Quantitative analysis of manual annotation of clinical text samples

Miñarro-Giménez, Jose A; Cornet, Ronald; Jaulent, M C; Dewenter, Heike; Thun, Sylvia; Gøeg, Kirstine Rosenbeck; Karlsson, Daniel; Schulz, Stefan

Published in:
International Journal of Medical Informatics

DOI (link to publication from Publisher):
10.1016/j.ijmedinf.2018.12.011

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

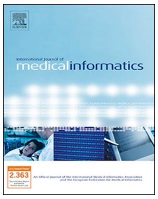
Citation for published version (APA):
Miñarro-Giménez, J. A., Cornet, R., Jaulent, M. C., Dewenter, H., Thun, S., Gøeg, K. R., Karlsson, D., & Schulz, S. (2019). Quantitative analysis of manual annotation of clinical text samples. *International Journal of Medical Informatics*, *123*, 37-48. https://doi.org/10.1016/j.ijmedinf.2018.12.011

International Journal of Medical Informatics 123 (2019) 37–48

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Quantitative analysis of manual annotation of clinical text samples

Jose A. Miñarro-Giménez[a,*], Ronald Cornet[b], M.C. Jaulent[c], Heike Dewenter[d], Sylvia Thun[e], Kirstine Rosenbeck Gøeg[f], Daniel Karlsson[g], Stefan Schulz[a]

[a] *Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *Department of Medical Informatics, Amsterdam Public Health Research Institute, The Netherlands*
[c] *Laboratoire d'Informatique Médicale et Ingénierie des Connaissances en eSanté, Institut National de la Santé et de la Recherche Médicale, Sorbonne Université, Université Paris 13, France*
[d] *University of Applied Sciences Niederrhein, Krefeld, Germany*
[e] *Charité Universitätsmedizin, Berlin Institute of Health, Germany*
[f] *Department of Health Science and Technology, Aalborg University, Denmark*
[g] *Department for Knowledge-Based Policy of Social Services, eHealth and Structured Information Unit, The National Board of Health and Welfare, Sweden*

ARTICLE INFO

Keywords:
Quantitative analysis
Medical terminology
SNOMED CT
Annotation experiment

ABSTRACT

Background: Semantic interoperability of eHealth services within and across countries has been the main topic in several research projects. It is a key consideration for the European Commission to overcome the complexity of making different health information systems work together. This paper describes a study within the EU-funded project ASSESS CT, which focuses on assessing the potential of SNOMED CT as core reference terminology for semantic interoperability at European level.
Objective: This paper presents a quantitative analysis of the results obtained in ASSESS CT to determine the fitness of SNOMED CT for semantic interoperability.
Methods: The quantitative analysis consists of concept coverage, term coverage and inter-annotator agreement analysis of the annotation experiments related to six European languages (English, Swedish, French, Dutch, German and Finnish) and three scenarios: (i) ADOPT, where only SNOMED CT was used by the annotators; (ii) ALTERNATIVE, where a fixed set of terminologies from UMLS, excluding SNOMED CT, was used; and (iii) ABSTAIN, where any terminologies available in the current national infrastructure of the annotators' country were used. For each language and each scenario, we configured the different terminology settings of the annotation experiments.
Results: There was a positive correlation between the number of concepts in each terminology setting and their concept and term coverage values. Inter-annotator agreement is low, irrespective of the terminology setting.
Conclusions: No significant differences were found between the analyses for the three scenarios, but availability of SNOMED CT for the assessed language is associated with increased concept coverage. Terminology setting size and concept and term coverage correlate positively up to a limit where more concepts do not significantly impact the coverage values. The results did not confirm the hypothesis of an inverse correlation between concept coverage and IAA due to a lower amount of choices available. The overall low IAA results pose a challenge for interoperability and indicate the need for further research to assess whether consistent terminology implementation is possible across Europe, e.g., improving term coverage by adding localized versions of the selected terminologies, analysing causes of low inter-annotator agreement, and improving tooling and guidance for annotators. The much lower term coverage for the Swedish version of SNOMED CT compared to English together with the similarly high concept coverage obtained with English and Swedish SNOMED CT reflects its relevance as a hub to connect user interface terminologies and serving a variety of user needs.

1. Introduction

Terminology systems provide standardized meaning of terms within a given domain. In medicine, many efforts have been undertaken into their development [1–3]. There are different ways to distinguish types of terminology systems, among which is the categorization of interface, reference, and aggregation terminologies [4,5]. Interface terminologies provide close-to-user descriptions of concepts, including colloquialisms,

[*] Corresponding author.
E-mail address: jose.minarro-gimenez@medunigraz.at (J.A. Miñarro-Giménez).

https://doi.org/10.1016/j.ijmedinf.2018.12.011
Received 5 June 2018; Received in revised form 9 October 2018; Accepted 30 December 2018

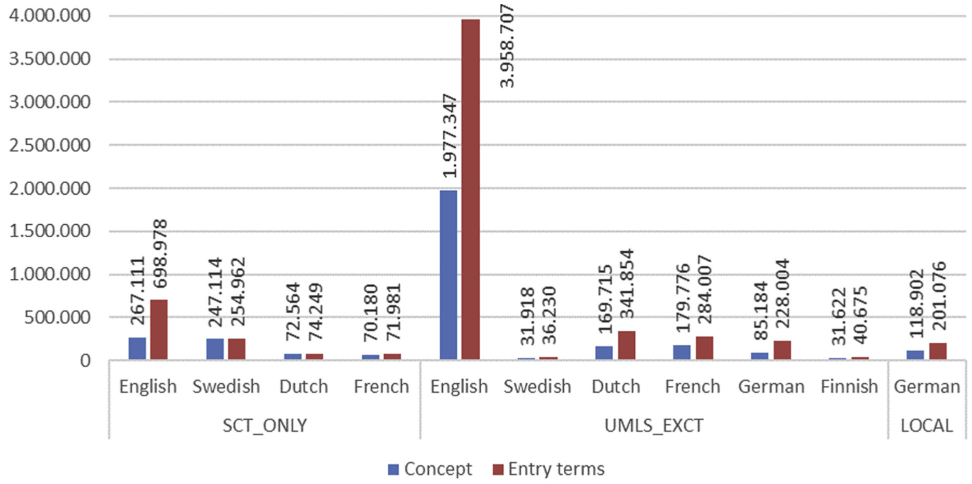


Fig. 1. Number of concepts and entry terms for each language in SCT_ONLY, UMLS_EXCT and LOCAL terminology settings.

various languages, and custom abbreviations. Interface terminologies can be linked to reference terminologies (e.g., SNOMED CT), which may provide logic-based definitions of concepts, enabling search and aggregation based on hierarchical ordering and properties. Aggregation terminologies are classifications (e.g. ICD-10), which provide classification rules to represent information as non-overlapping classes, used for statistics-oriented data processing.

Adequate capture of clinical information using (interface) terminologies depends on the terminology content, terminology representation, and on the user interface [6]. This means that, given a user interface, comparative analysis can be performed to assess the adequacy of various terminologies, or combinations thereof, for capturing clinical information. This can be measured by the indicator content coverage, which has been applied in numerous studies, among others for SNOMED CT [7].

SNOMED CT is distributed in English and Spanish, with other translations being provided by member countries, such as Swedish and Danish (completed), or French and Dutch (in development by the Dutch and Belgian Terminology Center at the time of the annotation experiment). However, other EU languages are not included, such as German, Italian, and Portuguese, and various EU member states, e.g., Germany, France, Finland, and Austria, have not, or not yet, joined SNOMED International.

To provide health terminology strategy recommendations to the European Commission whether SNOMED CT can play the role as core terminology in the EU, the ASSESS CT (Assessing SNOMED CT for Large-Scale eHealth Deployments in the EU) project [8] has set up several studies to deliver supporting evidence. This paper describes the quantitative analysis of the ASSESS CT study that determines the relative fitness for purpose of SNOMED CT for clinical documentation, focusing on the semantic annotation of clinical narratives. The ASSESS CT study is available in [9].

2. Materials and methods

The quantitative analysis is based on the calculation of the indicators for terminology coverage and quality measurement:

- Concept coverage: the degree of successful representation of the content in samples of clinical content.
- Term coverage: it measures the degree by which the language used in the source to represent that content shows a match with the terms used in the terminology setting under scrutiny.
- Inter-annotator agreement (IAA). Indicator of how easy it would be to code equivalent meanings from clinical narratives consistently across multiple coding scenarios.

2.1. Parallel clinical corpus

We selected the languages English, Swedish, French, Dutch, German and Finnish to provide a diverse set of languages with full, partial or lacking translation of SNOMED CT and their use in one or more than one European country.

The clinical corpus was acquired, consisting of 60 short samples (400–600 characters) of de-identified clinical texts, 10 from each selected language, which were translated into the other languages. Thus, we obtained a parallel corpus of 60 samples for each language. The corpora were balanced for their text types, document sections and clinical topics. Appendix A describes the acquisition and translation process of the samples.

2.2. Terminology settings

ASSESS CT addresses the three different annotation scenarios, i.e. ADOPT, ALTERNATIVE and ABSTAIN, by setting up three terminology settings respectively, SCT_ONLY, UMLS_EXCT and LOCAL. The differentiation between scenarios and settings stresses the limitations imposed by the terminologies to make them comparable in our experiment. These restrictions are related to the availability of the terminology content in other languages (see Appendix B):

- SCT_ONLY used exclusively SNOMED CT, with English, Swedish, French and Dutch descriptions respectively.
- UMLS_EXCT was a subset of terminologies from the 2015 UMLS release, excluding SNOMED CT.
- The LOCAL setting was configured to be used only for German. It contained the terminologies ICD10, LOINC, ATC, MeSH, ICD-O, the German Procedure Classification (OPS) and ABDAMED, a German drug catalog.

The terminology settings were designed to focus on typical medical concepts. To that end, we selected the UMLS semantic groups *Anatomy*, *Chemicals & Drugs*, *Concepts & Ideas*, *Devices*, *Disorders*, *Genes & Molecular Sequences*, *Living Beings*, *Objects* and *Procedures*. Fig. 1 shows the number of concepts and entry terms available for each terminology setting.

2.3. Annotation guidelines

We developed guidelines that support the annotation decisions so that annotations optimally represent the meaning of medical narratives for the three terminology settings. The guidelines were specifically adapted to ASSESS CT and did not depend on any real clinical coding

Table 1
Short glossary for the terms used in ASSESS CT with examples using SNOMED CT terminology.

Term	Definition	Example
Concept	Unit of specific meaning in a terminology	Normocytic anemia (disorder)
Code	Alphanumeric identifier for a concept	300980002
Token	Single word, numeric expression, or punctuation sign	Modest
Chunk	Single token or phrase delineated by the annotator to correspond to a clinical concept	Modest normocytic anaemia
Annotation group	Set of concept codes that jointly represent or approximate the meaning of the clinical concept related to a chunk	300980002 |Normocytic anemia (disorder), 255604002 |Mild (qualifier value)|

context. Table 1 provides a short glossary of terms defined in the annotation guidelines.

Excel spreadsheets were provided to annotators with a pre-filled column with one token per row from the text samples and followed by a column for chunk identifiers. Then, for each terminology setting, three columns were to be filled in: (i) set of annotation codes; (ii) concept coverage; and (iii) term coverage.

The annotation task comprised the following steps: (i) delimit and identify the chunks; (ii) find the smallest set of codes that best represents the meaning of each chunk; and (iii) provide the concept and term coverage scores. For all terminology settings, terminologies were uniformly displayed in the Averbis Terminology Platform (ATP) [10], a web-based service that supports navigation and search within customizable sets of terminologies.

Annotators rated the concept coverage using an assessment scale with five scores that represent full, inferred, partial or none-coverage of the meaning of a chunk, and out of scope (see Table 2). This scale was based on ISO/TR 12300:2014 “Health informatics – Principles of mapping between terminological systems” [11].

The term coverage is a binary Yes/No value that indicates whether the entry terms in a terminology approximately match the tokens in the text samples. Appendix C provides more details about the rules that restrict the annotation process and an annotation example. The complete annotation guidelines produced in ASSESS CT are available in [12].

2.4. Annotation experiment

Annotators were recruited to have similar medical domain knowledge and trained in our defined annotation guidelines to deliver comparable annotations. The spreadsheet provided to the annotators contains 20–60 text samples depending on the available effort of annotators. Texts were distributed among annotators in a way that a subset of 20 samples (the same samples for all languages) was annotated twice for computing the IAA. The resulting annotations were post-processed to avoid trivial annotation inconsistencies due to negligence of annotators, i.e. when an annotator provides a concept coverage score without providing any concept code or using a concept code from one terminology setting in the wrong column in the annotation spreadsheet. Appendix D describes the followed criteria for annotator recruitment and post-processing tasks.

2.5. Analysis of results

We developed a Java application to calculate the concept and term coverage, and to generate the input files that were required for the calculation of IAAs in R. The R scripts use agreestat functions [13] to calculate the IAA with Krippendorff’s alpha measure [14]. A complete description with examples of the calculation of these indicators is in Appendix E. All material and software produced are available in a GitHub repository [15].

The concept coverage is calculated as the percentage of codes within an annotation group. Two types of concept coverage measurements are calculated, viz. *Strict* coverage which only considers codes annotated with *Full coverage* and *Inferred coverage* scores, opposed to *Loose* coverage which considers, in addition, *Partial coverage* scores.

Term coverage is calculated as the percentage of tokens covered with the interface terms from the corresponding terminology setting.

IAA is evaluated with the weighted version of Krippendorff’s alpha. Krippendorff’s alpha is commonly used in content analysis [16–18] for measuring the agreement between coders. The alpha coefficient was selected in our experiments for its reliability, which considers the observed and the expected disagreement [19]. We calculated the IAA with the 20 samples annotated by every annotator in the two modes: *IAA Strict,* which considers the agreements in the codes and concept coverage score; and *IAA Loose,* which only considers the codes. The weight is calculated using the quadratic weight (see Eq. (1)) that ranges from 1 (full agreement) to 0 (total disagreement).

$$Quadratic\ W_{ij} = 1 - \frac{n_{ij}^2}{(k_{ij} - 1)^2} \tag{1}$$

Eq. (1) calculates the weight between the annotation units i and j where n represents the number of common unique annotations and k represents the total number of unique annotations between the two units.

3. Results

An overview of the annotation experiment is provided in Table 3. In summary, two annotators each were recruited for English, Swedish, Dutch, German and Finnish, and three for French due to availability of the annotators. The corpus consists of 60 clinical texts where a set of 20

Table 2
Definition of concept coverage scores for ASSESS CT manual annotation experiment.

Score	Definition
Full coverage	The meaning of a chunk is fully represented by a set of codes, e.g. the term “Heart attack” is fully covered by the SNOMED CT concept *Myocardial infarction (disorder)*.
Inferred coverage	The meaning of elliptic or ambiguous chunks of text can be inferred from the context and can be fully represented by a set of codes, e.g. a specific use of the term “hypertension” could mean “Renal arterial hypertension”, so the SNOMED CT concept *Renal arterial hypertension (disorder)* is justified.
Partial coverage	The meaning of the chunk comes close to the meaning of a set of codes, e.g. “Third rib fracture” is more specific than the SNOMED CT concept *Fracture of one rib (disorder)*. Yet the meaning is close enough to justify annotation with this set of codes.
None	There is no set of codes that has a sufficiently close meaning to the chunk, e.g. generic codes such as the SNOMED CT code *Fracture of bone (disorder)* for “third rib fracture” must not be used for partial coverage.
Out of scope	The meaning of the text fragment is not covered by any of the semantic groups selected for this study or the meaning of the text fragment is not clear to the annotators.

Table 3
Overview of annotation results for English, Swedish, French, Dutch, German and Finnish. The table shows for each language the number of annotators recruited, the number of annotated text samples, and the average number of sentences, tokens and chunks per text sample.

Language	Annotators	Annotated Samples	Sentences	Tokens	Chunks
English	2	80 (40 each; 20 overlap)	9.0	97.1	12.5
Swedish	2	80 (40 each; 20 overlap)	8.7	86.0	13.0
Dutch	2	80 (40 each; 20 overlap)	9.0	100.0	13.8
French	3	100 (34 by one; 33 by the other two; 20 overlap)	8.7	112.0	12.1
Finnish	2	74 (20 by one; 54 by the other; 20 overlap)	9.4	75.9	22.2
German	2	80 (60 by one; 20 by the other; 20 overlap)	9.3	91.4	15.6

text samples were annotated by every annotator, but Finnish annotators who annotated only 74 (54 by one annotator and 20 by the other with 20 samples overlap) due to limited availability. Thus, the average values shown in Table 3 were calculated taking into consideration the total number of text samples annotated for each language.

3.1. Concept coverage

The results of the concept coverage are provided in Table 4. It contains the mean concept coverage for each terminology setting. *Loose* coverage consistently surpasses the *Strict* coverage.

Tables 5 and 6 show the distribution of the annotation codes by their semantic groups and terminologies, respectively. These numbers represent the total amount of unique codes used by all annotators for each language.

3.2. Term coverage

The results of term coverage are given in Table 7. Term coverage is calculated for English, Swedish, French, Dutch, German and Finnish annotations and for their corresponding terminology settings.

3.3. Inter-annotator agreement

The obtained alpha coefficients are provided in Table 8 for *Strict* and *Loose* modes, for the three terminology settings and their corresponding languages.

4. Discussion and conclusions

The Fig. 2 shows a positive correlation of 0.73 (with p-value < 0.05) between the number of available concepts for each terminology setting (see Fig. 1) and their concept coverage (see Table 4). However, there are noteworthy exceptions: English UMLS_EXCT includes seven times as many concepts as SCT_ONLY, but this has no significant impact on the concept coverage. This is also reflected in the similar number of unique codes used for annotating English samples (see Table 5) with SCT_ONLY and UMLS_EXCT. Therefore, there could be a limit in the number of frequently used concepts that appear in medical texts. The concept coverage of German UMLS_EXCT was much higher than German LOCAL (see Table 4) with a higher number of unique codes used (see Table 6), in spite of the latter being larger (see Fig. 1). This can be explained by the lack of codes from certain semantic groups like Devices, Procedures or Genes in the LOCAL setting.

A correlation coefficient of 0.83 (with p-value < 0.05) was obtained for term coverage (see Fig. 2). The much lower term coverage for Swedish in SCT_ONLY, for which only one term exists per concept, shows that it lacks many important interface terms. In English, the higher number of interface terms in UMLS_EXCT compared to SCT_ONLY had only a low impact in the same way as the number of concepts. The results for both German LOCAL and UMLS_EXCT settings show how good term coverage can be achieved with a good mixture of localized terminologies. This selection of terminologies would constitute a good benchmark to evaluate SNOMED CT in Germany, with the necessary interface terminologies.

Table 8 shows a generally low IAA for terminology settings. We may expect an inverse correlation between concept coverage and IAA due to availability of a lower amount of choices, but we obtained a correlation coefficient of 0.33 (with p-value > 0.05). Therefore, we cannot claim or refute such correlation between both variables. IAA results were very close among terminology settings and languages, with the exception of French, for which we calculated the agreement between three annotators, and Finnish, which could be explained by the reported poor-quality translations of the text samples. Regarding *Strict* and *Loose* IAA, the *IAA Loose coefficients* are consistently higher than the *IAA Strict coefficients. This may indicate a disagreement in the understanding of the meaning of the concepts in terminologies*. We hypothesize that this could be mitigated with textual definition of the terminology concepts.

There are several factors that may have impacted the results and need highlighting for proper interpretations of the study. First, we have not assessed the extent to which the sample of text fragments was representative. However, we have aimed at constructing a realistic sample, covering various settings, and including realistic errors.

Table 4
Concept coverage for the SCT_ONLY, UMLS_EXCT and LOCAL settings and for English, Swedish, French, Dutch, German and Finnish languages. It shows the average concept coverage and the confidence interval (CI) with 95% significance.

Language	Concept coverage SCT_ONLY				Concept coverage UMLS_EXCT			
	Strict coverage		Loose coverage		Strict coverage		Loose coverage	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
English	0.86	[0.82;0.88]	0.92	[0.88;0.93]	0.88	[0.86;0.91]	0.94	[0.93;0.96]
Swedish	0.87	[0.84;0.89]	0.91	[0.88;0.93]	0.59	[0.55;0.63]	0.65	[0.61;0.69]
Dutch	0.43	[0.35;0.44]	0.52	[0.45;0.55]	0.60	[0.57;0.65]	0.67	[0.64;0.72]
French	0.45	[0.37;0.47]	0.57	[0.49;0.59]	0.64	[0.61;0.70]	0.75	[0.73;0.80]
Finnish					0.36	[0.30;0.40]	0.64	[0.60;0.69]
Language	Concept coverage LOCAL				Concept coverage UMLS_EXCT			
	Strict coverage		Loose coverage		Strict coverage		Loose coverage	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
German	0.64	[0.60;0.68]	0.74	[0.72;0.78]	0.82	[0.80;0.85]	0.90	[0.89;0.93]

Table 5
Number of unique codes used for annotating the corpus for each language with SCT_ONLY (SCT) and UMLS_EXCT (UMLS) settings. The numbers are classified by the UMLS semantic groups: Genes and Molecular Sequences (GENE), Anatomy (ANAT), Concepts and Ideas (CONC), Objects (OBJC), Procedures (PROC), Disorders (DISO), Living Beings (LIVB), Chemicals and Drugs (CHEM), Devices (DEVI).

Language	English		Swedish		Dutch		French		Finnish	German
	SCT	UMLS	SCT	UMLS	SCT	UMLS	SCT	UMLS	UMLS	UMLS
GENE	0	1	1	1	0	1	1	0	0	0
ANAT	170	159	222	123	122	107	133	128	90	131
CONC	189	197	123	7	64	7	81	13	10	8
OBJC	3	3	1	0	2	1	1	0	1	1
PROC	198	218	167	75	35	124	33	133	83	129
DISO	399	373	301	189	174	282	166	261	224	286
LIVB	5	5	5	3	4	4	4	4	0	4
CHEM	102	95	122	127	44	76	32	100	110	123
DEVI	3	5	6	3	9	2	5	4	4	3
Total	1069	1056	948	528	454	604	456	643	522	685

Table 6
Number of unique codes used for annotation from each terminology in the German LOCAL setting.

Terminology	German LOCAL
ABDAMED	114
OPS	38
ICD-10	213
ICD-O	5
MeSH Anatomy	128
MeSH Organism	4
Total	502

Second, recruitment and training of annotators was constrained by availability of suitable experts, which led to some heterogeneity in terms of skills and experience. However, in practice, training of clinicians may also vary, leading to similarly different results. Third, the ATP is a generic tool that misses some of the benefits, such as post-coordination, leading to possible underestimation of content coverage. This is in line with common practice, where post-coordination is often disregarded and clinicians select items from prepopulated interface terminologies. Fourth, as the IAA did not consider the semantic distance among concepts, but instead scored match versus no match, the inter-annotator agreement may be underestimated.

The main conclusion of this study is that, in general, there are no significant differences in concept coverage, term coverage and inter-annotator agreement when comparing annotations from SCT_ONLY with UMLS_EXCT and LOCAL settings. The much lower term coverage for the Swedish version of SNOMED CT compared to English shows that more terms per concept are required. However, the similarly high concept coverage obtained with English and Swedish annotations using SNOMED CT reflects its relevance as core reference terminology. These facts support ASSESS CT's recommendation [20] that "SNOMED CT should play the role of a hub, which connects user interface terminologies and value sets of different provenances, in different languages and dialects, serving a variety of user needs". Further investigations are needed to assess the use of interface terminologies for the support of SNOMED CT implementations in practice.

The overall low IAA results pose a challenge for interoperability. Apart from technology, this needs to be addressed by more elaborated guidelines as well as by improving the structure and content of the terminology. To what extent the current ontological foundation of SNOMED CT can be exploited to infer equivalences or at least semantic proximity between diverging annotations is currently being investigated [21].

Author statement

Jose A. Miñarro-Giménez: The author participated in the design of the annotation experiments and evaluation process. He conducted the research tasks and took the lead for carrying out the research work and writing, reviewing and revising the manuscript.

Ronald Cornet: The author was part of the conceptualization of the research project, as well as, its design and implementation. He was a main participant in writing, reviewing and revising the manuscript.

MC Jaulent: The author was involved in the conceptualization of the research project. She also participated in the implementation of the research tasks and collaborated in the revision of the manuscript.

Heike Dewenter: The author participated in the design and implementation of the research tasks. She collaborated in the revision of the manuscript.

Sylvia Thun: The author was involved in the conceptualization of the research project. She also participated in the design of the research tasks and collaborated in the revision of the manuscript.

Kirstine Rosenbeck Gøeg: The author was part of the conceptualization, design and implementation of the research tasks. She collaborated in reviewing and revising the manuscript.

Daniel Karlsson: The author was part of the conceptualization, design and implementation of the research tasks. He collaborated in reviewing the manuscript.

Table 7
Term coverage for SCT_ONLY, UMLS_EXCT and LOCAL settings and for English, Swedish, French, Dutch, German and Finnish annotations. The confidence interval (CI) is provided with 95% significance.

Language	SCT_ONLY		UMLS_EXCT		LOCAL	
	Term coverage	95% CI	Term coverage	95% CI	Term coverage	95% CI
English	0.68	[0.64;0.70]	0.73	[0.69;0.76]		
Swedish	0.47	[0.44;0.52]	0.35	[0.32;0.40]		
Dutch	0.35	[0.29;0.36]	0.44	[0.41;0.49]		
French	0.39	[0.34;0.43]	0.57	[0.55;0.64]		
Finnish			0.23	[0.20;0.29]		
German			0.72	[0.71;0.79]	0.56	[0.53;0.62]

Table 8
It indicates the Strict and Loose Krippendorff's alpha and their 95% confidence interval (CI) for each terminology setting, SCT_ONLY, UMLS_EXCT and LOCAL, with English, Swedish, French, Dutch, German and Finnish.

Language	SCT_ONLY				UMLS_EXCT			
	Strict		Loose		Strict		Loose	
	Alpha	95% CI	Alpha	95% CI	Alpha	95% CI	Alpha	95% CI
English	0.37	[0.33;0.41]	0.64	[0.60;0.69]	0.36	[0.32;0.40]	0.64	[0.60;0.68]
Swedish	0.30	[0.26;0.34]	0.55	[0.51;0.60]	0.49	[0.43;0.54]	0.74	[0.70;0.78]
Dutch	0.30	[0.25;0.35]	0.55	[0.49;0.62]	0.45	[0.40;0.50]	0.70	[0.65;0.75]
French	0.22	[0.17;0.27]	0.40	[0.34;0.47]	0.36	[0.30;0.41]	0.57	[0.51;0.62]
Finnish					0.30	[0.26;0.35]	0.47	[0.42;0.51]

Language	LOCAL				UMLS_EXCT			
	Strict		Loose		Strict		Loose	
	Alpha	95% CI	alpha	95% CI	alpha	95% CI	alpha	95% CI
German	0.46	[0.41;0.56]	0.70	[0.65;0.74]	0.49	[0.44;0.54]	0.73	[0.69;0.77]

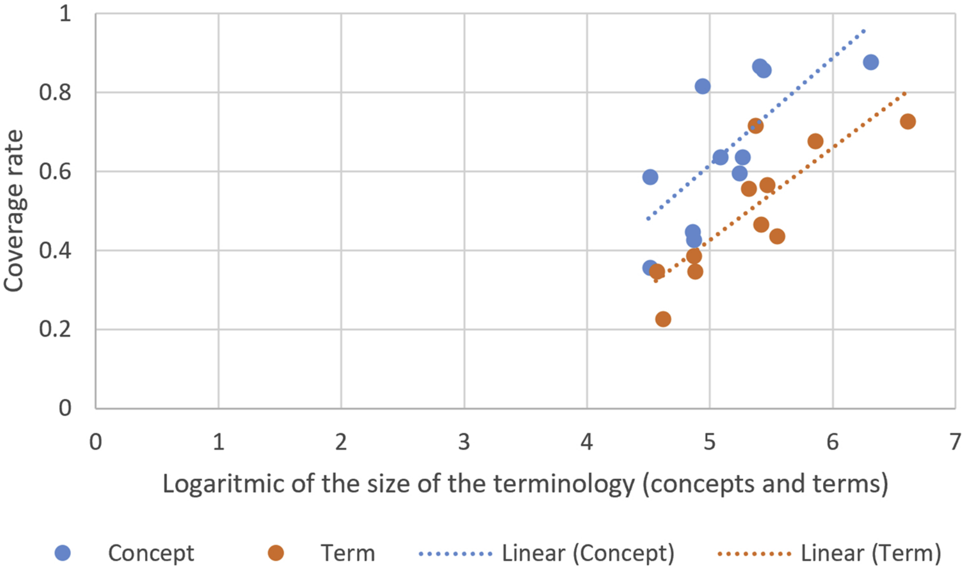


Fig. 2. It shows the correlation and linear regression of: (a) the concept coverage rate and the logarithmic number of concepts for each terminology setting and language; and (b) the term coverage rate and the logarithmic number of terms for each terminology setting and language.

Stefan Schulz: The author was the principal investigator of the research project. He was part of the conceptualization of the research project. He supported the first author in the design of the research tasks and participated in their implementation. Besides, he got involved in writing and reviewing the manuscript.

Conflict of interest

The authors declare they have no conflict of interests. It was funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643818.

Acknowledgments

This paper is produced as part of ASSESS CT which is funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643818. We thank our project colleagues: Anne Randorff Højen, Reza Fathollah Nejad, Vesna Kronstein Kufrin, Kornel Marko, Riikka Vuokko, Virpi Kalliokuusi, Benny Van Bruwaene.We thank the annotators Hester Albers, Erica Wilthagen, Chris Wroe, David Robinson, Xavier Aimé, Jean Charlet, Paul Maurice, Claudia Riede, Laura, Robert Liljenström, Sebastian Jarl, Hämäläinen Päivi and Mäkelä-Bengs Päivi.

Appendix A. Corpus acquisition

We summarized in Section 2.1 its main characteristics. In this section, we describe the guidelines to select the languages, collect and translate the parallel corpus.

The criteria for the selection of the languages of the corpus are the availability of a translation of SNOMED CT in that language and the number of countries that use the language in Europe. Table 9 shows the diversity of the languages in terms of the selection criteria. Such diversity is beneficial for the analysis of the fitness of terminology settings for semantic interoperability. We selected six European languages: English, Swedish, French, Dutch, German and Finnish. English and Swedish versions of SNOMED CT provide the full list of concepts, but Swedish version only translated one term per concept. French and Dutch were under development versions at the time of this study and provided by the Belgian and Dutch Terminology

Table 9
Criteria for the selection of languages for the construction of the corpus.

Criteria	English	Swedish	French	Dutch	German	Finnish
Full SNOMED CT translation	X	X				
Partial SNOMED CT translation			X	X		
No SNOMED CT translation					X	X
Official language only in one country		X		X		X
Official language in more than one country	X		X		X	

Table 10
Number of samples for each clinical domain in the corpora.

Clinical domains	Samples
Anaesthesiology	1
Dermatology	3
Gynaecology	2
Internal medicine	17
Neurology	3
Ophtalmology	1
Paediatrics	3
Pathology	12
Surgery	14
Urology	4

centres. There are no German or Finnish version available yet. Moreover, English, French and German are official languages in several European countries, i.e. English is official in UK, Ireland and Malta, French is official in France and Belgium, German is official in Germany, Austria and Switzerland; whereas Swedish, Dutch and Finnish are official in only one European country.

The parallel corpus contains for each of the six selected languages (English, Swedish, French, Dutch, German and Finnish) 60 short samples (400–600 characters) of clinical texts. Texts were provided by the partners of the ASSESS CT project and were completely de-identified. The acquisition of corpora was done in a way supposed to approximate representativeness in terms of clinical domains (see Table 10), document sections (see Table 11), and document types (see Table 12).

To construct the parallel corpora, all text samples were translated into the other languages by a professional translation service, following the predefined translation guidelines provided by the ASSESS CT consortium. The translation guidelines require (i) the translators to have some knowledge of medicine and have the target language as mother tongue, and (ii) the translation process preserves the meaning of the source text in a way it appears to a native speaker of the target language to have originally been written in that language. This includes maintaining the individual characteristics of the author, such as his/her command of the language. However, some problems had to be faced during translation and these rules were also defined: (a) drug trade names are replaced by their active ingredients, because the coverage of national drug names was considered out of scope for the terminology settings; (b) acronyms or abbreviations are translated into an equivalent acronym in the target language if possible; otherwise into a full term; (c) in case of spelling, grammar or style introduce similar errors are introduced into the translation. The raw translations were adjusted by the members of the ASSESS CT project due to obvious quality issues by the professional translation service. The corrections followed the same translation guidelines as the professional translation service.

Appendix B. Terminology setting restrictions

There are three annotation scenarios in ASSESS CT, namely ADOPT, ALTERNATIVE and ABSTAIN. Due to lack of availability of the selected medical terminologies in the six languages (English, Swedish, French, Dutch, German and Finnish) we defined the corresponding terminology setting

Table 11
Number of samples for each document section in the corpora.

Document sections	Samples
Conclusions	3
Diagnosis	2
Evolution	7
Findings	22
History	10
History & diagnosis	1
Imaging	1
Indication	1
Laboratory	4
Medication	2
Laboratory/medication	1
Order	1
Plan & finding	1
Recommendation	1
Summary	3

Table 12
Number of samples for each document type in the corpora.

Document types	Samples
Autopsy reports	3
Death certificate	1
Discharge summary	30
Microscopy report	2
Outpatient summary	1
Pathology report	3
Referral report	5
Finding report	4
Toxicology report	1
Visit report	10

SCT_ONLY, UMLS_EXCT and LOCAL.

SCT_ONLY consists of the English, Swedish, French and Dutch versions of SNOMED CT. The English international version (2015/07/31), the official Swedish translation and Dutch and French SNOMED CT work files of the National Release Center of Belgium were used, with the latter resource only covering parts of SNOMED CT. The German and Finnish translations of SNOMED CT are not available and such languages were omitted from SCT_ONLY.

UMLS_EXCT was a subset of the active release of the 2015 UMLS resources. This release has 128 active vocabularies, from which 71 were selected for our purposes, including the localized version of several terminologies: Anatomical Therapeutic Chemical (ATC) classification System, International Classification of Diseases (ICD-10), and Logical Observation Identifiers Names and Codes (LOINC). The criteria for selecting the vocabularies were the following: (i) exclude U.S. specific billing and reporting terminologies; (ii) include only terminologies currently used, excluding sources of uniquely scientific interest; (iii) exclude terminologies that are not currently maintained; (iv) exclude sources in languages other than English, Swedish, French, Dutch, German and Finnish; (v) exclude SNOMED CT and systems in its lineage, i.e. former SNOMED versions and READ codes; and (vi) exclude sources that represent disciplines that are considered out of scope, according to the decisions in ACCESS CT, viz. dentistry. The complete subset of selected terminologies is in Table 13.

The LOCAL setting was configured to be used for German texts. Table 14 shows the list of terminologies selected. All but the last two were localizations of international terminologies.

Appendix C. Annotation guidelines

The annotation guidelines describe the rules that guide the annotation decisions to obtain the best codes that represent the meaning of medical narratives from the corpus. The general annotation task consists of the following steps: (i) delimit and identify the chunks; (ii) find the smallest set of codes to better represent the meaning of each chunk; and (iii) provide the concept and term coverage scores to each token in a chunk.

To identify the chunks, annotators assign a unique chunk number for every row belonging to the same chunk. Relevant chunks are those that refer to medical concepts from the semantic types: findings, procedures, results of clinical and lab measurements (including related qualifiers), substances, organisms, medical devices and body structures. Chunk delimitation is user-specific, but it holds for all the terminology settings, according to the spreadsheet structure. After chunk delimitation, each setting requires three columns to be filled: (a) the set of codes; (b) concept coverage score; and (c) term coverage score. The idea is that each chunk is represented by an unordered set of codes. For technical reasons, three cases need to be further described:

- If a token is not part of any relevant chunk the corresponding cell content remains empty.
- If the meaning of the chunk is fully represented by one single code, then this specific code is added into every row of the chunk.
- If the representation of a single token requires more than one code, these codes are entered as list separated by comma.

The main purpose of the annotation process is the selection of the codes that more accurately cover the meaning of text fragments for each terminology setting. Thus, the annotators must select, first, the codes that cover the meaning of the chunk with *Full coverage* and *Inferred coverage* scores. Both scores have the same coverage level.

Whenever a chunk requires more than one code, the annotator should find a minimal list of codes that, together, better cover the meaning of the chunk. Partial overlap of tokens (e.g. "Sarcoma of rib" + "third rib") is allowed in case there is no better way to express its meaning. It is mandatory to use ATP with the corresponding language and terminology setting for retrieving the codes. Annotators should try several synonyms and substring matches before assigning a *Partial coverage* or *None score*. External services, such as Wikipedia and other Web resources can be used to find synonyms and discover the meaning of unknown terms. In case of doubt, the corresponding annotation code and coverage score cell must be left empty.

Regarding the term coverage, if a token exists literally or with minor variants (inflection, word order, typing error) and with the same meaning in an entry term of a terminology, then it is considered a full match and is annotated with *Yes*, in any other situation with *No*.

The general rules for the annotation process have some exceptions due to the scope of the selected UMLS semantic groups. Thus, annotators should not annotate the content related to:

- Proper names, professional roles, social groups, geographic entities, institutions, non-medical devices, non-medical events.
- Context information such as diagnostic certainty, plans, risks, clinical history or family history, e.g. in the phrase "high risk for lung cancer" only "lung cancer" should be annotated, as well as in "father died from lung cancer" or "suspected lung cancer"
- Temporal information, e.g. in the phrase "lung cancer, first diagnosed in Oct 2014" only "lung cancer" should be annotated. The only case where time-related information is annotated is in drug prescription such as "1-1-1" or "t.i.d.".
- Residuals, e.g. "Arterial hypertension NEC", "Tuberculosis, unspecified"," other complications of unspecified head injuries".

Table 13
List of terminologies that were selected from active release of the 2015 UMLS resources.

Terminology in UMLS_EXCT
AOT (Authorized Osteopathic Thesaurus)
ATC (Anatomical Therapeutic Chemical (ATC)-classification system)
CCC (Clinical Care Classification)
CHV (Consumer Health Vocabulary)
CPM (Medical Entities Dictionary)
CSP (CRISPy Thesaurus)
CVX (Vaccines Administered)
DMDICD10 (ICD-10 German)
DSM4 (DSM-IV)
FMA (Foundational Model of Anatomy)
HGNC (HUGO Gene Nomenclature Committee)
HL7V2.5 (HL7 Version 2.5)
HL7V3.0 (HL7 Version 3.0)
ICD10 (ICD-10)
ICD10AE (ICD-10 Am Engl)
ICD10CM (ICD-10-CM)
ICD10DUT (ICD10, Dutch Translation)
ICD10PCS (ICD-10-PCS)
ICD9CM (ICD-9-CM)
ICF (International Classification of Functioning, Disability and Health)
ICF-CY (International Classification of Functioning, Disability and Health for Children and Youth)
ICNP (International Classification for Nursing Practice)
ICPC2EDUT (ICPC2E Dutch)
ICPC2EENG (ICPC2E)
ICPC2ICD10DUT (ICPC2-ICD10ENG Thesaurus Dutch)
ICPC2ICD10ENG (ICPC2-ICD10 Thesaurus)
ICPC2P (ICPC-2 PLUS)
LCH_NW (Library of Congress Subject Headings, Northwestern University subset)
LNC (LOINC)
MDDB (Master Drug Data Base)
MDR (MedDRA)
MDRDUT (MedDRA Dutch)
MDRFRE (MedDRA French)
MDRGER (MedDRA German)
MEDCIN (MEDCIN)
MEDLINEPLUS (MedlinePlus)
MMSL (Multum)
MMX (Micromedex)
MSH (MeSH)
MSHDUT (MeSH Dutch)
MSHFIN (MeSH Finnish)
MSHFRE (MeSH French)
MSHGER (MeSH German)
MSHSWE (MeSH Swedish)
MTHHH (HCPCS Hierarchical Terms (UMLS))
MTHICD9 (ICD-9-CM Entry Terms)
MTHICPC2EAE (ICPC2E Am Engl)
MTHICPC2ICD10AE (ICPC2-ICD10 Thesaurus, Am Engl)
NCBI (NCBI Taxonomy)
NCI (NCI Thesaurus)
NCI_BRIDG (Biomedical Research Integrated Domain Group Model)
NCI_CDISC (Clinical Data Interchange Standards Consortium)
NCI_CRCH (Cancer Research Center of Hawaii Nutrition Terminology)
NCI_CTCAE (Common Terminology Criteria for Adverse Events)
NCI_CTEP-SDC (Cancer Therapy Evaluation Program - Simple Disease Classification)
NCI_DCP (NCI Division of Cancer Prevention Program)
NCI_DICOM (Digital Imaging Communications in Medicine)
NCI_DTP (NCI Developmental Therapeutics Program)
NCI_ICH (International Conference on Harmonization)
NCI_NCI-GLOSS (NCI Dictionary of Cancer Terms)
NCI_NCI-HL7 (NCI Health Level 7)
NCI_NCPDP (National Council for Prescription Drug Programs)
NCI_NICHD (National Institute of Child Health and Human Development)
NCI_PID (National Cancer Institute Nature Pathway Interaction Database)
NCI_RENI (Registry Nomenclature Information System)
NCI_UCUM (Unified Code for Units of Measure)
NDDF (FDB MedKnowledge)
NEU (Neuronames Brain Hierarchy)
PDQ (Physician Data Query)
PSY (Psychological Index Terms)
UWDA (Digital Anatomist)

Table 14
List of terminologies in the LOCAL setting.

Terminology in LOCAL
ICD10
LOINC
ATC
MeSH
ICD-O
OPS (German Procedure Classification)
ABDAMED (German drug catalog)

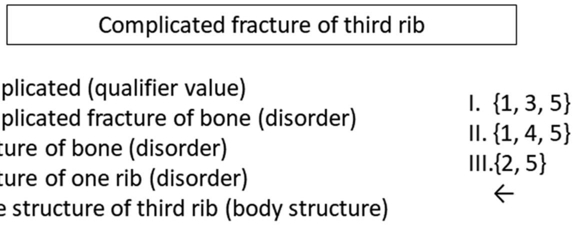

Complicated fracture of third rib
1. Complicated (qualifier value)
2. Complicated fracture of bone (disorder)
3. Fracture of bone (disorder)
4. Fracture of one rib (disorder)
5. Bone structure of third rib (body structure)
I. {1, 3, 5}
II. {1, 4, 5}
III. {2, 5}
←

Fig. 3. Example that shows the annotation process and the different possibilities of how to combine available codes for coding the text fragment "Complicated fracture of third rib" with SNOMED CT. SNOMED CT concepts (on the left) are used in three combinations (on the right).

- Digits, e.g. "eight", "8".

The annotation guidelines also include some recommendations of concept preference. Anatomy concepts that contain the word "Structure" should be given preference about those that are distinguished by the term "Entire". Finding /Disorder concepts should be given preference over corresponding "Body structure" concepts. For all lab values, preference should be given to those concepts that include the term "measurement", such as "Measurement of potassium in serum". "Substance" concepts are given preference over "Drug product" concepts.

Fig. 3 shows an example of groupings of SNOMED CT codes that could be used to annotate the text fragment "Complicated fracture of the third rib" and which is the best according to the ASSESS CT annotation guidelines. The first code "Complicated" fully covers the token "complicated". The second code "Complicated fracture of bone" partially covers the text fragment. The third code "Fracture of bone" partially covers the tokens "fracture of third rib" but it is more generic than the fourth code "Fracture of one rib", which also partially covers the same tokens. The fifth code "Bone structure of third rib" fully covers the tokens "third rib". Because there is no single code that fully covers the whole text fragment the annotators must use a set of codes. Fig. 3 presents three examples of groupings which fully cover the text fragment. The first annotation group contains the first, third and the fifth codes. The second annotation group contains the same codes but the fourth code instead of the third code, which has a more detailed meaning. The third annotation group contains the second and fifth code and fully covers the meaning of the text fragment like the other two annotation groups but also has the fewest number of codes. Consequently, annotators should use the third annotation group. In cases were the full meaning of the text cannot be covered by any set of codes, annotators should repeat the same process but try to achieve the highest coverage possible with fewer codes.

Appendix D. Annotation experiment setting

We assigned one spreadsheet to each annotator with the same set of text samples (from 20–60) for their corresponding terminology settings. The number of text samples in a spreadsheet depended on the available effort declared by each annotator. Texts were distributed among annotators in a way that a subset of 20 samples (the same ones for all languages) was annotated twice for computing the inter-annotator agreement.

We defined the following criteria for the recruitment of annotators: (i) annotators must be physicians or medical students in their last year; (ii) experience with medical terminologies is desirable but not compulsory; (iii) at least two annotators should be recruited for each language; and (iv) the mother tongue of annotators must coincide with the language of the corpus. The recruited annotators were trained in our annotation guidelines by providing them training material and online seminars about their use in the diverse terminology settings. As training material, a clinical text sample, different from those in the text corpora, was provided to the annotators before the seminar so that they could practice before and address doubts during the seminar. The annotators were also provided with a contact from the ASSESS CT WP2 group to ask further questions related to the annotation guidelines, but not related to specific problems about text samples.

The resulting annotations were post-processed to facilitate the analysis and evaluation. The main goal of post-processing task is to avoid trivial annotation inconsistencies due to carelessness of annotators. We used the resulting annotations to automatically correct such inconsistencies. Usually, we completed the annotation spreadsheet with missing information and corrected annotation contradictions. We performed the following tasks for missing annotations:

- If there was a code without coverage score, then the score was set to Partial coverage.
- If there was a coverage score without code, then the coverage score was set to None.
- If there was a concept coverage score without term coverage, then the term coverage was set to No.
- If there was a term coverage value without concept coverage score, the term coverage value was removed.
- If a token in a chunk should have been annotated but the annotator only included the annotation for one terminology setting, then the concept coverage of the token was set to None.

Table 15
Example of annotated chunk in a text sample. The chunk is annotated with two different codes, two different concept coverage scores and two term coverage values.

Line	Token	Chunk	Code	Coverage score	Term coverage
1	He				
2	is				
3	currently				
4	on				
5	a				
6	120				
7	cc	8	414738006	Full coverage	No
8	per	8	414738006	Full coverage	Yes
9	kg	8	414738006	Full coverage	No
10	per	8	414738006	Full coverage	Yes
11	day	8	414738006	Full coverage	Yes
12	of				
13	Enfamil 8			None	No
14	20				
15	calories	8	258790008	Full coverage	Yes
16	.				

Trivial annotation issues were those inconsistencies that could be automatically detected and fixed. The following methods were applied:

- When a code did not belong to its terminology setting, then, an equivalent code was found in the corresponding terminology setting. For example, a UMLS CUI was used to annotate a token in the SCT_ONLY setting and via the UMLS mappings the corresponding SNOMED CT code was identified. In case of no direct mapping, the codes were removed, and the concept coverage score was set to None.
- When a code did not belong to any terminology, it was checked whether the value in the cell above was valid, which explains that the error was committed by the Excel's auto-increment mechanism. Here, the wrong code was replaced by the preceding value of the same chunk. In all other cases, the code was removed, and the concept coverage was set to None.

Appendix E. Calculation of the study endpoints

Some guidelines are used to calculate the concept coverage: (i) unannotated tokens in a chunk are omitted from the analysis, i.e. we ignore tokens which have not any concept coverage score; (ii) we obtain the set of unique codes per chunk, i.e. duplicated codes in a chunk are considered only once; and (iii) consecutive tokens with None-coverage are treated as a single concept coverage score. For example, the annotations (lines 7–15) in Table 15 are evaluated as an annotation group with the codes "414738006—Milliliter/kilogram/day (qualifier value)" and "258790008—calorie (qualifier value)" with *Full coverage* and one annotation with *Non-coverage* associated with the token "Enfamil". Therefore, the resulting concept coverage is 75%. The *Strict* and *Loose* coverage is the same due to lack of annotations with *Partial coverage*. Only the most restricted concept coverage scores are considered in case different coverage scores are attached to the same code in a chunk. In our analysis, the rank of coverage scores from the most to the least restricted is *Non-coverage, Partial coverage, Inferred coverage and Full coverage*.

Term coverage, in contrast to concept coverage, is analyzed at token level. It represents the percentage of matched tokens. The tokens which are not part of a chunk are omitted from the term coverage analysis. For example, the term coverage of the annotation example in Table 9 is 57%.

IAA is evaluated applying the weighted Krippendorff's alpha coefficient with the results from the annotators of the same language. Alpha requires equivalent annotation units among annotators in terms of scope. Therefore, we defined units as the set of annotations that are enclosed in a sentence. The number and scope of the sentences are the same for annotators of the same language since they must annotate the same text samples. This is not possible between annotators of different languages because the text samples are translated, and the distribution of the tokens and sentences could be different. The rating table represents the units in the rows and the annotators in the columns; the coincidence table is a square matrix with the list of unique units as the dimension and provides the number of times one unit coincides with another; the weight table have the same structure as the coincidence table, but it represents the overlap between two units. The overlap is measured based on the total number of unique annotations between two units. The overlap is calculated using the quadratic weight (see Eq. (1)).

References

[1] N.F. de Keizer, A. Abu-Hanna, J.H.M. Zwetsloot-Schonk, Understanding terminological systems I: terminology and typology, Methods Archive 39 (1) (2000) 16–21.

[2] J. Ingenerf, W. Giere, Concept-oriented standardization and statistics-oriented classification: continuing the classification versus nomenclature controversy, Methods Archive 37 (1998) 527–539.

[3] F. Freitas, S. Schulz, E. Moraes, Survey of current terminologies and ontologies in biology and medicine, RECIIS Electron. J. Commun. Inf. Innov. Health 3 (2009) 7–18.

[4] S. Schulz, J.-M. Rodrigues, A. Rector, C.G. Chute, Interface Terminologies, Reference terminologies and aggregation terminologies: a strategy for better integration, Stud. Health Technol. Inform. 245 (2017) 940–944.

[5] S.T. Rosenbloom, R.A. Miller, K.B. Johnson, P.L. Elkin, S.H. Brown, Interface terminologies: facilitating direct entry of clinical data into electronic health record systems, J. Am. Med. Inform. Assoc. 13 (3) (2006) 277–288.

[6] J.J. Cimino, V.L. Patel, A.W. Kushniruk, Studying the human-computer-terminology interface, J. Am. Med. Inform. Assoc. 8 (2) (2001) 163–173.

[7] D. Lee, N. de Keizer, F. Lau, R. Cornet, Literature review of SNOMED CT use, J. Am. Med. Inform. Assoc. 21 (e1) (2014) e11–e19.

[8] ASSESS CT Consortium. ASSESS CT - Assessing SNOMED CT for Large Scale eHealth Deployments in the EU. Available from: http://assess-ct.eu/start0/.

[9] ASSESS CT Consortium. ASSESS CT Deliverables; Available from: http://assess-ct.eu/deliverables0.html.

[10] Averbis GmbH. Averbis Terminology Platform, (2014) Available from: http://apps.averbis.de/atp/.

[11] ISO/TC 215 Health informatics technical comittee, ISO/TR 12300:2014 Health Informatics – Principles of Mapping Between Terminological Systems, ISO, 2014.

[12] Resources for ASSESS CT - Quantitative Analysis, (2017) Available from: http://user.medunigraz.at/jose.minarro-gimenez/assessct_index.html.

[13] Advanced Analytics LLC. Agreestat: Inter-Rater Reliability With R, (2010) Available from: http://www.agreestat.com/r_functions.html.

[14] K. Krippendorff, Content Analysis: an Introduction to Its Methodology, 2nd ed., Sage, Thousand Oaks, Calif, 2004.

[15] ASSESS CT Consortium. ASSESS CT WP2.3 analysis resources; Available from: https://github.com/joseminya/ASSESSCT_WP2_T2-3.

[16] J.E. Andrews, R.L. Richesson, J. Krischer, Variation of SNOMED CT coding of clinical research concepts among coding experts, J. Am. Med. Inform. Assoc. 14 (4) (2007) 497–506.

[17] J.Y. Ng, H.S. Boon, A.K. Thompson, C.R. Whitehead, Making sense of “alternative”, “complementary”, “unconventional” and “integrative” medicine: exploring the terms and meanings through a textual analysis, BMC Complement. Altern. Med. 16 (2016).

[18] D. Karlsson, K.R. Gøeg, H. Örman, A.R. Højen, Semantic Krippendorff’s alpha for measuring inter-rater agreement in SNOMED CT coding studies, Stud. Health Technol. Inform. 205 (2014) 151–155.

[19] K. Krippendorff, Computing Krippendorff’s Alpha Reliability, Available from: (2011) http://repository.upenn.edu/asc_papers/43.

[20] ASSESS CT Consortium. ASSESS CT Final deliverable; Available from: http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf.

[21] J.A. Miñarro-Giménez, C. Martínez-Costa, P. López-García, S. Schulz, Building SNOMED CT post-coordinated expressions from annotation groups, Stud. Health Technol. Inform. 235 (2017) 446–450.