A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC

RECEIVED 12 November 2014
REVISED 11 March 2015
ACCEPTED 29 March 2015
PUBLISHED ONLINE FIRST 6 May 2015

Jan A Kors,[1] Simon Clematide,[2] Saber A Akhondi,[1] Erik M van Mulligen,[1] Dietrich Rebholz-Schuhmann[2]

ABSTRACT

Objective To create a multilingual gold-standard corpus for biomedical concept recognition.

Materials and methods We selected text units from different parallel corpora (Medline abstract titles, drug labels, biomedical patent claims) in English, French, German, Spanish, and Dutch. Three annotators per language independently annotated the biomedical concepts, based on a subset of the Unified Medical Language System and covering a wide range of semantic groups. To reduce the annotation workload, automatically generated preannotations were provided. Individual annotations were automatically harmonized and then adjudicated, and cross-language consistency checks were carried out to arrive at the final annotations.

Results The number of final annotations was 5530. Inter-annotator agreement scores indicate good agreement (median *F*-score 0.79), and are similar to those between individual annotators and the gold standard. The automatically generated harmonized annotation set for each language performed equally well as the best annotator for that language.

Discussion The use of automatic preannotations, harmonized annotations, and parallel corpora helped to keep the manual annotation efforts manageable. The inter-annotator agreement scores provide a reference standard for gauging the performance of automatic annotation techniques.

Conclusion To our knowledge, this is the first gold-standard corpus for biomedical concept recognition in languages other than English. Other distinguishing features are the wide variety of semantic groups that are being covered, and the diversity of text genres that were annotated.

Keywords: gold-standard corpus, multilinguality, inter-annotator agreement, concept identification, semantic enrichment

RESEARCH AND APPLICATIONS

BACKGROUND AND SIGNIFICANCE

Introduction

Huge amounts of biomedical information are only available in textual form, such as in scientific publications, electronic health records, and patents.[1] The sheer volume of these unstructured sources makes it impossible for researchers, physicians, or database curators to keep abreast of all information that is being poured out. Natural language processing systems hold promise for facilitating the time-consuming and expensive manual information extraction process, or even for automatically generating new hypotheses and other insights.

An important step in the information extraction task is the recognition and normalization of relevant terms in a text.[2] Term recognition aims at finding text strings that refer to entities or concepts, and marking each term with a semantic type, like "gene," "drug," or "disease." Term normalization or concept recognition is more complex than term recognition only. It assigns a unique identifier to the recognized term, which links it to a source that contains further information about the concept, such as its definition, its preferred name, and synonyms, and its relationships with other concepts. While many terminological resources are available for English, other languages are far more underresourced.

To train and evaluate automated concept recognition methods, manually annotated "gold-standard" corpora (GSCs) are essential. However, the creation of a GSC is a cumbersome and complex task. This is already true for the annotation of term boundaries, but the difficulty is compounded when terms have to be mapped to concepts in terminological resources. Complexity further increases when annotations are to be done in different languages, by several annotators. As a consequence, there are few GSCs available that contain annotations of concepts, and they mostly annotate only concepts that belong to a limited set of semantic groups, such as "disorder" or "gene or protein." The assessment of concept recognition systems for a broad range of semantic groups, and the evaluation of ensemble approaches that combine the results of multiple annotation systems into a "silver standard,"[3,4] require new GSCs.

Furthermore, the available GSCs only contain documents in the English language. To develop and evaluate natural language processing methods for biomedical concept recognition in non-English languages, GSCs in these languages are needed. Moreover, if such GSCs are based on parallel corpora, they will be helpful in assessing methods for automatic enrichment of terminologies, especially for the non-English languages.

In this paper we describe the generation of a GSC for biomedical concept recognition in five different languages (English, French, German, Spanish, and Dutch), based on parallel corpora representing different biomedical subdomains. The annotations are based on a subset of the Unified Medical Language System (UMLS)[5] and cover a wide variety of semantic groups. We present an elaborate annotation process that involves the use of automatically generated preannotations to help reduce the annotation overload, and the harmonization

Correspondence to: Jan A. Kors, Department of Medical Informatics, Erasmus University Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands
j.kors@erasmusmc.nl; Tel: +31-10-7043045

For numbered affiliations see end of article.

and curation of the annotations of multiple annotators for each of the languages. We also provide concept annotation statistics and inter-annotator agreement scores. This work has been part of the EU-funded Mantra project, aimed at providing enriched multilingual biomedical terminologies and semantically annotated multilingual documents for a wide range of semantic types.[6,7]

Related work

There are several biomedical corpora that provide concept annotations. The Arizona Disease Corpus[8] contains 2784 sentences from Medline abstracts annotated with disease mentions and mapped to UMLS concept unique identifiers (CUIs). Gurulingappa *et al.*[9] annotated mentions of diseases and adverse events and their corresponding UMLS CUIs, in a set of 4272 sentences from Medline abstracts describing case reports. The Colorado Richly Annotated Full-Text corpus[10] consists of 97 full-text biomedical articles with concept annotations from nine ontologies and terminologies, including Chemical Entities of Biological Interest, Gene Ontology, and National Center for Biotechnology Information Taxonomy. The Shared Annotated Resource corpus[11] is composed of 298 clinical notes annotated for disorder mentions and normalized to CUIs from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). In several BioCreative challenges, annotated corpora of Medline abstracts and full-text articles were created for gene normalization tasks, linking genes, and gene products to Entrez Gene database identifiers.[12–14] However, the annotations of the gene mentions in BioCreative are not incorporated into the sentences, that is, they are provided at the document level.

All of these corpora are in English, and concern only one or a limited set of semantic types. In an attempt to overcome this latter limitation, the Collaborative Annotation of a Large Biomedical Corpus project automatically generated silver standard corpora by combining the annotations of different concept recognition systems on a set of about one million Medline abstracts.[3] While silver-standard corpora offer a number of advantages, in particular a large amount of annotations and a wide coverage and variety of semantic groups, for the task of performance assessments they cannot replace gold-standard annotations yet.[4]

Very few biomedical corpora with concept annotations are available in languages other than English. The MuchMore corpus contains about 9000 bilingual (English-German) medical abstracts, which were annotated with Medical Subject Headers (MeSH) CUIs, albeit by automatic means.[15] We are not aware of any non-English biomedical GSC that contains concept annotations.

MATERIALS AND METHODS

Corpus selection

The GSC is based on parallel corpora for three text types that were collected in the Mantra project: scientific abstract titles, drug labels, and biomedical patent claims.[6] The languages of interest in the Mantra project were English, German, French, Spanish, and Dutch. Abstract titles have been taken from Medline and are bilingual, always in English and one of the other languages. The drug label corpus consists of parallel documents from the European Medicines Agency, and are available through the Open Source Parallel Corpus collection.[16] The drug labels are available in all five languages. Patents of the European Patent Office were selected from the IFI CLAIMS patent database[17] by querying for the International Patent Classification code A61K ("Preparations for medical, dental, or toilet purposes"). The patents are available in English, German, and French in parallel.

Table 1: Number of units and words (in parentheses) for the different languages in the Mantra GSC.

Language	Medline titles				Drug labels	Patents
English	100 (1119)	100 (1165)	100 (1112)	100 (1020)	100 (1995)	50 (3224)
French	100 (1218)				100 (2391)	50 (3597)
German		100 (947)			100 (1956)	50 (3117)
Spanish			100 (1256)		100 (2245)	
Dutch				100 (922)	100 (2055)	

Each document in the Mantra corpora has been decomposed into one or more units of text, where a unit can be a title (Medline abstracts, varying between 54 483 units for Dutch and 1 593 546 units for English) or a sentence (drug labels: 129 567 units for each language; patents: 154 836 units for English, French, and German). From each Mantra corpus, parallel units were randomly selected for constructing the GSC (Table 1): 100 units from each set of bilingual abstract titles (400 parallel units in total), 100 units from the drug labels, and 50 units from the patents. For English, this resulted in a total of 550 units, for French and German in 250 units, and for Spanish and Dutch in 200 units. For English, the average number of words per unit was about 11 for the Medline titles, 20 for the drug labels, and 64 for the patents. These average numbers were slightly higher for French and Spanish, and lower for German and Dutch. A separate set of 20 English units (11 titles, 5 labels, 4 patents) was selected for the development of annotation guidelines.

Terminology

The annotators had to make their annotations in conformity to the terminology that was used in the Mantra project. Briefly, the Mantra terminology contains a subset of the UMLS, consisting of all concepts from three terminologies: MeSH, SNOMED-CT, and the Medical Dictionary for Regulatory Activities (MedDRA). MeSH is a comprehensive controlled vocabulary used to index and search articles and books in the life sciences, SNOMED-CT is the most extensive clinical healthcare terminology currently available, and MedDRA is a terminology used for classification of medical products and adverse events. For each concept from these three terminologies, all terms together with their semantic type and CUI were included in the Mantra terminology if the semantic type of the concept belonged to any of the following semantic groups[18]: Anatomy, Chemicals and drugs, Devices, Disorders, Geographic areas, Living beings, Objects, Phenomena, Physiology, and Procedures. Note that other terminologies in the UMLS may be covered in part by the Mantra terminology in as far as the concepts in these terminologies are also contained in MeSH, SNOMED-CT, or MedDRA. The Mantra terminology includes 591 918 concepts with a total of 3 238 015 terms, most of which are English (2 039 988), followed by Spanish (785 083).

Annotation guidelines

Annotation guidelines were established based on the 20 units that were selected for development purposes. A detailed description of the guidelines with annotation examples is provided as supplementary material. Briefly, the annotators were supplied with automatically generated preannotations (see Annotation process for details). In case of alternative preannotations for the same span of text, the annotators had the task to disambiguate for a single annotation. For example, in

the phrase "intraocular pressure should be monitored," the term "intraocular pressure" was preannotated with the CUIs C0021888 (preferred term "Intraocular pressure") and C0595921 ("Disorder of intraocular pressure"). Since the context did not indicate a disorder, the latter concept was removed. If the semantic difference between the suggested concepts could not be resolved, all annotations were kept. For example, "thyroid cancer" was preannotated as C0007115 ("Malignant neoplasm of thyroid") and as C0549473 ("Thyroid carcinoma"). Since "thyroid cancer" is synonymous with both concepts, the annotations were kept.

When one term was nested within another term, only the most specific and informative term was annotated. For instance, in " . . . subjected to partial resection of the small intestine," "partial resection" was annotated (as C0184908 "Partial excision"), while "resection" (C0728940 "Excision") was not.

A subword (part of a word) was annotated if the subword mapped to a concept in the Mantra terminology and the full word did not. This could happen for compound terms, as are common in German and Dutch. For example, in the German word "Arzneimittelüberwachungsplan" ("plan for drug monitoring"), the subword "Arzneimittelüberwachung" was annotated as C0085421 ("Drug monitoring").

Again, only those concepts have been annotated that could be resolved to the Mantra terminology. For example, the term "postoperative hypovolemia" refers to a concept in the UMLS (C1409762), which is only based on the International Classification of Primary Care, second edition (ICPC-2). Since this concept is not part of the Mantra terminology, the term is not annotated. Instead, "hypovolemia" (C0546884), which is included in the Mantra terminology, is annotated.

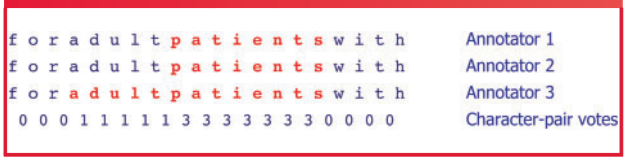
Discontiguous spans of text could be mapped to a single concept. For instance, in the phrase "swelling of the face and/or lips," the part "swelling of the face" is annotated as C0151602 ("Facial swelling"), and the two text spans "swelling of the" and "lips" are annotated as fragments that map to the single concept C0240211 ("Lip swelling").

Annotation process

Annotators independently annotated the units of each language using the brat rapid annotation tool.[19] Brat was configured in different ways for the various steps of the annotation process described below.

To reduce as much as possible the annotators' workload, preannotations of concepts were provided for each unit. A preannotation provides the span of text together with the assigned CUI, the concept's name, and its semantic type and group (all given by the Mantra terminology). The preannotations were constructed by harmonizing the annotations from five concept recognition systems (four systems[20–23] covered all five languages, one[24] only English). These systems participated in the Conference and Labs of the Evaluation Forum-Entity Recognition (CLEF-ER) challenge[25] and provided concept annotations for the multilingual Mantra corpora, from which the GSC was drawn. Harmonization, that is, combining the annotations of multiple annotators into a single annotation, was performed with the e-centroid method.[26,27] In short, the text is tokenized at the character level, spaces are ignored, and votes are counted over pairs of adjacent inter-term characters in the set of annotations. Centroids are defined as the substrings over character pairs with votes equal to or above a first threshold. In addition, the left and right boundaries of the centroids may be extended subject to a second threshold, yielding the extended centroid or e-centroid (Figure 1 illustrates the method). For the preannotations, we used low, recall-oriented harmonization thresholds (ec21: e-centroids with a first threshold of 2 and a second threshold of 1).

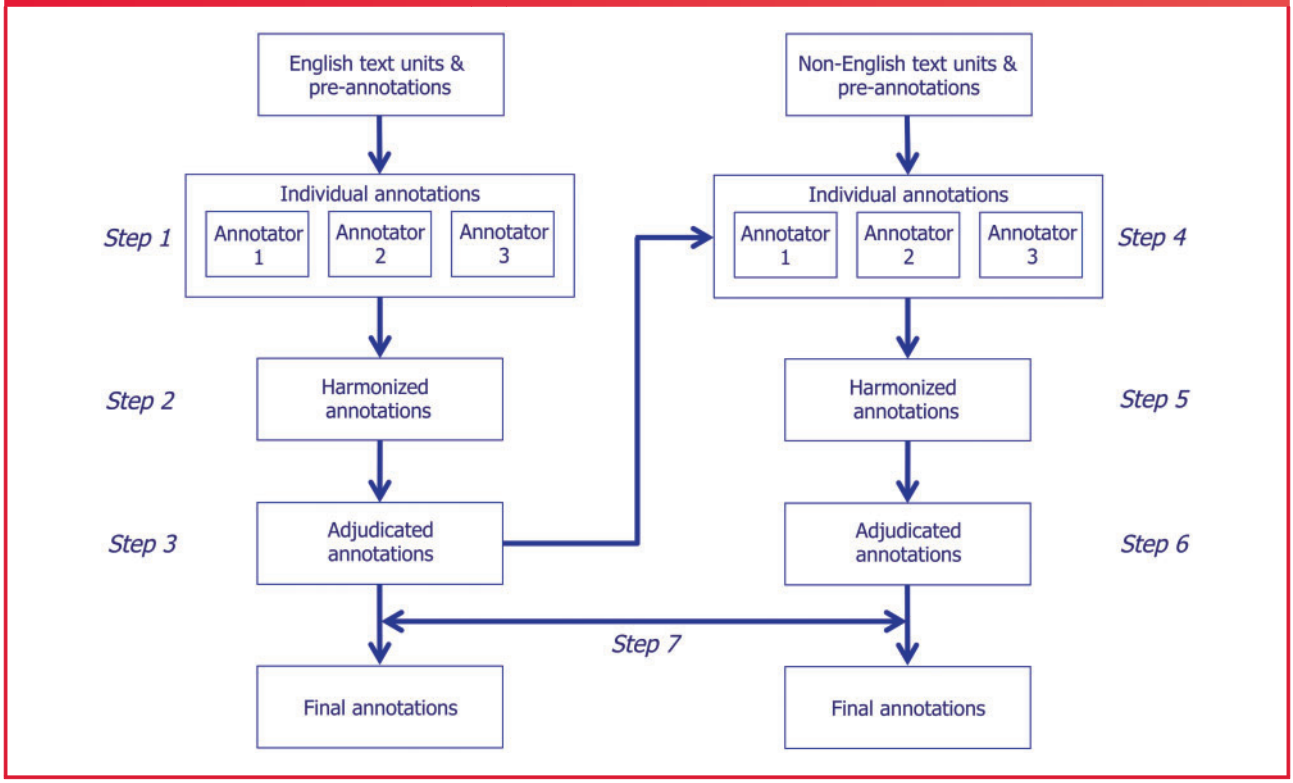
Figure 1: Example of harmonization by the e-centroid method. Two annotators annotated "patients," one annotated "adult patients," resulting in the character-pair votes shown on the last line. With a centroid threshold of 2 and a boundary threshold of 2, the harmonized annotation is "patients"; the same centroid threshold with a boundary threshold of 1 results in "adult patients."



The annotation process consisted of the following steps (see also Figure 2):

1. The English units were independently annotated by three annotators. Each had to correct wrong preannotations and had to add missing annotations. For background information on spans of text or terms, annotators had access to the UMLS Terminology Services[28] and the Mantra terminology.
   Figure 3 shows two screenshots of the brat tool for the annotation of an English unit. Information on the (pre-)annotated concepts is presented when hovering the cursor over the annotations. A double-click on a word or phrase shows a window that allows to make modifications or to link out to further information.
2. A harmonized set of English annotations was produced by calculating the e-centroids from the individual annotations; both thresholds were set to 2 (i.e., annotations were accepted if there was a majority vote).
3. All discrepancies between the individual English annotations and the harmonized annotations were discussed and, where necessary, the harmonized annotations were changed or removed, or new annotations were added.
   Figure 4 shows the annotations that have been made by all three annotators for the same text unit as in Figure 3. The aggregated view was helpful to identify and resolve the annotation discrepancies between the annotators.
4. The non-English units were now independently annotated by three annotators per language. The annotators received preannotations made in the specific language and, in addition, could view the curated harmonized annotations of the corresponding English unit (see previous step).
   As an example, Figure 5 shows the English unit together with the harmonized English annotations, and the corresponding German unit, with the German preannotations. Obviously, the annotations for "erwachsenen" (C0001675) and "eingeschränkter Nierenfunktion" (C0341697) are lacking and have to be added.
   It was common that unit pairs had more English than non-English annotations, which suggests that new annotations had to be added to the non-English units. However, a non-English annotation could also initiate a new annotation in the English unit. For example, in Spanish "En caso de deterioro de la función renal, . . . ", the term "deterioro de la función renal" was preannotated (C1278220, "Deteriorating renal function"). In the corresponding English unit ("In the case of renal function deterioration, . . . "), the term "renal function" (C0232804) resulted from the harmonized annotation and

Figure 2: Flow diagram of the annotation process. The different annotation steps are described in the text. The steps on the right-hand side are done for each non-English language separately.



was now replaced by "renal function deterioration" (C1278220), which is more specific. Note that the term "renal function deterioration" was not part of the synonym list for this concept in the Mantra terminology.

5. For each non-English language, a harmonized set of annotations was generated from the individual annotations.
6. For each non-English language, all discrepancies between the individual non-English and the harmonized annotations were discussed and, where necessary, the harmonized annotations were changed or removed, or new annotations were added.
7. In the last step, across all English and non-English units, any remaining CUI discrepancies in parallel units were discussed and resolved.

For the same example unit as above, Figure 6 shows the curated annotations for all languages.

RESULTS

For each of the five languages, three annotators independently annotated the units in that language. In total 12 annotators were involved: nine annotated one language (two annotators worked as a team, each annotating a different part of the units), two annotated two languages, and one annotated three languages. All annotators had a biomedical background and were fluent in the languages they worked on. The curation of the harmonized annotations and the final cross-language checking was done by two annotators (J.K. and D.R.S.).

For each language, we computed the inter-annotator agreement scores. In addition, we determined the agreement of the three annotators, of the preannotated set, and the automatically harmonized annotation set with the final gold-standard set (Table 2). Two annotations were considered in agreement if the CUIs as well as the annotated term boundaries were exactly the same. We used the *F*-score between two annotators,[10,11] since other agreement measures – in particular the kappa coefficient – require additional categorical data and do not apply to concept annotation agreements. Note that the *F*-score (harmonic mean of recall and precision) is invariant to the choice of annotator serving as the reference when computing precision and recall.

Overall, the annotators showed good agreements between each other (median *F*-score is 0.79). Similarly good agreements are

Table 2: Inter-annotator agreement and agreement against the final gold standard set for the different languages.

Annotators	Agreement (*F*-score)				
	English	French	German	Spanish	Dutch
1/2	0.83	0.78	0.84	0.79	0.85
1/3	0.87	0.80	0.63	0.78	0.74
2/3	0.86	0.83	0.64	0.79	0.75
Preannotated/Final	0.73	0.52	0.50	0.60	0.43
1/Final	0.80	0.77	0.86	0.76	0.79
2/Final	0.79	0.82	0.84	0.78	0.85
3/Final	0.85	0.86	0.66	0.86	0.79
Harmonized/Final	0.84	0.86	0.85	0.84	0.83

Figure 3: Example of an English unit from the Mantra GSC with preannotated concepts, color-coded by semantic group. Upper screen: hovering the cursor over an annotation delivers the corresponding CUI, preferred term, semantic type, and semantic group (DISO is "Disorders"). Lower screen: double-clicking a text opens a pop-up window to edit the annotation or to link out to other resources, such as the UMLS Technology Services.

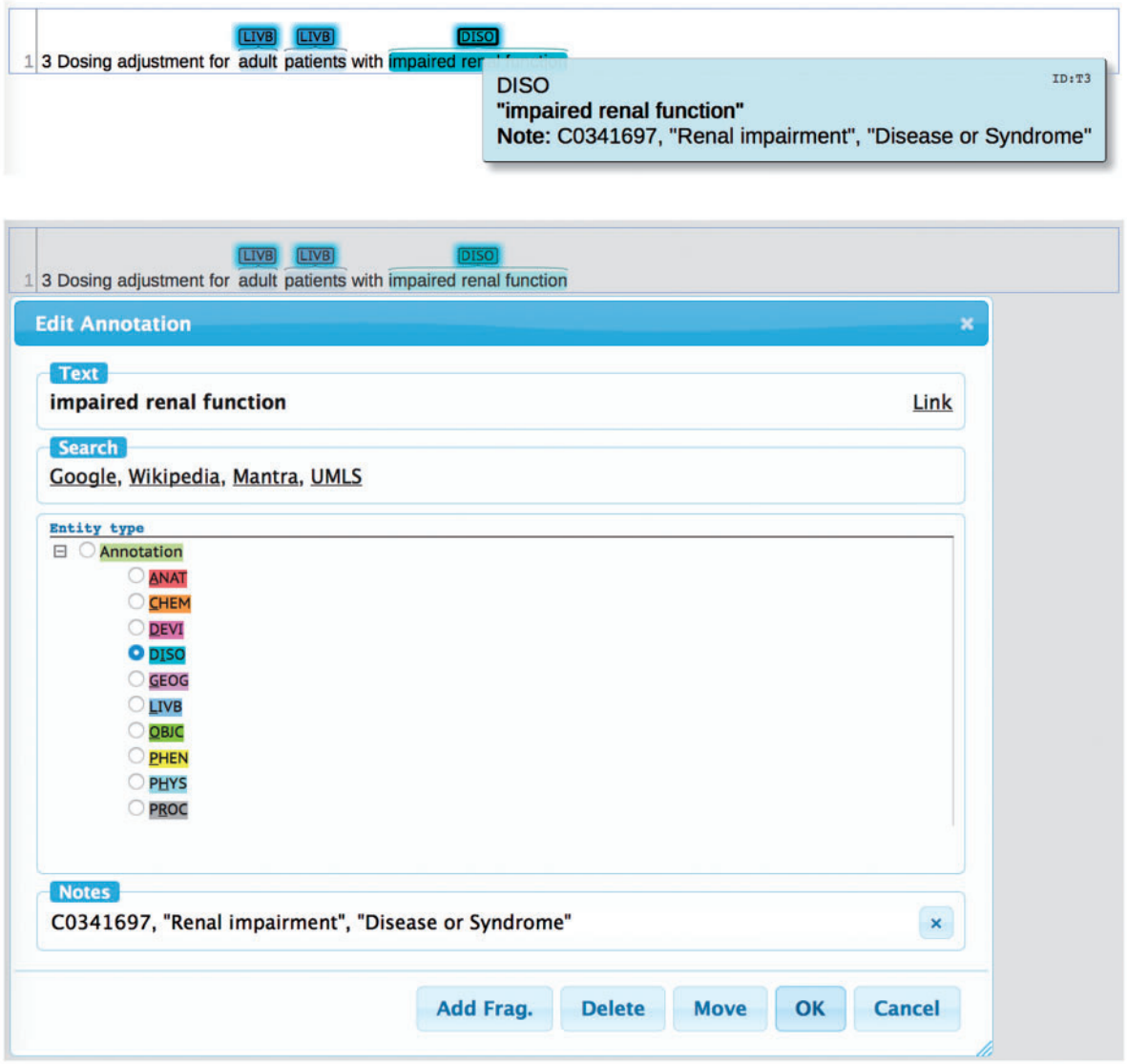
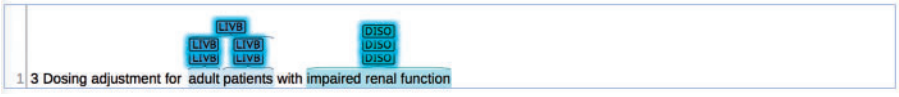


Figure 4: English unit from the Mantra GSC with the aggregated annotations of three independent annotators. LIVB denotes semantic group "Living beings," DISO "Disorders."



observed between the individual annotators and the gold standard for all languages (median *F*-score 0.80). The low performance of the third German language annotator (*F*-score 0.66) is a combination of low recall (0.72) and low precision (0.61). The low precision is partly due to the annotator's noncompliance with the Mantra terminology, mainly because he annotated concepts that belong to semantic groups that were excluded from the Mantra terminology (e.g., the semantic group "Concepts and ideas").

Figure 5: Corresponding German and English units from the Mantra GSC. The German unit has one German preannotation, the English unit shows the CUIs of the curated English harmonized annotations. LIVB denotes semantic group "Living beings."

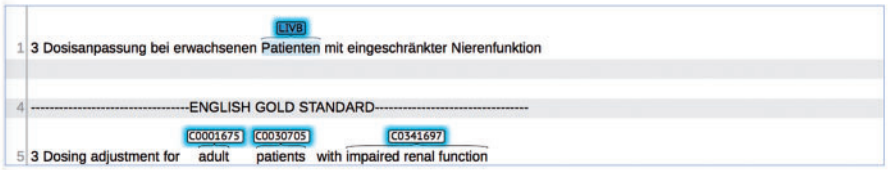
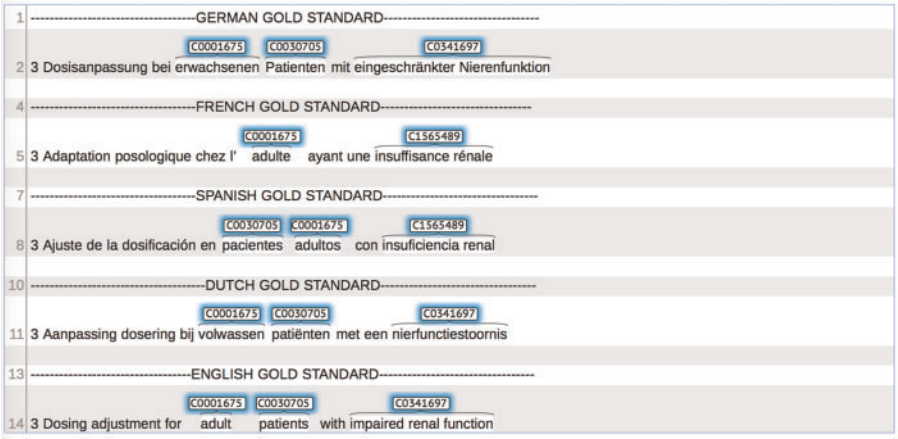


Figure 6: Corresponding units from the Mantra GSC in different languages, with curated harmonized annotations. Note that the concept for "patients" (C0030705) could not be annotated in French.



The agreement between the automatically constructed harmonized set and the gold standard is consistently high across all languages (*F*-scores from 0.83 to 0.86). These scores are comparable with those of the best annotators against the gold standard, and substantially better than some of the other annotators (e.g., for French, German, and Spanish). By contrast, the preannotation sets show moderate to low agreement scores against the gold standard. Precision of the preannotations varied between 0.60 and 0.73 across languages, but recall was considerably lower for French (0.47), German (0.38), and Dutch (0.33). This can be explained by the limitations in the terminological resources for these languages leading to a restricted number of preannotations. Furthermore, the annotation solutions for these languages may fall behind the ones for English, and compound words, which are common in German and Dutch, may add further complexity by occasionally requiring two concepts, for example, "Eisenstoffwechsel" ("iron metabolism") is not a concept in the Mantra terminology, and both "Eisen" and "stoffwechsel" have to be annotated. The recall for Spanish (0.61) is higher, probably due to SNOMED-CT being available in Spanish in contrast to the other non-English languages. The recall for English was 0.83, which may be attributed to the wide range of English synonyms in the Mantra terminology.

Table 3 shows the final numbers of total and unique concepts in the GSC and the percentage of annotated terms consisting of one word, two words, or three or more, for each of the five languages. The number of annotations correlates with the number of units in the different languages (cf. Table 1): 550 for English, 250 for French and German, and 200 for Spanish and Dutch. The differences between French and German, and between Spanish and Dutch largely stem from the differences in Medline titles between these languages (since the Medline titles were bilingual). The distribution of the number of words of the annotated terms shows that German and Dutch annotations tend to have a smaller number of words than the other languages, reflecting the fact that word compounding more frequently occurs in German and Dutch.

Some terms had to be annotated with more than one concept because the UMLS provided insufficient information to distinguish the concepts. For example, the difference between concepts C0038351 ("stomach") and C1278920 ("entire stomach") was unclear, and both concepts were annotated when the term "stomach" occurred in text. The average number of concepts per term was highest for the drug labels (1.19) and lowest for the patents (1.09); the bilingual Medline titles had intermediate values. The averages hardly differed (at most 0.01) between the languages in each subcorpus.

Table 4 shows the distribution of the annotations in terms of semantic groups for each text type in the Mantra GSC. The figures for Medline titles and drug labels are rather similar, although not surprisingly drug labels contain relatively more annotations belonging to the group "Chemicals and drugs" and less to "Procedures." Patents show a still larger representation of "Chemicals and drugs," and relatively few "Procedures" and "Living beings."

Table 3: Number of total and unique annotations (CUIs) and the percentage of annotated terms with a given number of words in the Mantra GSC for the different languages.

Language	No. of annotations		Word length of annotated terms (%)		
	Total	Unique	1	2	≥3
English	1963	1301	63.8	27.2	9.0
French	1052	710	66.8	19.8	13.4
German	1082	729	82.2	13.1	4.7
Spanish	756	550	63.4	18.0	18.6
Dutch	677	490	77.7	15.2	7.1

In a truly parallel corpus, the CUIs that are annotated in one language should be the same as those in the other language. To assess how parallel the text units in the Mantra GSC are, we determined the agreement between the final annotations for all available pairs of languages per text types. The agreement scores on the patents were high (*F*-scores 0.95 or 0.96); for Medline titles, the average agreement was 0.95, the lowest for English/Spanish (0.93) and the highest for English/German (0.96); and for drug labels, the agreement scores ranged from 0.88 (German/Spanish) to 0.94 (English/French), with an average agreement of 0.91. Overall, the patents and Medline titles are considered highly parallel.

DISCUSSION

The creation of a GSC is an extensive task, especially if the GSC covers different languages, the annotations have to be furnished at the concept level, and the concepts belong to a broad range of semantic types. Our approach to create a multilingual, wide-scoped GSC for biomedical concept recognition reduced the manual curation effort and increased the annotation quality through several means. First, we primed the annotators with automatically generated preannotations thus minimizing the time-consuming identification of appropriate CUIs for relevant terms. Although it remains open whether the preannotations biased the annotation results, the low agreement scores between preannotations and final annotations demonstrate that the annotators made substantial changes, thus following their own judgment.

Second, we automatically harmonized the individual annotations, which showed similar performance against the final annotations as the best annotator for each language and outperformed the other annotators. Harmonizing the individual annotations reduces the curation effort required for reaching the final annotation set, and also served as high-quality input (i.e., as English preannotations) for the annotation of the non-English units, ameliorating the low recall of non-English preannotations. Furthermore, the harmonization of multiple annotations appears to be a suitable approach for obtaining high-quality annotations if the performance of the individual annotators is unknown.

Finally, the use of parallel text units that should contain the same concept in different languages, greatly facilitated the annotation process, although – in practice – small variations in language use could slightly modify the meaning. For instance, the "attending physician" (C1320929) mentioned in an English unit about a possible ophthalmologic adverse drug reaction, was referred to as "médico" ("physician", C0031831) in the corresponding Spanish unit, and as "oogarts" ("ophthalmologist", C1704292) in the Dutch unit. The availability of parallel corpora also allowed consistency checks across the different languages for the given annotations leading to higher annotation quality.

Table 4: Distribution (%) of the total number of annotations (CUIs) per text type in the Mantra GSC over the different semantic groups.

Semantic group	Medline titles ($n = 2332$)	Drug labels ($n = 2155$)	Patents ($n = 1043$)
Anatomy	13.3	7.4	8.9
Chemicals and drugs	9.1	23.5	43.0
Devices	1.2	0.7	1.2
Disorders	30.6	35.4	29.5
Geographic areas	1.4	0.0	0.0
Living beings	13.3	11.0	3.2
Objects	1.8	1.9	4.0
Phenomena	2.0	1.5	0.9
Physiology	6.0	4.6	3.2
Procedures	21.4	14.0	6.2

Three annotators per language instead of a single annotator increase the total curation effort and the complexity of the approach, but also induces several benefits. The variability amongst the annotators in combination with the harmonized annotation set is a key element in obtaining high-quality annotations. Furthermore, different annotators allow the computation of inter-annotator agreement scores as well as agreements between annotators and the final gold standard. Such agreement scores provide important reference standards for gauging the performance of automatic annotation solutions or silver standard approaches.

The annotators showed very few discrepancies in their annotation of concept boundaries. Most of these were due to a definite or indefinite article (or in French a partitive article) being included in the annotated term by one annotator and not by the others. Occasionally, an annotator missed marking the initial or last character of the term to be annotated. Discrepancies were more frequently seen in French, possibly because a definite particle in French contracts with a term that starts with a vowel or mute h. The far majority of disagreements between annotators stem from differences in the annotated CUIs. For the English units, which had many preannotations, disagreements mainly resulted from ambiguous preannotations where annotators disagreed on the concepts to remove, or from unambiguous but incorrect preannotations that an annotator forgot to delete. For the non-English units, which had fewer preannotations, disagreements also occurred if the annotators added annotations, in particular if the term in the non-English unit had a slightly different meaning than in the corresponding English unit.

There are many linguistic differences between the languages covered in this study. For example, while all five languages generally follow the basic word order of subject-verb-object in main clauses, for the non-English languages word order may change in subordinate clauses or to emphasize words. French and Spanish are much more inflected than the other languages, especially in verb conjugations, whereas German has four cases for the declination of nouns and depending adjectives and articles. Languages also differ in word compounding, which is more common in German and Dutch than in the other languages. In our experience these linguistic differences did not

affect the annotation process for the Mantra GSC, where the annotated terms mainly consist of nouns and adjectives. For more complicated annotation schemes, for example, involving relationships, or a more diverse set of languages, linguistic variation may have a larger impact.

Our study has several limitations. First, the GSC is still of rather limited size. A possible future extension should profit from the acquired expertise and the infrastructure that has been developed (guidelines, annotation tools). Second, our GSC covers different document types, but not electronic health records, an important data source for text-mining applications.[1] Privacy issues complicate public availability of such data. Moreover, our annotation approach exploits parallelism, but parallel corpora of electronic health records do not exist and are unlikely to become available. A third limitation is the incompleteness and ambiguity of the Mantra terminology. Although we based our terminology on a large subset of the UMLS, sometimes a concept that is present in the UMLS could not be annotated because it was not contained in our selection of vocabularies and semantic groups. For instance, "dental" (C0226984) was not annotated because it belongs to the semantic group "Concepts and ideas," which we excluded from the Mantra terminology as this group also contains many general and unspecific terms. In some cases, terms had to be annotated with more than one concept because the UMLS provided insufficient information to distinguish the concepts. For example, the difference between concepts C0038351 ("stomach") and C1278920 ("entire stomach") was unclear, and both concepts were annotated when the term "stomach" occurred in the text.

Our approach to the creation of a multilingual annotated corpus can be applied to more languages than we have covered in this study. Parallel corpora for additional languages are readily available. The drug labels in the EMEA corpus are available in 22 European languages, bilingual Medline titles can be obtained for many European and non-European languages, and bilingual patent claims can be retrieved, for example, for Japanese or Chinese.

CONCLUSION

To our knowledge, the Mantra GSC is the first gold-standard corpus for biomedical concept recognition in languages other than English. Other distinguishing features of the Mantra GSC are the wide variety of semantic groups that are being covered, and the diversity of text genres that were annotated.

CONTRIBUTORS

J.K., S.C., E.V.M., and D.R.S. designed the study. J.K., S.C., and D.R.S. developed the annotation guidelines. S.C. determined the preannotations and computed the corpus statistics. S.A. set up and configured the annotation tool, determined the harmonized annotations, and calculated the agreement scores. J.K., S.C., E.V.M., and D.R.S. provided concept annotations and J.K. and D.R.S. resolved annotation discrepancies. J.K. and D.R.S. drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Mantra project (STREP project grant 296410) under the EU's 7th Framework Programme within theme "Technologies for Digital Content and Languages" (FP7 ICT-2011.4.1).

COMPETING INTEREST

None.

AVAILABILITY

The Mantra GSC can be viewed online and downloaded in brat format, at http://biosemantics.org/mantra/. It is also available in XML format and can be downloaded from https://files.ifi.uzh.ch/cl/mantra/gsc/GSC-v1.1.zip.

ACKNOWLEDGEMENTS

We greatly thank our annotators: Leonardo Campillos, Ronald Cornet, Antonio Jimeno-Yepes, Luise Modersohn, Moritz Mohr, Antonio Moreno Sandoval, Christel Olivares, Daniël Westerbeek.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Ohno-Machado L. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. *J Am Med Inform Assoc.* 2014;21:193.
2. Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform.* 2004;37:512–526.
3. Rebholz-Schuhmann D, Jimeno Yepes AJ, Van Mulligen EM, *et al.* CALBC silver standard corpus. *J Bioinform Comput Biol.* 2010;8:163–179.
4. Rebholz-Schuhmann D, Jimeno Yepes A, Li C, *et al.* Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *J Biomed Semantics.* 2011;2 (Suppl 5):S11.
5. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32:D267–D270.
6. Rebholz-Schuhmann D, Clematide S, Rinaldi F, *et al.* Entity recognition in parallel multi-lingual biomedical corpora: the CLEF-ER laboratory overview. In: Forner P, Müller H, Paredes R, *et al.*, eds. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization.* Berlin Heidelberg: Springer; 2013:353–367.
7. Mantra project website. http://www.mantra-project.eu Accessed April 17, 2015.
8. Leaman R, Miller C, Gonzalez G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine (LBM); Jeju Island, South Korea*, 2009: 82–89.
9. Gurulingappa H, Rajput AM, Roberts A, *et al.* Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform.* 2012;45:885–892.
10. Bada M, Eckert M, Evans D, *et al.* Concept annotation in the CRAFT corpus. *BMC Bioinformatics.* 2012;13:161.
11. Pradhan S, Elhadad N, South BR, *et al.* Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc.* 2015;22:143–154.
12. Hirschman L, Colosimo M, Morgan A, *et al.* Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics.* 2005;6 (Suppl 1):S11.
13. Morgan AA, Lu Z, Wang X, *et al.* Overview of BioCreative II gene normalization. *Genome Biol.* 2008;9 (Suppl 2):S3.
14. Lu Z, Kao HY, Wei CH, *et al.* The gene normalization task in BioCreative III. *BMC Bioinformatics.* 2011;12 (Suppl 8):S2.
15. Volk M, Ripplinger B, Vintar S, *et al.* Semantic annotation for concept-based cross-language medical information retrieval. *Int J Med Inform.* 2002;67:97–112.
16. Open Source Parallel Corpus (OPUS), European Medicines Agency documents. http://opus.lingfil.uu.se/EMEA.php Accessed April 17, 2015.
17. IFI CLAIMS patent database. http://www.ificlaims.com/index.php?page=-products_claims_databases2 Accessed April 17, 2015.
18. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform.* 2003;36:414–432.
19. Stenetorp P, Pyysalo S, Topić G, *et al.* brat: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations Session at EACL 2012; Association for Computational Linguistics; 2012: 103-107.
20. Schuemie MJ, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. *Proceedings of the BioCreAtIvE II Workshop; Madrid, Spain*; 2007: 131–133.
21. Hahn U, Buyko E, Landefeld R, *et al.* An overview of JCoRe, the JULIE lab UIMA component repository. *Proceedings of the Language Resources and Evaluation Conference (LREC); Marrakech, Morocco*; 2008: 1–7.

22. Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al*. Text processing through Web services: calling Whatizit. *Bioinformatics*. 2008;24:296–298.
23. Averbis Extraction Platform. http://www.averbis.de/en/technologies/text_analytics Accessed April 17, 2015.
24. Linguamatics I2E text mining software. http://www.linguamatics.com/welcome/software/I2E.html Accessed April 17, 2015.
25. Rebholz-Schuhmann D, Clematide S, Rinaldi F, *et al*. Multilingual semantic resources and parallel corpora in the biomedical domain: the CLEF-ER challenge. Conference and Labs of the Evaluation Forum (CLEF) 2013. CLEF-ER working notes. http://www.clef-initiative.eu/edition/clef2013/working-notes Accessed April 17, 2015.
26. Lewin I, Kafkas S, Rebholz-Schuhmann D. Centroids: gold standards with distributional variation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012); European Language Resources Association*; 2012: 3894–3900.
27. Lewin I, Clematide S. Deriving an English biomedical silver standard corpus for CLEF-ER. Conference and Labs of the Evaluation Forum (CLEF) 2013. CLEF-ER working notes. http://www.clef-initiative.eu/edition/clef2013/working-notes Accessed April 17, 2015.
28. UMLS Terminology Services. https://uts.nlm.nih.gov/home.html Accessed April 17, 2015.

AUTHOR AFFILIATIONS

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland