Collaboratively Annotating Multilingual Parallel Corpora in the Biomedical Domain—some MANTRAs

Johannes Hellrich[1], **Simon Clematide**[2], **Udo Hahn**[1], **Dietrich Rebholz-Schuhmann**[2]

[1]Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany
{johannes.hellrich,udo.hahn}@uni-jena.de
[2]Institute for Computational Linguistics
University of Zürich, Zürich, Switzerland
{siclemat,rebholz}@cl.uzh.ch

Abstract

The coverage of multilingual biomedical resources is high for the English language, yet sparse for non-English languages—an observation which holds for seemingly well-resourced, yet still dramatically low-resourced ones such as Spanish, French or German but even more so for really under-resourced ones such as Dutch. We here present experimental results for automatically annotating parallel corpora and simultaneously acquiring new biomedical terminology for these under-resourced non-English languages on the basis of two types of language resources, namely parallel corpora (i.e. full translation equivalents at the document unit level) and (admittedly deficient) multilingual biomedical terminologies, with English as their anchor language. We automatically annotate these parallel corpora with biomedical named entities by an ensemble of named entity taggers and harmonize non-identical annotations the outcome of which is a so-called silver standard corpus. We conclude with an empirical assessment of this approach to automatically identify both known and new terms in multilingual corpora.

Keywords: Named Entity Recognition, Multilingual Terminologies, Silver Standard Corpus

1. Introduction

Biomedical terminologies assemble a huge amount of semantic metadata descriptors which span the whole range of conceptualizations relevant for the life sciences. They have shown their versatile usefulness and great importance in many application scenarios—ranging from biological database curation in molecular biology, e.g. gene/protein annotation (Camon et al., 2004), to clinical disease encoding (Spackman and Campbell, 1998) and patient record management (Campbell et al., 1997).

Despite the reasonable claim that terminologies should be designed in a language-independent way, in reality, they all rely on verbalizations in a specific natural language. Actually, the vast majority of these terminological systems are phrased in English. This can be beneficial e.g. for terminological homogenization, when sciences converge on an internationally shared *lingua franca* such as English for molecular biology. But clearly for hospitals, health insurance companies and (mostly non-expert) patients the medical sublanguage will always remain their own nation's native language—in the English-speaking as well as the non-English-speaking countries. Hence, there is an enormous need for interlingual communication beyond the limits of the English language within Europe and also worldwide.

There is, however, a striking lack of balance in the linguistic coverage of biomedical terminologies. Whereas English is very well covered in most of the relevant thematic areas in the life sciences, even otherwise well-resourced languages, such as Spanish, German or French, fall short of acceptable proportions of coverage in those areas, with loss rates of 60-90% (compared with the English coverage). Even worse, the wide range of definitely under-resourced languages (European ones such as Czech, Dutch, Turkish, Swedish or Polish and also many Non-European ones such as Hindi, Thai, Bengal, etc.) and, furthermore, the remaining low- and non-resourced languages (such as Bulgarian, Greek, etc.) have coverage loss rates between 95% to 99%, some of them even have no coverage at all (e.g. Croatian, Maltese, Latvian) for the life sciences. In essence, this means that the health care system of these countries is severely decoupled not only from the English-speaking biomedical community, and thus the much warranted interoperability of medical data (e.g. required in an age of increasing cross-border mobility of people and goods) is clearly out of sight.

That is the reason for massive investments into multilingual biomedical terminological resources. The classical approach—manual terminology development—is not only resource-costly in terms of time and money but obviously doomed to failure since the coverage loss data have not changed much for decades so that the terminology gaps have not been closed despite the necessity of such resources. Also due to the conceptual dynamics in the life sciences this situation is likely to get worse rather than get better in the future.

The MANTRA project[1] targets this scenario in that its main goal is the automatic enhancement of biomedical terminology resources for some selected non-English European languages. Starting from a massively trimmed version of the Unified Medical Language System (UMLS),[2] one of the most authoritative broad-band collections of terminology resources for the life sciences, and its English verbalizations of terms, in the MANTRA project methodological procedures are under development which help increase the more than limited coverage of Spanish, French, German and Dutch language terms within the UMLS.

[1]http://www.mantra-project.eu/
[2]http://www.nlm.nih.gov/research/umls/

The key idea is here to exploit three kinds of parallel corpora which contain sets of manually supplied pairwise direct translations of documents—titles from biomedical journal articles, drug product descriptions and claim sections from biomedical patents —for different kinds of lexical processing to generate translation equivalents from these sources.

To gather results for a wide array of approaches the MANTRA project organized the CLEF-ER challenge competition[3] within the framework of CLEF (Conference and Labs of the Evaluation Forum) 2013.[4] Participants were asked to provide biomedical entity annotations, grounded in a stripped down version of the current UMLS, for the parallel corpora. A multitude of approaches, ranging from dictionary-based term extraction over named entity recognition to phrasal alignment within statistical machine translation, was used by the participants.

The major methodological challenge for us was to harmonize the in-coming proposals for named entities and concepts—we defined a character-based metric which computes the term-wise overlap between all annotation contributions (Lewin et al., 2012; Lewin and Clematide, 2013).

Our work resulted in an entirely new type of language resource: a set of parallel corpora in English, French, Spanish, German and Dutch, all annotated for biomedical terms of a large variety. We call this outcome a *silver standard corpus* (SSC) (see also our previous work on an English-only annotated corpus within the CALBC project (Rebholz-Schuhmann et al., 2010; Rebholz-Schuhmann et al., 2011)), since, unlike human-developed gold standards, this collection of semantic metadata has automatically evolved on the basis of an ensemble of entity taggers. In the following, we will describe the resources required and procedures crucial for the construction of the silver standard (Section 2.), as well as the annotations contained in the SSC, both for known and new terms (Section 3.).

2. Multilingual Language Resources

The preparation work for the CLEF-ER challenge comprised the compilation of the parallel corpora and the multilingual terminological resources.

2.1. Multilingual Parallel Texts

Our parallel corpora which contain manually translated text units were compiled from three publicly available document repositories. They were chosen in order to increase the diversity of text genres and phrasings. The MEDLINE collection[5] contains bilingual titles from biomedical journal articles, which can be searched via PUBMED.[6] The multilingual EMEA documents (Tiedemann, 2009) provide consumer-oriented information on the usage of drugs.[7] The multilingual patent claims from the European Patent Office[8] focus on the technical and legal aspects of biomedical information. With the exception of Spanish and Dutch for patent claims, we were able to compile parallel documents from all three text genres mentioned above. Table 1 gives the basic statistics of the text units for each text genre. The available data from MEDLINE is not evenly distributed across the different languages, especially Dutch and to a lesser degree Spanish are not well represented there.

Lang	EMEA	MEDLINE	PATENT	All
Unit counts				
en	141k	1,594k	121k	1,856k
de	141k	719k	121k	981k
fr	141k	572k	121k	834k
es	141k	248k		389k
nl	141k	54k		195k
Word counts				
en	2,236k	15,776k	6,034k	24,046k
de	2,100k	5,997k	5,194k	13,291k
fr	2,598k	6,024k	6,690k	15,312k
es	2,504k	2,573k		5,077k
nl	2,263k	435k		2,698k

Table 1: Unit and word counts per language in all corpora. The MEDLINE titles are strictly bilingual, German/English and French/English titles are more frequent than Spanish/English. The multilingual EMEA corpus covers all languages. The patent claims are multilingual, however, they do not cover Spanish and Dutch. Patent units are whole paragraphs from the patent claims, all other units are segments of the size of sentences.

MEDLINE titles have an average length of about 8 to 10 words per unit. EMEA units (sentence-like segments) are a bit longer on average: 15 to 20 words per unit. Patent claim units are whole paragraphs (often in the form of a bullet list containing several sentences).

We expected all three text genres to be highly parallel regarding their semantic content. The translations of patent texts and EMEA drug labels should reflect the original content for legal or regulatory reasons. However, in the case of EMEA, we detected a substantial amount of non-parallelism in the original EMEA text collection due to imperfect conversion from PDF to text. Using a filtering approach based on the number of characters in potentially parallel text units, we had to remove about 243k units of the 364k original EMEA units that we started working with from (Tiedemann, 2009). Medline titles were partly translated into English by the original authors of the article, partly they were translated by third parties. Non-ASCII characters such as accented vowels in French or Spanish as well as German umlauts were not well represented in the original MEDLINE data. Therefore, we used a technique based on character n-grams to reconstruct the original orthography of the non-English MEDLINE titles as much as possible.

2.2. Multilingual Biomedical Terminology

The shared multilingual terminological resource (MTR)[9] (Rebholz-Schuhmann et al., 2013a) for the identification of

[3] http://www.mantra-project.eu/clef-er-challenge

[4] http://www.clef-initiative.eu

[5] http://mbr.nlm.nih.gov/Download/

[6] http://www.ncbi.nlm.nih.gov/pubmed

[7] http://opus.lingfil.uu.se/EMEA.php

[8] Source: IFI claims (http://ificlaims.com)

[9] The MTR is accessible (UMLS licence restrictions apply) through the submission site of the CLEF-ER challenge: http://www.clefer.org.

novel terms from the parallel corpora has been derived from the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004). The UMLS Metathesaurus incorporates over 100 biomedical terminologies, from which we selected the Medical Subject Headings (MESH), the Medical Dictionary for Regulatory Activities Terminology (MEDDRA, (Brown et al., 1999)) and the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT, (Stearns et al., 2001)).

In UMLS, terms are organized in synsets that are identified by a conceptual fix point, the so-called *Concept Unique Identifier (CUI)*. Each concept (or CUI) may have multiple names per language, these are called *synonyms* which also cover the different translations of a term. CUIs are categorized into 15 broader semantic groups.

We certainly did not want to provide the entire terminology, since it contains sets of terms that are either not relevant for the annotation of concepts in the biomedical literature or were deemed too problematic for the identification of multilingual biomedical terms. For example, the terms in the UMLS semantic group "Concepts & Ideas" (CONC) denote common English entities and concepts such as "contract" or "contract agreement" with less or low relevance for the annotation and translation of specific biomedical terminologies.

In order to choose the relevant semantic groups for inclusion in our MTR, all English corpora have been annotated with the full biomedical terminology and then all those semantic groups have been removed from the terminological resource that either contributed only a very small number of annotations (e.g., terms linked to genes), or that generated very unspecific annotations according to the manual inspection.

For the CLEF-ER challenge, the semantic groups "Activities and Behaviors" (ACTI), "Anatomy" (ANAT), "Chemicals and Drugs" (CHEM), "Devices" (DEVI), "Disorders" (DISO), "Geographic Areas" (GEOG), "Living Beings" (LIVB), "Objects" (OBJC), "Phenomena" (PHEN), and "Physiology" (PHYS) were kept. The MTR contains 531,466 concepts with 2,839,277 synonyms.

Table 2 shows a detailed breakdown of the multilingual coverage of the MTR. Some of the resources already have a very high coverage in one or more non-English languages. For instance, SNOMED-CT in Spanish, or MEDDRA in German, French and Spanish. However, for MESH all non-English languages are strongly under-resourced.

Terms	MESH	SNOMED-CT	MEDDRA
en	764,000	1,184,005	56,061
de	77,249	-	50,128
fr	105,758	-	49,586
es	59,678	1,089,723	49,499
nl	40,808	-	-

Table 2: Multilingual terminological resource: The English part of the TR contains most terms. Only Spanish is covered in SNOMED-CT. MEDDRA terms have been translated in all languages.

Individual annotations:

```
...and visceral adipose tissue is...
...and visceral adipose tissue is...
...and visceral adipose tissue is...
...and visceral adipose tissue is...
...and visceral adipose tissue is...
```

Inter-entity character counts and centroid:

```
a n d v i s c e r a l a d i p o s e t i s s u e i s
 0 0 0 2 2 2 2 2 2 2 2 5 5 5 5 5 5 5 4 4 4 4 4 4 0 0
```

Extended centroids with varying boundary thresholds:

Boundary Thresholds	E-Centroid
1 or 2	visceral adipose tissue
3 or 4	adipose tissue
5	adipose

Figure 1: Individual annotations, their centroids and extended centroids

2.3. CLEF-ER Challenge for Semantically Annotating Multilingual Corpora

In order to enrich the non-English part of our MTR with new synonyms and/or new translations, we followed a collaborative, corpus-based approach, the so-called CLEF-ER challenge. The objective of the challenge was the identification of mentions of named entities and biomedical concepts in multilingual biomedical corpora, including the attribution of CUIs from our MTR to these mentions.

2.3.1. Input Resources for the Challenge

The participants of the CLEF-ER challenge received the following input data from the organizers. First, the MTR in the OBO exchange format.[10] Second, the unannotated non-English parallel corpora. Third, the automatically annotated and harmonized English Silver Standard Corpus (SSC).

The creation of the English SSC for CLEF-ER and its properties are described in detail by Lewin and Clematide (2013). There are several reasons why an English SSC is useful for the enhancement of multilingual terminological resources. First, expert annotations for a broad-coverage gold standard annotation are costly and time-consuming and do not scale up to large corpora. Second, the coverage of English terminology resources and the performance of biomedical named entity taggers for English allow for an automatic annotation in a quality that alleviates the need of a gold standard. Third, an even more satisfactory level of automatic named entity annotation can be reached if the output of several systems is harmonized into an ensemble annotation, the so-called harmonized SSC. The harmonization avoids the inevitable biases and errors of any individual annotation solution.

For the alignment and harmonization of the output of several different entity taggers, we applied and adapted the centroid approach originally described in (Lewin et al., 2012). Figure 1 illustrates the character-based centroid harmonization. Each annotation adds one vote to the inter-entity pairs of adjacent characters (spaces are ignored). If a pre-determined *voting threshold* is reached, the span with

[10] http://www.geneontology.org/GO.format.obo-1_2.shtml

	EMEA			MEDLINE			PATENT		All
	de	es	fr	de	es	fr	de	fr	
A1		1							**1**
A2		1			1				**2**
A3				1	1	1			**3**
A4		1	1		1	1			**4**
A5	1		1	1		1	1	1	**6**
A6	1	1	1	1	1	1	1	1	**8**
A7	1	1	1	1	1	1	1	1	**8**
All	**3**	**5**	**4**	**4**	**5**	**5**	**3**	**3**	**32**

Table 3: Distribution of the challenge contributions (A1-7) for the non-English SSCs. Some contributors provided more than one annotation run for a corpus but only one run was selected for the SSC in order to prevent harmonization biases.

the highest number of votes is considered the centroid. The *boundary distribution* of a centroid is given by the character offsets to the left and right of the centroid where the number of votes changes. The value of a boundary is the difference in number of votes.

Although centroids and their boundary distributions are maximally informative, they could have been too complex and discouraging for the challenge participants. Therefore, we decided to transform the centroids into a classical markup format with single boundaries. In general, the boundaries of centroids cannot be taken as adequate mention boundaries for the enhancement of a terminology, because they represent only the shared core of an ensemble annotation. In order to include more lexical content, we decided to extend the centroids (*e-centroids*) to the left and right according to a pre-determined *boundary threshold*.

For the English SSC, 6 different annotations were available from the MANTRA project partners. A voting threshold of 3 and a boundary threshold of 2 was finally chosen. This setting kept 45% of all possible concept centroids (voting threshold 1). On average, 19% (standard deviation 14%) of the original annotations were removed. 97.8% of the partner annotations that went into the SSC had exactly the same boundaries as their e-centroids.

2.3.2. Exploiting the Challenge Outcome

Each challenge participant had to deliver at least one annotated non-English corpus. In total, seven annotation solutions were submitted to the challenge (Rebholz-Schuhmann et al., 2013b). Almost all contributing solutions exploited publicly available resources (UMLS, WordNet, Wikipedia), and – in addition – applied lexical lookup solutions or indexing of the terminological resources. Two groups translated the terms through public resources (i.e. BabelNet, Google Translate), and four systems made use of statistical machine translation methods or multilingual word alignment. Altogether, the used solutions showed high heterogeneity. Table 3 shows the distribution of annotation contributions across languages and corpora. Unfortunately, only two system annotated Dutch corpora which is the reason that we excluded this language for the terminology enhancement evaluations described below.

The challenge contributions were evaluated in two different ways. Evaluation A measured the annotations of an individual contribution against a non-English SSC built from all contributions on the level of mentions. Evaluation B compared the bag of CUIs in one unit against the bag of CUIs annotated in the unit of the parallel English SSC.

The exploitation of the challenge outcomes for the enhancement of the non-English terminology relies on non-English SSCs that were harmonized from the challenge contributions. However, for the purpose of terminology enhancement we are more interested in the subset of annotations that cannot be trivially linked to already existing entries in our provided MTR. Therefore, we produced a partially deannotated version of the challenge contributions where we removed such annotations. This material was then used to create deannotated SSCs according to two different voting threshold schemas. The *majority voting* schema requires a threshold of $V := \lfloor N/2 \rfloor + N \bmod 2$ where N is the number of contributions for a given corpus. The *fixed threshold voting* schema requires a minimal amount of votes. For our non-English corpora, a voting threshold of 2 was set.

3. Results

We performed both quantitative and selected analyses of the annotations in the SSC, investigating effects of harmonization methods, corpus types, corpus sizes and languages (focusing on German, French and Spanish).

3.1. Number of Annotations

We counted for each class the number of concepts (i.e. CUIs), terms and term occurrences, and calculated the ratios thereof, as well as counts normalized for corpus size. Findings have been normalized by removing diacritics and non-letter characters, and transforming them to a lowercase representation. Tables 5 and 6 in the appendix provide an overview on the number of annotations contained in the SSCs generated with threshold voting and majority voting, respectively. For our analysis we distinguish three annotation classes:

- **known**, i.e. the UMLS *contains* the annotated text as a term for the concept and language in question.
- **entirely new**, i.e. the UMLS *does not contain* the annotated text as a term for the concept, neither for English, nor for the language in question.
- **new, as English**, i.e. the UMLS *does not contain* the annotated text as a term for the concept and language in question, yet *contains* it for English—many of these terms are of Latin origin, e.g. the name of the fungus *Cephalosporium acremonium*.

Comparing harmonization methods: The largest portion of annotations by all metrics results from the *new* class if using threshold harmonization, yet for majority harmonization the *known* class dominates, except for Spanish concepts and terms. The overall numbers for the *known* class are comparable for both harmonization methods, threshold harmonization producing slightly higher numbers (about 10 percent). In contrast, numbers for the classes *as English* and *new* are far lower for the more conservative majority

voting. This difference is especially dramatic for the *new* class, with the majority harmonized SSC containing only about half as many concepts, a sixth of the terms and a quarter of the occurrences present in the threshold harmonized SSC. This is also reflected in the ratio of terms/concept, being very similar for the *known* (about 1.3) and *as English* classes (about 1.1) over all languages and corpora.

In contrast, results for the *new* class depend strongly on the harmonization method used—majority voting results in numbers around 1.2, whereas threshold voting results in ratios of 3.5 to 4.2. The ratio of occurrences/concept for the *new* class is also diverging based on the harmonization method, majority harmonization resulting in about half the value provided by threshold harmonization. In general, majority harmonization seems to result in *new* annotations behaving similar to those of the *known* or *as English* class, while threshold harmonization *new* annotations behave atypically, having both far more terms and occurrences per concept.

Comparing corpora: The German EMEA corpus and the French PATENT corpus provided surprisingly few *new* concepts and terms relative to their number of *new* occurrences, independently of the harmonization method being used; the inverse is true for Spanish MEDLINE (cf. the occurrences/concept column of Tables 5 and 6). MEDLINE is the dominant source of annotations in all three classes, probably due to its high corpus size, broad thematic spectrum and the annotation-friendly simple syntactic structure of the titles.

Regarding languages: Spanish has, independently of the harmonization method being used, about three times the *known* and twice the *new* concepts and terms per thousand words as other languages, thus its absolute number of annotated concepts and terms is comparable to those of German and French, despite its combined corpora having only 5M words, whereas German and French have 13M and 15M, respectively. The absolute number of *known* concepts and terms is similar for French and Spanish, while German is about 10 percent lower, again independently of the harmonization method used. *As English* terms and concepts are more frequent in German, especially for majority harmonization or MEDLINE titles, which could be caused by a greater openness to English loan words or more reliance on Greek and Latin medical terms.

Overall, the analysis of the SSC annotations leads to two questions: Why are there so many more *as English* synonyms in German than in other languages and is threshold voting too lax or is the abnormal number of terms and occurrences in the *new* class an accurate reflection of the corpora? While the latter question can only be answered by the creation of and evaluation against a GSC, the former can be answered by sampling the annotations of the *as English* class.

3.2. Breakdown of *as English* annotations

To better understand the occurrence of *as English* annotations in non-English texts and the comparatively high number of *as English* terms and concepts in German texts we sampled 100 randomly selected terms each for German, Spanish and French from the threshold harmonized SSC. We suspected internationally used Latin and Greek loanwords to be the main reason for the appearance of *as English* terms in general and a greater openness to English loanwords as the reason for the abnormally high rate in German corpora. We found the following explanations for *as English* terms occurring in non-English texts:

- Latin or Greek terms used internationally, e.g. "decubitus"; used only for terms which are inflected according to the original language and not for compounds or words formed by derivation with non-Latin/Greek material.
- Names of drugs, chemical compounds, persons or places, e.g. "Valoron".
- English words used internationally, e.g. "suspension".
- Abbreviation used internationally, e.g. "PCP" for pneumonia.
- Other, e.g. random similarity like "perimeters" which could be both an English plural or a German genitive of the Greek loanword.

German behaved according to our expectation, with Latin/Greek words making up the majority of *as English* terms, whereas those made up only a minority of the *as English* terms for French and Spanish (cf. Table 4). French *as English* terms are quite often French terms which are missing in the terminology, yet appear, due to diacritics being removed during normalization, to be English terms. Spanish *as English* terms are most often real English terms. Some of these cases are due to wrong language identification in the EMEA corpus, e.g. the following sentence being listed as Spanish: "Dogs Treatment of pain".

Overall no clear explanation for the differences in the frequency of *as English* terms could be found, and surprisingly German seems to be much more open to Latin and Greek terms than the two Romance languages. A possible explanation are differences in the existing terminologies, e.g. the Spanish terminology already containing many Latin/Greek terms leading to few new ones being found, yet further investigating the etymology of UMLS entries is out of scope for this paper.

Cause	de	fr	es
Latin/Greek	34	18	18
Name	31	16	10
English	20	33	51
Abbreviation	13	10	17
Other/Native	2	23	4

Table 4: This table lists the frequency of explanations for the occurrence of *as English* terms in German, French and Spanish texts, based on a sample of 100 terms each. We distinguish the following explanations: Latin/Greek term used internationally, name (of e.g. a drug), English word used internationally, abbreviation used internationally and other (e.g. random similarity).

4. Conclusions

The exploitation of parallel SSCs for the generation of multilingual terminological resources is a new approach which enables normalization of the term candidates against an existing terminological resource.

Future work will include the creation of a small GSC, allowing us to assess the quality of the SSC, and to refine our harmonization process to find a good balance between the number and quality of new terms. We also plan to use the multilingual annotations to enrich the underlying terminological resource with new non-English entries and assess the impact of an enhanced terminology on other applications, e.g. machine translation.

The SSCs described in this paper will be made publicly available in Summer 2014 via ELRA.

5. Acknowledgements

This work was funded by the EU STREP project grant 296410 ("MANTRA") under the 7th EU Framework Programme within Theme "Information Content Technologies, Technologies for Digital Content and Languages" [FP7-ICT-2011-4.1]. Patent data supplied by IFI Claims Patent Services.

6. References

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32:D267–270, Jan.

Brown, Elliot G, Wood, Louise, and Wood, Sue. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117.

Camon, Evelyn, Magrane, Michele, Barrell, Daniel, Lee, Vivian, Dimmer, Emily, Maslen, John, Binns, David, Harte, Nicola, Lopez, Rodrigo, and Apweiler, Rolf. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266.

Campbell, James R., Carpenter, Paul, Sneiderman, Charles A., Cohn, Simon, Chute, Christopher G., Warren, Judith, and CPRI Work Group on Codes and Structures. (1997). Phase II evaluation of clinical coding schemes: Completeness, taxonomy, mapping, definitions, and clarity. *Journal of the American Medical Informatics Association*, 4(3):238–250, May.

Lewin, Ian and Clematide, Simon. (2013). Deriving an english biomedical silver standard corpus for CLEF-ER. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, 23-26 September, Valencia - Spain*.

Lewin, Ian, Kafkas, Senay, and Rebholz-Schuhmann, Dietrich. (2012). Centroids: Gold standards with distributional variations. In *LREC '12 – Proceedings of the 2012 Language Resources and Evaluation Conference*, Istanbul, Turkey.

Rebholz-Schuhmann, D., Yepes, A. J., Mulligen, E. M. Van, Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., and Hahn, U. (2010). CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology*, 8(1):163–179, Feb.

Rebholz-Schuhmann, D., Jimeno-Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J.B., Baker, C.J.O., Kuo, C.-J., Clematide, S., Rinaldi, F., Farkas, R., Mra, G., Hara, K., Furlong, L., Rautschka, M., Lara Neves, M., Pascual-Montano, A., Wei, Q., Collier, N., Mahbub Chowdhury, Md. F., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J.L., van Mulligen, E., Kors, J., and Hahn, U. (2011). Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *Journal of Biomedical Semantics*, 2(Suppl 5):S11, Oct.

Rebholz-Schuhmann, Dietrich, Clematide, Simon, Rinaldi, Fabio, Kafkas, Senay, van Mulligen, Erik M., Bui, Chinh, Hellrich, Johannes, Lewin, Ian, Milward, David, Poprat, Michael, Jimeno-Yepes, Antonio, Hahn, Udo, and Kors, Jan A. (2013a). Multilingual semantic resources and parallel corpora in the biomedical domain: the CLEF-ER Challenge. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, 23-26 September, Valencia - Spain*.

Rebholz-Schuhmann, Dietrich, Clematide, Simon, Rinaldi, Fabio, Kafkas, Senay, van Mulligen, Erik M., Bui, Chinh, Hellrich, Johannes, Lewin, Ian, Milward, David, Poprat, Michael, Jimeno-Yepes, Antonio, Hahn, Udo, and Kors, Jan A. (2013b). Entity recognition in parallel multi-lingual biomedical corpora: the CLEF-ER laboratory overview. In Forner, Pamela, Müller, Henning, Paredes, Roberto, Rosso, Paolo, and Stein, Benno, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings*, Lecture Notes in Computer Science, pages 353–367. Springer.

Spackman, Kent A. and Campbell, Keith E. (1998). Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. In Chute, Christopher G., editor, *AMIA '98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century. Orlando, FL, November 7-11, 1998*, pages 740–744, Philadelphia/PA. Hanley & Belfus.

Stearns, Michael Q, Price, Colin, Spackman, Kent A, and Wang, Amy Y. (2001). SNOMED Clinical Terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Tiedemann, Jörg. (2009). News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. Amsterdam, Philadelphia: John Benjamins.

Corpus	Language	Class	Concepts	Conc./1k words	Terms	Terms/1k words	Occurrences	Occ./1k words	Terms/Conc.	Occ./Conc.
EMEA	de	known	4,576	2.18	5,497	2.62	123,469	58.79	1.20	26.98
		new	5,344	2.54	7,370	3.51	354,715	168.91	1.38	66.38
		as English	2,396	1.14	2,552	1.22	71,896	34.24	1.07	30.01
	es	known	7,082	2.83	8,774	3.50	198,947	79.45	1.24	28.09
		new	12,717	5.08	36,936	14.75	630,039	251.61	2.90	49.54
		as English	1,890	0.75	2,014	0.80	52,623	21.02	1.07	27.84
	fr	known	5,014	1.93	6,326	2.43	136,125	52.40	1.26	27.15
		new	8,655	3.33	30,503	11.74	539,677	207.73	3.52	62.35
		as English	2,373	0.91	2,574	0.99	91,149	35.08	1.08	38.41
Medline	de	known	15,743	2.63	19,954	3.33	412,498	68.78	1.27	26.20
		new	42,630	7.11	185,017	30.85	1,573,407	262.37	4.34	36.91
		as English	11,161	1.86	12,575	2.10	103,033	17.18	1.13	9.23
	es	known	19,231	7.47	24,423	9.49	313,022	121.66	1.27	16.28
		new	33,065	12.85	109,120	42.41	688,953	267.76	3.30	20.84
		as English	3,301	1.28	3,538	1.38	35,043	13.62	1.07	10.62
	fr	known	18,373	3.05	25,370	4.21	530,095	88.00	1.38	28.85
		new	45,325	7.52	145,005	24.07	1,731,777	287.48	3.20	38.21
		as English	9,308	1.55	10,446	1.73	204,305	33.92	1.12	21.95
Patent	de	known	4,135	0.80	4,612	0.89	75,889	14.61	1.12	18.35
		new	5,514	1.06	7,511	1.45	301,380	58.02	1.36	54.66
		as English	2,404	0.46	2,560	0.49	46,615	8.97	1.06	19.39
	fr	known	4,891	0.73	5,823	0.87	223,155	33.36	1.19	45.63
		new	6,110	0.91	9,185	1.37	794,671	118.78	1.50	130.06
		as English	2,690	0.40	2,872	0.43	182,602	27.29	1.07	67.88
all	de	known	17,057	1.28	21,861	1.64	611,856	46.04	1.28	35.87
		new	45,100	3.39	190,631	14.34	2,229,502	167.75	4.23	49.43
		as English	13,071	0.98	14,799	1.11	221,544	16.67	1.13	16.95
	es	known	21,286	4.19	27,563	5.43	511,969	100.84	1.29	24.05
		new	39,218	7.72	141,022	27.78	1,318,992	259.80	3.60	33.63
		as English	4,527	0.89	4,926	0.97	87,666	17.27	1.09	19.37
	fr	known	20,020	1.31	28,152	1.84	889,375	58.08	1.41	44.42
		new	49,515	3.23	171,561	11.20	3,066,125	200.24	3.46	61.92
		as English	11,320	0.74	12,738	0.83	478,056	31.22	1.13	42.23

Table 5: This table gives an overview on the identification of terms in the non-English SSCs harmonized by **threshold** voting, distinguishing three classes of annotations: *known* i.e. contained in the UMLS as a term for this language, *new* i.e. not contained in the UMLS as a term, neither for this language nor English and *as English* i.e. not contained in the UMLS as a term for this language, yet for English—mostly Latin terms. We list for each combination of class, corpus and language the number of concepts, terms and occurrences annotated in the SSC, both in absolute numbers and normalized per thousand words. We also list the ratios of terms and occurrences per concept.

Corpus	Language	Class	Concepts	Conc./1k words	Terms	Terms/1k words	Occurrences	Occ./1k words	Terms/Conc.	Occ./Conc.
EMEA	de	known	4,513	2.15	5,408	2.58	123,076	58.61	1.20	27.27
		new	1,907	0.91	2,059	0.98	87,975	41.89	1.08	46.13
		as English	1,130	0.54	1,165	0.55	46,344	22.07	1.03	41.01
	es	known	6,307	2.52	7,800	3.12	138,243	55.21	1.24	21.92
		new	3,887	1.55	4,680	1.87	174,920	69.86	1.20	45.00
		as English	1,022	0.41	1,066	0.43	34,455	13.76	1.04	33.71
	fr	known	5,275	2.03	6,673	2.57	147,460	56.76	1.27	27.95
		new	5,409	2.08	6,448	2.48	184,789	71.13	1.19	34.16
		as English	1,397	0.54	1,465	0.56	57,459	22.12	1.05	41.13
Medline	de	known	15,874	2.65	20,066	3.35	448,442	74.78	1.26	28.25
		new	24,585	4.10	29,988	5.00	318,494	53.11	1.22	12.95
		as English	8,956	1.49	9,607	1.60	54,286	9.05	1.07	6.06
	es	known	17,464	6.79	22,045	8.57	276,586	107.50	1.26	15.84
		new	14,973	5.82	17,329	6.73	105,516	41.01	1.16	7.05
		as English	1,919	0.75	2,003	0.78	12,839	4.99	1.04	6.69
	fr	known	17,121	2.84	22,984	3.82	489,760	81.30	1.34	28.61
		new	22,580	3.75	25,825	4.29	311,403	51.69	1.14	13.79
		as English	3,776	0.63	3,924	0.65	53,876	8.94	1.04	14.27
Patent	de	known	4,092	0.79	4,560	0.88	75,185	14.48	1.11	18.37
		new	1,739	0.33	1,849	0.36	62,649	12.06	1.06	36.03
		as English	457	0.09	464	0.09	7,284	1.40	1.02	15.94
	fr	known	4,992	0.75	5,918	0.88	217,885	32.57	1.19	43.65
		new	2,017	0.30	2,266	0.34	204,705	30.60	1.12	101.49
		as English	702	0.10	750	0.11	78,634	11.75	1.07	112.01
all	de	known	17,102	1.29	21,851	1.64	646,703	48.66	1.28	37.81
		new	25,436	1.91	31,050	2.34	469,118	35.30	1.22	18.44
		as English	9,590	0.72	10,293	0.77	107,914	8.12	1.07	11.25
	es	known	19,260	3.79	24,804	4.89	414,829	81.71	1.29	21.54
		new	17,558	3.46	20,855	4.11	280,436	55.24	1.19	15.97
		as English	2,734	0.54	2,872	0.57	47,294	9.32	1.05	17.30
	fr	known	18,933	1.24	26,019	1.70	855,105	55.85	1.37	45.16
		new	26,034	1.70	30,527	1.99	700,897	45.77	1.17	26.92
		as English	4,600	0.30	4,835	0.32	189,969	12.41	1.05	41.30

Table 6: This table gives an overview on the identification of terms in the non-English SSCs harmonized by **majority** voting, distinguishing three classes of annotations: *known* i.e. contained in the UMLS as a term for this language, *new* i.e. not contained in the UMLS as a term, neither for this language nor English and *as English* i.e. not contained in the UMLS as a term for this language, yet for English—mostly Latin terms. We list for each combination of class, corpus and language the number of concepts, terms and occurrences annotated in the SSC, both in absolute numbers and normalized per thousand words. We also list the ratios of terms and occurrences per concept.