation. All participants received feedback on their contribution prior to the CLEF-ER workshop.

The table 5.1 lists all submissions to the challenge, and table 5.2 all the submissions that have been considered for the SSC generation.
Spanish was the most popular language, i.e. the Spanish corpora have been annotated by the largest number of participants and the largest number of submissions was linked to Spanish. French was more popular than German and the least contributions – as expected – were delivered for Dutch. These figures are relevant for the evaluation of the challenge, since a larger number of contributions leads to a larger set of annotated corpora that can be considered for the generation of a SSC in a given language.

Not all submissions have been included into the SSC. More in detail, only one submission has been considered from a team that provided two submissions. This is required to avoid any biases from individual solutions. Furthermore, not enough annotated Dutch corpora have been provided from the challenge participants to produce a reliable SSC for the Dutch evaluation. Although the Dutch corpora have been processed and evaluated, the numbers for this evaluation won’t be reported in this deliverable since the numbers are deemed less reliable than for the other non-English languages.

In comparison to the CALBC challenge, the SSC generation based on centroids has been modified to identify the maximum boundaries about a centroid that fulfill the following conditions:

Table 5.1 (Submissions to the CLEF-ER challenge): The table gives an overview on the submissions to the CLEF-ER challenge. For all corpora and for all languages at least one annotated corpus has been contributed.
(1) the centroid is still fully part of the overall boundaries after harmonisation,
(2) the left and the right boundary of the mention annotation have minimum support (= minimum threshold) by the contributing systems (e.g. at least two-vote agreement), and
(3) all contributing annotations support either the same semantic type or even the same concept unique identifier.

It is important to distinguish the mono-lingual non-English SSCs for the evaluation A („mention annotation”) from the English SSCs for the evaluation B („entity normalisation”). The former uses the submissions from the challenge participants as input, and the latter is based on annotated corpora from the project partners prior to the challenge (PPSSC, for SSC from project partners). Both SSCs serve as a consensus annotation standard between the different annotation solutions.

Performing the evaluation
All files, i.e. all submission, have been analysed against the silver standard corpus (SSC, in English). The first evaluation is concerned with the correct identification of CUI’s in the non-English corpus, where the reference annotations stem from the English corpus. In addition, the mention annotations in the non-English corpus will be assessed against an SSC, which will be produced from all available annotated non-English corpora. Note that – in this case – the annotations from the challenge participants will be included into the SSC.

The evaluation of the contributing solutions is performed against the corpora that have been prepared prior to the challenge. All evaluations have been performed automatically and a number of parameters have been determined that help to judge the challenge and the challenge outcome.

Since the evaluation is performed against a SSC, the performance evaluation can always be interpreted into two opposite directions: first, we can judge the performance of the contributing systems in the experimental setup, and second, we can judge the consistency of the experimental setup by comparing the performances of the contributing solutions. For example, we can expect that in a consistent experimental setting, the same annotation solution would always perform in the same way, i.e. the comparison of the different solution leads to the result that they are ranked the same or in a similar way in the same experiments. If we subdivided the whole annotations into different disjoint groups and rank the performances of the solutions, then we would expect that the profiles of the contributing solutions composed from the ranks of the solutions against the different semantic groups are correlated across the different systems.

For the evaluation of the solutions two schemes have been used, which are called Evaluation A for the mention evaluation and Evaluation B for the normalization task.

Mono-lingual mention evaluation (Evaluation A)
In order to assess the quality of the annotations in all non-English corpora, a mention agreement evaluation against a harmonized silver corpus built from the monolingual contributions of the challenge participants and from annotations from project partners was
performed. Table 5.1 shows the number of contributions from the challenge participants that have been included into the generation of the SSC.

Additional annotated corpora have been made available from the MANTRA project partners prior to the challenge. A subset from these annotated corpora has been used for the development of the centroid-based SSCs (see below).

**Cross-lingual concept evaluation (Evaluation B)**

Given the fact that the English terminology covers a lot more concepts and provides more synonyms for them compared to the non-English terminologies, a second evaluation of concept coverage against a harmonized English silver standard corpus built from the Mantra project partners was performed. For each corpus there are 6 different annotations that are harmonized into a centroid-based silver Standard using a voting threshold of 3 [Lewin et al., 2013].

**Harmonised means of evaluation A+B (Evaluation AB)**

Evaluation A tests the mention annotation and evaluation B tests the normalization of the entities (or concepts) against the reference corpus. We also measured the performance in using the harmonized means from the evaluation A and B to determine those solutions that performed well in both approaches.

**Mono-lingual de-annotated mention evaluation (evaluation A/-t)**

Since the Mantra project aims at the detection of novel terms from the annotated corpora, we also evaluated an approach where the annotations have been reduced by those terms that are already contained in the provided terminology, i.e. the non-English SSC contains novel terms with regards to the reference terminology. This approach measures whether an annotation solution, i.e. a contribution to the CLEF-ER challenge, identifies terms in the corpus that are not yet available from the reference terminology.

It becomes clear that the performances in evaluation A and evaluation A/-t differ, i.e. a solution can be a good „reproducer“ of annotations against the harmonised corpus (evaluation A for one corpus, one language, one semantic type), but would not be necessarily a good „innovator“ (evaluation A/-t for the same set).
6. Results

The CLEF-ER challenge offers a number of opportunities to assess the performance of the contributing solutions and all of them have been used to assess the quality of the annotation solutions, but also to determine the consistency of the CLEF-ER approach and to judge the relevance of the text corpora and the Mantra seed terminology for the challenge setup.

For the different measurements we can distinguish as relevant experimental parameter the following components to the CLEF-ER challenge:

- the corpora (´EMEA', `Medline', `Patent');
- the different languages (´de', `es', `fr', `nl');
- the semantic groups according to the UMLS:
  anat, chem, devi, diso, geog, livb, objc, phen, phys;
- the evaluation against mention boundaries (evaluation A) and against the CUIs in the English corpus (evaluation B);
- the identification of terms that are already known in the provided terminology (evaluation A) and the identification of novel term candidates in the non-English language for the Mantra terminology („evaluation A/-t‟);
- it is possible to calculate the harmonized means between the F-measure from evaluation A and evaluation B, giving a combined measure for both performances („evaluation AB‟);
- the correlation between the different annotation solutions, where the correlation can be based on different features from the assessment outcome.

**Best overall systems according to the F-measure**

The most straightforward approach is the evaluation according to the F-measure across all annotations for a single CLP, i.e. all TP, FP and FN results across all semantic groups in the terminological resource have been gathered for the calculation of the precision, recall and the F-measure.

In table 6.1 the three evaluations are shown: evaluation A (left), evaluation A/-t (middle) and evaluation B (right). For all three corpora, the best performing solution in a given language is named and the F-measure determined on the overall figures of the TP results and the error figures is shown there as well.
Table 6.1: The best performing solutions for overall F-measure (“F-m”) for the different CLPs are shown in the table. The color-encoding of the different systems is maintained throughout this deliverable.
(A1/A2: light-/dark-green; B1: light-violet; C1/C2: light-/dark-red; D1: dark-violet; F1/F2: light-/dark-blue; G1/G2: light-/dark-orange)

This leads to the result that
- for example, E1 is the best performing solution for the EMEA/de CPL (see top left corner of table 6.1);
- for evaluation B:
  o D1 is the leading solutions, since it has the best performance on the CPLs for 6 out of 10 possible contests (all EMEA and Patents CPLs, and EMEA/es)
  o G2 and G1 show good performances on the Medline CPLs
  o B1 shows the best performance for the Medline/es CPL
- for evaluation A:
  o E1 is overall the best performing solution on evaluation A with 5 out of 8 possible CPL experiments (all German CLPs, French CLPs except for Medine)
  o D1 also performs well on the Medline/fr and the Medline/es CPL, and finally A2.
  o A2 is the best performing on the EMEA/es CPL (third column under evaluation A)
- for evaluation A/-t:
  o D1 is the dominating solution with six out of eight best performances
  o E1 is the best performing solution on the EMEA/de and the EMEA/fr CPLs

A more detailed analysis of the performances of the challenge contributions will be given in the following chapters. Furthermore, table 4.1 gives an overview on the components of the annotation solutions. The two solutions D1 and E1 made use of term translation techniques to produce their results, whereas A2, G1 and G2 are mainly machine learning techniques that have been trained on different corpora.

Assessment of the FP, FN, TP result across all solutions
The evaluation of the challenge contributions requires calculating the true positive results as well as the false positive errors and the false negative errors leading to the precision and recall
figures, and finally into the calculation of the F-measure. Since the latter does not give any clues anymore on the underlying figures, we have generated diagrams that show the performances of the annotation solutions according to their error profile (see table 6.2, evaluation A; table 6.3, evaluation B).

Table 6.2: The diagrams show the FN (light blue, bottom part), TP (dark blue, middle part) and the FP (green, top section) results of the challenge contributions against the SSC in evaluation A; for EMEA, Medline titles and patent claims (from top to bottom). The FN counts and the TP counts result to the positive entries in the annotated corpus and is recognizable as a straight line going across the upper limit of the TP counts in all solutions.

Taking the first diagram in the top-left corner of the table 6.2, we see the results from the systems D1, E, F1/2 and G1/2 against the EMEA/de CPL. F1 and F2 has the smallest portion of FN errors (bottom section, light blue) and the biggest portion of TP results. F2 and G2 have the largest number of FP results (green), which is displayed to the top of the diagram.
A few findings can be drawn from table 6.2:

- The annotation solutions produced on the patent corpora a larger number of FP results than against all the other corpora which is probably due to the complex nature of the patent claims, in particular the language use in legal text. It is also the case that the selection of patent documents which were used in the challenge did not cover well the biomedical domain knowledge. This was noticed after the challenge and a more appropriate selection of documents was produced. We conclude that the performance drop is partially due to the fact that we have a mismatch between the content of the patents and the domain terminology.

- Annotation solutions with a lower FN rate tend to generate a higher FP rate, since they overall annotate a larger number of candidate terms. This phenomenon is known and leads to the result that the F-measure between the solutions may not differ significantly.

- The annotation solutions show on Medl/de CPL very similar error type profiles, whereas the same profiles vary to a large extend on the EMEA/es CPL and on the EMEA CPLs overall.

- It can be expected that the language diversity is higher on the EMEA CPLs than on the Medline CPLs, but the contrary appears to be the case, i.e. the error profiles of the annotation solutions on Medline seem to be more homogeneous than on the EMEA CPLs.

- Only the solutions C1 and C2 seem to fail on identifying a significant number of the annotations from the SSC.

Altogether, all solutions (possibly apart from C1 and C2) produced results that indicate that the annotation solutions can cope with the challenge task, i.e. the tagging solutions produced the annotations that have been expected for the task.

Again, it has to be kept in mind that all annotation solutions from the challenge make their contribution into the SSC for the different non-English languages. Therefore, the best performing solution in evaluation A shows the highest compliance with the consensus amongst all tagging solutions represented by the SSC.
Table 6.3: The 8 diagrams show the same results as covered in table 6.2, i.e. the performances from evaluation A, but in these calculations the known non-English terms from the Mantra terminology have been removed to only judge the performances of the NER taggers against the term candidates that stem from the SSC.

In Table 6.3, the same data is shown as in table 6.2, but now the counted annotations have been reduced to only cover those terms that are not already provided by the Mantra terminology. This measurement gives an assessment of annotations that are deemed to be novel in the SSC over the existing terminological resource. The performance of the different annotation solutions in this assessment is different from the previous one in the sense that the number of annotations (TP + FP) is lower overall, and that the FP rate of the systems is higher than in the previous evaluation leading to lower precision and lower F-measures. Apart from this finding, the heterogeneity of the results is similar to the regular evaluation A.

Again, it is important to keep in mind that the evaluation A/t judges the annotations against the consensus annotations from the different challenge participants (Silver Standard approach)
and that therefore novel terms are defined by the annotation contributions from the challenge participants. A FP result could still be a valid term – with a lower likelihood though – which may only be suggested by the current tagger. On the other side, all those terms that form true positives are still terms that have been identified by entity recognition without using the Mantra terminology.

The ground truth for the novel term candidates can only be identified either (1) by the development of a gold standard approach suitable to judge the novel term candidates (which is under way), or (2) by evaluating the term candidates through external judges, which is an option for the future.

The performances of the tagging solutions in the evaluation B show very similar results as given in table 6.2 (data not shown). Overall the error rates are higher and thus the precision, recall and F-measure performances are recorded at lower levels, but the distribution of results is very similar to the one displayed for evaluation A (see following results).

**F-measure across the different semantic groups**

To give a better overview on the distribution of the results, we have generated boxplots that list the performances of the annotation solutions in the evaluation A and B.

In the first diagram the F-measures as determined in the evaluation A across the different semantic groups are shown. Each diagram represents one CPL, and each boxplot in one diagram comprises the performances of one tagging solution against the different semantic groups covered in the CPL. The color encoding is kept for each system across the different challenges and even in the other diagrams that show the same comparisons. This analysis shows – more than the error type representation – the coherence of the annotation solutions for the different semantic groups.

The top-left diagram in table 6.4 displays the results against the EMEA/de CPL, i.e. this diagram has the same structure as the tables 6.2 and 6.3 showing from the top to the bottom the EMEA, Medline and patent CPLs and from the left to the right the German, French and Spanish results. The top-left diagram shows the distribution of F-measures for four annotation solutions (D1, E1, F1 and F2) against the semantic groups, i.e. E1 shows a higher diversity of F-measures across the semantic groups than D1 and F1 does.
Table 6.4: The boxplots all display the F-measure performance of the different annotation solutions against the CPLs in the evaluation A. (A1/A2: light-/dark-green; B1: light-violet; C1/C2: light-/dark-red; D1: dark-violet; F1/F2: light-/dark-blue; G1/G2: light-/dark-orange)

The following results can be extracted from the diagrams:
- The performance against the German corpora is more coherent, i.e. there is less variability in the F-measure distribution than can be seen from the performances against Medl/fr and the Spanish corpora.
- The F-measure variability for C1 and C2 is highest amongst all solutions.
- The results on the Patent CPL show very similar distributions, although the overall performance is lower.
- The solutions D1 and to a lesser degree E1 show a narrow distribution of F-measures across the corpora.
Even related solutions (e.g., F1 and F2) do show quite different spreads, but overall about half of the solutions have a limited spread with only a few outliers. The same analyses have been performed for the evaluation A/-t as shown in table 6.5.

Table 6.5: The boxplots in this table – in contrast to the previous table 5.4 – have been determined against the de-annotated SSC of the evaluation A, i.e., in the evaluation A/-t. (A1/A2: light-/dark-green; B1: light-violet; C1/C2: light-/dark-red; D1: dark-violet; F1/F2: light-/dark-blue; G1/G2: light-/dark-orange)

As seen from the previous analyses, the overall performance is reduced and the spread in the F-measure across the different semantic groups is increased. For the Medline CPLs, the G1 and G2 solution show a small spread, but other solutions too on selected CPLs (e.g., F1/F2).
7. Evaluation A

The evaluation of the mention annotations has been performed against monolingual SSCs that have been generated from the annotated corpora contributed through the CLEF-ER challenge by the challenge participants. From the previous analyses based on descriptional statistics it became already clear that the annotation solutions differed in their error profiles and the resulting performances. Now we will compare the different solutions according to statistical parameters that help to rank the performances of the tagging solutions.

The performance evaluation indicates that – with a few exceptions – the annotation of the EMEA documents can be achieved with better results than the annotation of the Medline, or the patent documents. This results is true for all languages except for Dutch. The mention annotation of the patent documents shows a mixed picture, since in general the performance for the annotation in German and French resembles the performance produced on the other two corpora, but when comparing the different semantic groups it becomes clear that for selected groups the performance is deteriorated (e.g., phenotype – `phen', `anat', `livb' and `chem')

Different measures

The following table gives an overview on the best ranked solutions in the evaluation A according to different measures which comprise (1) the standard F-measure calculated on all errors without considering the semantic group annotation, and two measures based on the F-measures determined per semantic group, i.e. (2) average F-measure, and (3) the median of the F-measure across the semantic groups.

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1 avg</td>
<td>F1 avg</td>
<td>F1 avg</td>
</tr>
<tr>
<td>EMEA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>74.1</td>
<td>67.0</td>
<td>66.9</td>
</tr>
<tr>
<td>B1</td>
<td>57.5</td>
<td>61.7</td>
<td>66.0</td>
</tr>
<tr>
<td>C1</td>
<td>74.2</td>
<td>55.5</td>
<td>68.1</td>
</tr>
<tr>
<td>C2</td>
<td>72.4</td>
<td>54.9</td>
<td>55.6</td>
</tr>
<tr>
<td>D1</td>
<td>68.2</td>
<td>78.6</td>
<td>68.9</td>
</tr>
<tr>
<td>E1</td>
<td>86.1</td>
<td>79.4</td>
<td>87.4</td>
</tr>
<tr>
<td>F1</td>
<td>75.6</td>
<td>75.1</td>
<td>77.0</td>
</tr>
<tr>
<td>E1</td>
<td>71.2</td>
<td>68.8</td>
<td>78.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.1:** The table gives three different measures to judge the performance of the tagging solution against a CPL: (1) F-measure across all annotations (“F1”), (2) average F-measure across all semantic groups (“avg.”), and (3) the median of the F-measures across all semantic groups.

Version 1.0 (28.11.2013)
In table 7.1 only the best ranked solution according to the F-measure has been marked up. Only in a few instances this measure deviates from the other two measures, e.g. EMEA/es (A2 vs. D1) and Patent/fr (E1 vs. D1), leading to the conclusion that the overall F-measure but also the average F-measure across all semantic groups should serve as a reliable means to judge the different challenge contributions.

Overall E1 is the most successful solutions on the evaluation A task, since this solutions shows the best F-measure in the majority of the CPLs.

<table>
<thead>
<tr>
<th>Test-A</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>C1</th>
<th>C2</th>
<th>E1</th>
<th>E2</th>
<th>G1</th>
<th>G2</th>
<th>Tot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-1st</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>All-2nd</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>19</td>
<td>4</td>
<td>68</td>
</tr>
<tr>
<td>EMEA</td>
<td>14</td>
<td>11</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>Medline</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Patent</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>anat</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>chem</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>diso</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>livb</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>dmnt</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>geog</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>objc</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>phen</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>phys</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>All-1+2</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>EMEAMedl</td>
<td>4</td>
<td>22</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>144</td>
</tr>
<tr>
<td>Big-4-Grp</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Oth-Grp</td>
<td>15</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 7.2: Each count in the table is the scoring of a tagging system in a CPL for a single semantic group except for “All-2nd”, which considers the second rank for a solution.

Table 7.2 gives several clues on the judgement of the different annotation solutions. In general it shows the frequencies for any annotation solution to score as the first ranked solution in one CPL. The y-axis lists the groups that have been formed to judge these performances. For example, the challenge contribution D1 has scored 25 times as the best ranked solution in all CPLs (top row), and has scored 11 times as second best solution over all CPLs (2nd row). Both figures together (“All 1+2”, bottom part) leads to the result that D1 scored 36 times as best solution on rank 1 or 2 for all CPLs. Considering only the EMEA and Medline CPLs, it scores 8 times best on EMEA, 14 times best on Medline and 22 times best on both solutions.

It was noticed during the analysis of the challenge, that four semantic groups cover most of the annotations: “anat”, “chem”, “livb”, and “diso”. This group is called “Big-4-Grp” in the diagram and the remaining semantic groups are summarized as the “Oth-Grp”. Please notice that the solution E1 scores well on the Big-4-Grp whereas D1 performs better on the other semantic groups, which could indicate that Babelnet contributes useful terms for the four “bigger” semantic groups, whereas the terms from the other semantic groups are better covered by Google translate than by Babelnet.
In the next analysis, we have determined the frequency of a tagging system to be ranked first according to its F-measure on a single semantic group for a given CPL. It can be expected that the NER solutions have different annotation profiles concerning the semantic groups on a given CPL, and that even the best tagging solution does not perform evenly well on all semantic groups in a CPL. The 2nd position was also considered across all CPLs and semantic groups to deliver additional input on judging the outcome of this analysis.

Again, E1 and D1 show the best performances, and the tagging solution F1 also ranks first in a number of experiments. D1 is the best tagging solution in Spanish and E1 for French and German, which may also be due to the fact that E1 did not participate in the Spanish CPL experiments.

**F-measures across semantic groups**

The following three diagrams display the F-measure performances of the challenge contributions against the different corpora and the semantic groups. The systems with the best average of the F-measures for all semantic groups have been listed on the bottom, and the ones with the worst average on the top.

![Diagram 1](image1.png)

**Table 7.3:** F-measure performances against the non-English EMEA corpora. The systems have been stacked according to the average F-measure performance, with the best performance at the bottom.

(A1/A2: light-/dark-green; B1: light-violet; C1/C2: light-/dark-red; D1: dark-violet; F1/F2: light-/dark-blue; G1/G2: light-/dark-orange)

Table 7.3 covers the results for the EMEA corpus. The left diagram shows the results from E1 (in yellow, bottom), F1 (light-blue, second), F2 (dark-blue, below the top) and D1 (purple, top). E1 is in the bottom section, since it has the best average F-measure performance across the different semantic groups followed by F1, F2 and finally D1.

The y-axis has been scaled with the number of contributing solutions and in the ideal case the F-measure of the different solutions would each be 1 adding up to the total number of contributing solutions. This way the maximum bar length is an indicator for the overall performances of all solutions against the same semantic group for a given CPL.

For example, all solutions in the EMEA/de CPL have a good performance in the “livb” semantic group and the performance in the “objc” semantic group is lower than in the other semantic groups, where F2 in particular has a poor performance in this contribution. In the
semantic groups “phen” and “devi” we find also rather low performances, whereas F1 shows a high F-measure in the “anat” semantic group.

Table 7.4: The same data representation but now for the Medline CPLs as provided in the previous table 7.3.

Table 7.4 shows the results of the different annotation solutions against the Medline CPLs. The most contributions have been received for the Medl/fr and the Medl/es CPL. D1 shows strong F-measure performances, but also G1 and G2 on Medl/fr and Medl/es. Again, performances drop for the “objc” semantic group. Furthermore, the best annotation results have been achieved in the semantic groups ‘livb’, ‘geog’ and ‘chem’. This is not surprising, since these groups contain terms which are well standardized in the biomedical domain. The lower performances for e.g. in ‘phys’ or ‘objc’ can be explained by the fact that these types of terms are not as well standardized as the previous ones and the ones from other categories in general.

Table 7.5: The same data representation but now for the Medline CPLs as provided in the previous tables 7.4 and 7.5.

The performances on the patent CPLs were lower than on the other CPLs. This is again due to the fact that the patents deviated from the domain knowledge as represented in Medline and EMEA. Surprisingly, the identification of “geog” in patents lead to poor performances.
Table 3.6 (Evaluation A) – The tables denote the F-measure results of the annotation solutions against the EMEA and the Medline CPL sorted by semantic groups.

A number of solutions demonstrate very similar performance in the different corpora regarding the same semantic groups. For example, the system C1/C2 (from team “C”), has only been applied to the Spanish and French corpora and shows a very similar profile in EMEA and in Medline. A similar result can be derived for the system F1/F2 although the performance varies between different languages.
8. Evaluation B

The challenge participants had to produce annotations for their preferred corpus in their preferred languages, which should cover at least one non-English language. The annotations had to comprise the assignment of a CUI to the entity mention. Please refer to [Rebholz-Schuhmann et al., 2013c] for an overview on the combined results from all challenge contributions.

**Different measures**

<table>
<thead>
<tr>
<th></th>
<th>German F1</th>
<th>French F1</th>
<th>Spanish F1</th>
<th>Dutch F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg.</td>
<td>med.</td>
<td>avg.</td>
<td>med.</td>
</tr>
<tr>
<td>A1</td>
<td>41.3</td>
<td>54.7</td>
<td>50.5</td>
<td>50.5</td>
</tr>
<tr>
<td>A2</td>
<td>45.7</td>
<td>50.0</td>
<td>50.6</td>
<td>50.4</td>
</tr>
<tr>
<td>B1</td>
<td>54.9</td>
<td>63.0</td>
<td>61.6</td>
<td>59.9</td>
</tr>
<tr>
<td>C1</td>
<td>54.8</td>
<td>62.6</td>
<td>61.0</td>
<td>57.3</td>
</tr>
<tr>
<td>C2</td>
<td>54.1</td>
<td>62.1</td>
<td>61.5</td>
<td>57.7</td>
</tr>
<tr>
<td>D1</td>
<td>60.9</td>
<td>53.0</td>
<td>50.1</td>
<td>50.2</td>
</tr>
<tr>
<td>E1</td>
<td>50.7</td>
<td>45.2</td>
<td>50.9</td>
<td>50.9</td>
</tr>
<tr>
<td>D2</td>
<td>54.2</td>
<td>48.0</td>
<td>50.6</td>
<td>50.4</td>
</tr>
<tr>
<td>E2</td>
<td>56.1</td>
<td>54.1</td>
<td>50.6</td>
<td>50.4</td>
</tr>
</tbody>
</table>

**Table 7.6**: The table gives an overview on the different F-measures for the evaluation B in the CLEF-ER challenge.

The table 7.6 shows the performances of the different solutions in the evaluation B. All measures lead to the same best ranked solutions except in the case of the CPL EMEA/es, where “F1” has the best performance for the measures based on the semantic groups (“avg.”, “med.”) in contrast to “B1” which has the best performance according to the F-measure.

On EMEA, the system D1 has the best performance except for EMEA/ex, where B1 performs best. For patents again D1 shows the best performance, and for Medline D1 scores best for Medl/nl, whereas G1 and G2 perform best for Medline in the languages de, fr and es. The performances given as F-measures are lower than for the evaluation A, which is partially due to the fact that the contributions from the challenge participants are not used for the SSC, where they have been judged against, i.e. the contributing solutions have to match the consensus annotations from the systems provided by the Mantra project partners.

The next table for the evaluation B is based on the distribution of the best ranked solutions across the different sets of CPLs, i.e. it counts how often a solution scores best across the different corpora, languages and semantic groups (see also table 7.2 from evaluation A).
Table 7.7: The overview shows the best performing solutions counted for the different corpora, languages and semantic groups.

In this analysis, the solutions F1 and F2 are frequently judged to be the best ranked solutions for a given semantic group and a given CPL, followed again by D1 and E1 (see “All 1+2” in table 7.7).

Different from previous results, the table 7.7 indicates that the solution F2 and also F1 frequently gets the best rank on an evaluation in a semantic group and in a CPL. This result seems to be in contradiction to the result from table 7.6, where G1, G2 and D1 are most prominent, but can be explained by the way how both evaluations have been performed: G1/G2, i.e. the average performance based on the semantic groups whereas this approach counts all CPLs where the solutions has been judge to be the best performing solution. Form the Boxplots.

F2 outperforms the other solutions on the patent CPLs

F-measures across semantic groups

Table 7.8: F-measure performances in the evaluation B. The systems have been stacked according to the average F-measure performance, with the best performance at the bottom. (A1/A2: light-/dark-green; B1: light-violet; C1/C2: light-/dark-red; D1: dark-violet; F1/F2: light-/dark-blue; G1/G2: light-/dark-orange)
The three diagrams (table 7.8, 7.9, and 7.10) again show the stacked F-measure performance of the different tagging solutions against the semantic groups, the best performing solution to the lower parts (according to the average F-measure across the semantic groups) and the systems with lower performances to the top. Overall the performances are lower than in the evaluation A. The solutions D1 and E1 show the best performance in the EMEA/de and EMEA/fr CPLs whereas F1 has the best average performance on EMEA/es. The performance on the semantic group “anat” tends to be lower than for the other three groups with larger number of annotations (“diso”, “chem”, “livb”). Performances on “objc” are in particularly low over all annotation solutions.

Table 7.9: The diagrams show the same systems and performances now measured on the Medline CPLs instead of the EMEA CPLs as in table 5.12.

Against the Medline title CPLs, the solutions G2, D1, and G1 show the best average performances. It is remarkable that the identification of “livb” produces lower results than the identification of “diso” and “chem”, although the species mentions are quite universal in the biomedical domain.

Table 7.10: The diagrams show the same systems and performances now measured on the Patent CPLs instead of the EMEA CPLs as in table 5.12.
The performance against the German Patent corpus is the worst in comparison to the other CPLs and the other challenges again showing the high heterogeneity of the language used in the patent texts (see table 7.10). In particular, the annotation of the CPL Patent/de appeared to be more complex than all the other CPLs. This is again partially due to the fact that the patent documents have been less well selected than the other CPLs.

Overall, we conclude that – apart from selected semantic groups – the CUI annotation of the corpus could be performed in similar ranges across the different CPLs. A few solutions, such as C1 and C2 did not comply similarly well with the harmonized SSC as other solutions such as G1 and G2 that contributed to the annotation of the entities to a larger degree than the other annotation solutions.

The last table 7.11 lists all the measured performances of the tagging solutions against the CPLs.

Table 7.11 (Evaluation B) – The table lists all performances of all systems sorted by the semantic groups.

The two tables again show again that C1/C2 produce low performances in comparison to the other solutions and that “chem”, “diso”, and “livb” are the semantic groups that produce the best performance results across the different tagging solutions.
9. Other evaluations

Further evaluations have been performed to better describe the outcomes of the CLEF-ER challenge. The following questions have been addressed:

- Can we produce a measure that helps to judge the annotation solutions for both evaluations together to suggest solutions that have produced good results in both approaches?
- Can we compare the results between the tagging solutions directly?
- What performances do we see, if we want to judge the novelty of the identified terms?

Combining evaluation A and evaluation B

(The same text as in the annual report D1.1.2, here in a shortened and slightly modified form)

The figure 9.1 compares the results from both evaluations in a single diagram. The performances from evaluation A have been averaged across all challenge contributions for a single corpus in a single language (corpus-language pair, CLP). Only one contribution from each challenge participant has been accepted to avoid any biases. This also leads to the result that the final figure for Spanish (for any of the corpora) makes use of more challenge contributions than the resulting average for any of the other languages and any of the other CLPs.

The resulting average figures enable the comparison of the results from evaluation A (dashed lines) and B (solid lines) in the same CLP, but also the comparison of the performances in one CLP against another CLP, where either the corpus can be different and the language the same, or the language may change and the corpus stays the same. The averages across the recall and precision performances are given in the diagram in addition to the average of the F-measure performances.

![Diagram showing performances of different contributions to CLEF-ER challenge](image_url)

**Fig. 9.1 (Evaluation A and B)** – The diagram shows the performances of the different contributions to the CLEF-ER challenge averaged for precision, recall and F-measure in evaluation A (dashed line) and in evaluation B (solid line). The average involves all contributions for one corpus-language pair (CLP).
The results from figure 9.1 demonstrate that the average F-measure in the evaluation B is lower than in the evaluation A. The F-measure performance is lowest on the Patent/de CLP, and even when comparing the Patent/fr CLP against the other two French CLPs, the Patent/fr CLP gives lower performance than the other two.

For evaluation B the recall performances are lower than the precision performances, except for the EMEA/es CLP which shows almost even recall and precision performances. For evaluation A we see a different picture. Here the average recall performances are high and the average precision performances limit the average F-measure performances. Again we find cases, where the average recall and the average precision performance are better than the average F-measure performances (Mdl/fr, EMEA/es, Mdl/es, Mdl/nl, see above).

For evaluation A, the average F-measure performances on the Medline CLPs are lower than the ones on the EMEA CLPs. We conclude that EMEA uses terminology that is better standardized than the high diversity of terminology in the Medline CLPs. This is supported by the fact that a larger number of annotations have been achieved with the Mantra terminology on the Medline titles in contrast to the EMEA corpora, although the Medline corpus and EMEA corpus have similar sizes according to their word counts.

These results are not at all surprising, if we consider the origin of the different documents and resources. EMEA documents provide the same content in the different languages and this is the reason, why we would expect the same annotations in the different CLPs and not many deviating annotations between the English and the non-English documents. For the Medline titles we have bilingual parallel corpora, i.e. the units in the German-English document set do not have any similar documents in the Spanish or French document set. In other words, the French-English parallel Medline corpora are completely separate from the other parallel Medline corpora, leading to a different selection of terms and CUIs.

Considering these arguments we can also conclude that – although EMEA and Medline have different structural characteristics – the average F-measure performances are still in the same range.

For evaluation B we find that the average F-measure performances on the Medline CLPs in comparison to the EMEA CLPs are better for French and Spanish, and worse for German. This could be explained by the following reason: different synonyms in the non-English language are translated into the same CUI in the evaluation against the English SSC increasing the likelihood to identify the correct CUI.

**Harmonised means of evaluation A+B (Evaluation AB)**

In the following we have calculated the harmonized means of the F-measure performances according to evaluation A and B to be able to rate the annotation solutions for both approaches in a single measurement. More in detail, the statistical parameters from the evaluation A and B have been harmonized and are given in the table 9.1 below.
**Table 9.1:** The table provides the harmonized means of the F-measures from evaluation A and B leading to a judgment, which solution performed best “on average” in both challenges.

<table>
<thead>
<tr>
<th>Eval-A/B</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>C1</th>
<th>C2</th>
<th>D1</th>
<th>D2</th>
<th>E1</th>
<th>E2</th>
<th>F1</th>
<th>F2</th>
<th>G1</th>
<th>G2</th>
<th>Tot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-1st</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>11.0</td>
<td>11.0</td>
<td>3.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>All-2nd</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>11.0</td>
<td>11.0</td>
<td>3.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>EMEA</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>11.0</td>
<td>11.0</td>
<td>3.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Medline</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>11.0</td>
<td>11.0</td>
<td>3.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Patent</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>11.0</td>
<td>11.0</td>
<td>3.4</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The evaluation shown in the table 9.1 leads to the result:

- Solutions D1 and F1 have the best overall performances, F1 slightly ahead of D1 according to the evaluation “All-1+2”, which summarizes the frequencies of the best ranked solutions on rank 1 and 2.
- Only taking EMEA and Medline into consideration, this puts E1 into the best position.
- For the different languages:
  - F1 and F2 perform best in German,
  - D1 in Spanish and
  - F1 again in French.
- D1 is the best performing solution across the big four annotation groups.

Since the number of systems contributing into all the evaluations is not evenly distributed across the different CPL experiments, it is the case that experiments with less participants – in principle – have a higher weight, since the likelihood to obtain the first rank is larger. Therefore, the statistical parameters have been corrected and reduced for those CPL challenges, where lesser participants took part, i.e. the relative number of participants in comparison to the maximum number of participants has been used as the weight for adaptation.
Table 9.2: The table provides the same values as the previous table 9.1, but now the F-measure performances from evaluation A and B have been weighted differently: CLPs with less participants have now a lower weight according to the percentage of participation. These figures are fairer for those contributions that have been made in very popular CLPs.

The results in table 9.1 and 9.2 are very similar to the results from the previous evaluations with the exception that the solution F1 appears as one of the most performing solutions. This is due to the fact that it shows consistently high performances across a large number of challenges, but this performance did not show in the previous analyses. We assume that any solution that delivers contributions to most of the experiments may perform well on average, but may not perform necessarily best on individual experiments. Due to the harmonization of the performances, a solution that performs well on both tasks may get a better rating than a solution that has quite diverse performances on the different tasks.

Altogether the following results can be derived from the evaluation A+B according to table 9.2:

- As before: solutions D1 and F1 have the best overall performances, F1 slightly ahead of D1.
- E1 performs best on EMEA+Medline.
- For the different languages:
  - F1 and F2 perform best in German,
  - D1 in Spanish and
  - F1 again in French, now only slightly ahead of E1.
- D1 is the best performing solution across the big four annotation groups.
**Evaluation A/-t**

The solutions of the challenge participants have also been evaluated for their contribution of novel terms (see above, evaluation chapter). For this assessment, the known terms in the non-English corpus have been ignored and the annotation solutions have only been evaluated against those entries in the SSC that have not a correlate in the Mantra terminology. This evaluation is called evaluation “A/-t”, since it is a variant of evaluation A.

It becomes clear that the performances in evaluation A and evaluation A/-t differ, i.e. a solution can be a good „reproducer“ of annotations against the harmonised corpus (evaluation A for one corpus, one language, one semantic type), but would not necessarily serve as a good „innovator“ (evaluation A/-t for the same set).

Table 9.3 shows again the table of the best ranked solutions in different evaluation scenarios. The F-measure of the tagging solutions has been determined on the de-annotated corpora, then the different solutions have been ranked and the frequency of best-ranked solutions has been determined on the basis of different setups.

<table>
<thead>
<tr>
<th>Eval-A/t</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>C1</th>
<th>C2</th>
<th>D1</th>
<th>E1</th>
<th>F1</th>
<th>F2</th>
<th>G1</th>
<th>G2</th>
<th>Tot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-1st</td>
<td>5</td>
<td>6</td>
<td>65</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-2nd</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>14</td>
<td>16</td>
<td>21</td>
<td>4</td>
<td>10</td>
<td>2</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>EM/EA</td>
<td>3</td>
<td>4</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medline</td>
<td>2</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patent</td>
<td>20</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>1</td>
<td>16</td>
<td>7</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>anat</td>
<td>6</td>
<td>2</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chem</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diio</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lib</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>devl</td>
<td>1</td>
<td>7</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>geog</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>objx</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phen</td>
<td>7</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phys</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-1+2</td>
<td>10</td>
<td>2</td>
<td>11</td>
<td>29</td>
<td>25</td>
<td>16</td>
<td>3</td>
<td>2</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM/EA+Medl</td>
<td>6</td>
<td>6</td>
<td>35</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG:4-grp</td>
<td>3</td>
<td>14</td>
<td>4</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oth-grp</td>
<td>2</td>
<td>22</td>
<td>3</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 9.3:** This evaluation is based on the de-annotated corpus for the mention evaluation (Evaluation A).

The best solutions in this evaluation according to their listing as first ranked and second ranked solutions are D1 and E1 (see “All-1+2”, bottom part). Both approaches make use of term translation methods and existing semantic resources to identify novel terms. Other solutions that are based on machine-translation techniques did not perform as well, which may be explained by the assumption that these methods also produce uncommon terms. So altogether:

- Now D1 has the best overall performances (and not F1 as before).
- Again D1 performs best on EMEA+Medline (and not E1 as before).
- For the different languages:
  - D1 perform best in German (and not F1 and F2 as before),
  - as before D1 in Spanish and
- D1 is the best performing solution across the big four annotation groups.

Again, these measures mainly indicate that D1 and E1 do provide annotations that are in agreement with the SSC even if the known non-English terms have been removed from the corpora.

**Correlation of challenge contribution solutions**

In the last experiments, we have determined which annotation solutions correlate pairwise in their output. The table 9.4 shows the results for the performances in evaluation A and the table 9.5 for evaluation B.

In the first evaluation, the TP results and the error types (FN, FP) from each annotation solution against a CPL have been ranked, where the TP results and the error types have been ranked independently each, i.e. the annotation solutions have been ranked according to their TP results against a given CPL independently from ranking the same annotation solutions separately against the FP results. All ranks for one annotation solution have then been gathered in a single vector which has been used to measure the correlation. Measuring the correlation based on the F-measures was simply using the F-measures across all semantic groups for a given CPL to then measure the correlation for two annotation solutions.

![Table 9.4](image)

As can be seen from table 9.4, systems from the same team do correlate to a high degree, but also solutions that stem from different teams but use similar technologies, e.g. F-solutions and G-solutions. Only on the patent CPLs, correlations are lower than on the other corpora, since

---

Version 1.0 (28.11.2013)
the patent CPLs do not cover the biomedical domain knowledge in the same way as the other CPLs and the units in the patent CPLs are larger than on the EMEA and Medline CPLs.

If the correlation is based on the F-measure over the semantic groups for a given CPL, it becomes clear that correlated methods according to the first approach are actually rather anti-correlated, e.g. D1 versus F1, C1 and C2. The correlation according to the F-measure gives a better measure for the judgement of the similarity of two annotation solutions, whereas the correlation based on the ranking of the results and the error types shows consistency of the challenge approaches.

| Table 9.5: This table shows the same correlation results as in the previous table for the ranking of the results and the error types (left table) and for the F-measures, but now the correlation results is covering the data from evaluation B. |

The same analysis for the evaluation B shows a very similar result, but the correlation values are lower than reported for the evaluation A.

The correlation analysis, mainly the results based on the F-measure lead to the conclusion that:

- D1, E1 and F1 are well correlated on the German EMEA and Medline CPLs. This may be due to less term variability in German in comparison to the other languages.

- In the French CPLs, F1 and E1 correlate better than D1 and the other two annotation solutions.

- For Spanish, D1 and F1 correlate, and C1 and B1 both correlate.

- Surprisingly for Spanish CPLs, A1 and A2 show negative correlation.
10. Conclusion and Final Remarks

The Mantra challenge has been planned and executed as described in the project plan. In addition, the challenge has been executed in a faster pace than expected to comply with the timelines given by the CLEF conference cycle. The challenge had to be closed in May to achieve timely submission of the CLEF conference contributions, to offer the challenge participants to provide their input for the CLEF-ER workshop and to be able to discuss the results from the challenge at the workshop.

The project partners have achieved to reach the research community, to motivate members of the research domain from around the world to contribute to the challenge and received a significant number of submissions as part of the challenge. In preparation of the challenge, a number of corpora have been prepared and processed to deliver them as input to the challenge, and following the challenge further corpora have been developed to enable performance measurements of the challenge contributions. As a result, 10 Challenge-Language pairs are available:

1. EMEA in ‘de’, ‘fr’, ‘es’, and ‘nl’;

For all corpora an English SSC has been provided and now after the challenge the non-English SSCs have been produced for all corpora that help to judge the challenge contributions in evaluation A.

For the evaluation a number of analyses have been performed to determine whether

1. the evaluation measures show the same results leading to the conclusion that we measure the performances correctly,
2. alternative measurement scenarios have been tested to judge the annotation solutions according to different criteria and experimental settings, and
3. comparisons between the annotation solutions have been done to determine whether the annotation solutions differ in their results according to their performance profiles.

Overall, a few solutions turned out to be the highlights of the CLEF-ER challenge:

- D1 is a solution that is using mainly term translation and is the best ranked solution in a large number of CPLs. In particular for the identification of novel terms, D1 is the best performing solution, showing that this solution produces non-English terms in a very reliable way.
- F1 is the runner-up solution that also performs well across both evaluations, but is not one of the best performing solutions in the evaluation A and B only.
- The solution E1 performs well in the evaluation A and also – to a lesser degree – in the evaluation A/t, which shows that still a significant portion of relevant terms can be extracted from public resources such as Babelnet.
- Other solutions, such as G1 and G2, and also B1 and A2 have been well adapted to the CLEF-ER challenge without using lot of background knowledge about the biomedical
domain knowledge, and each deliver best ranked solutions on selected CPLs. In particular, G1 and G2 provide convincing results in the overall assessment in evaluation B.

Further work will be invested in the evaluation of the results to determine the relevance of components in the different tagging solutions that play an important role to achieve high performances in the CLEF-ER challenge.

Furthermore, the development of the Gold Standard Corpus (WP 3) will contribute to the understanding of the performances of the different solutions against the SSC and will give more details on the performances in the boundary detection and the assignment of the correct CUIs to the multi-lingual corpora.

Altogether, the CLEF-ER challenge (Mantra challenge) is certainly already now a significant contribution to the scientific community, since this challenge is the first challenge that contributed in a single round a large number of documents that have been annotated with a large number of entities in 5 different languages. The challenge participants had to choose from a wider selection of corpora and languages and had to take a decision on the tasks that they wanted to fulfill. Furthermore they had to investigate into the challenge how to annotate the multilingual documents.

Certainly it would be advantageous to extend this challenge to a larger number of languages, to different domains and to specific topics such as medical patient records. Future will tell, to which extend this can be achieved.
11. References


