DELIVERABLE

Project Acronym: **ASSESS CT**
Grant Agreement number: **643818**
Project Title: **Assessing SNOMED CT for Large Scale eHealth Deployments in the EU**

D2.4 – Use of terminologies for representing structured and unstructured clinical content

Stefan Schulz  
Medical University of Graz (Austria)

Jose Antonio Miñarro-Giménez  
Medical University of Graz (Austria)

Marie-Christine Jaulent  
Päivi Hämäläinen

INSERM, Paris (France)

National Institute for Health and Welfare (THL) (Finland)

Daniel Karlsson  
University of Linköping (Sweden)

Kirstine Rosenbeck Gøeg  
University of Ålborg (Denmark)

Catherine Chronaki  
Veli Stroetmann

HL7 Europe

Empirica, Bonn

Kornél Markó  
Averbis GmbH, Freiburg (Germany)

Benny Van BruwaeneRonald Cornet  
Brussels (Belgium)  
Academic Medical Center, Amsterdam (The Netherlands)

Dipak Kalra  
Eurorec, Brussels
Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Author</th>
<th>Organisation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28 Jul 2016</td>
<td>Stefan Schulz</td>
<td>Medical University of Graz</td>
<td>Update</td>
</tr>
<tr>
<td>2</td>
<td>28 Jul 2016</td>
<td>Kornél Markó</td>
<td>Averbis GmbH</td>
<td>Revision of NLP-Chapter</td>
</tr>
<tr>
<td>3</td>
<td>29 Jul 2016</td>
<td>MC Jaulent</td>
<td>Inserm</td>
<td>Revision of the document</td>
</tr>
<tr>
<td>4</td>
<td>3 Aug 2016</td>
<td>Ronald Cornet</td>
<td>Academic Medical Center</td>
<td>Revision of the document</td>
</tr>
<tr>
<td>5</td>
<td>9 Aug 2016</td>
<td>Stefan Schulz</td>
<td>Medical University of Graz</td>
<td>Final Revision</td>
</tr>
<tr>
<td>6</td>
<td>11 Nov 2016</td>
<td>Stefan Schulz, Daniel Karlsson</td>
<td></td>
<td>Post-review revision</td>
</tr>
</tbody>
</table>

Date of delivery

<table>
<thead>
<tr>
<th>Contractual:</th>
<th>Actual:</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 Jul 2016</td>
<td>10 Aug 2016</td>
</tr>
</tbody>
</table>

Status: final /draft

Abstract (for dissemination)

Deliverable 2.4 introduces the notion of user interface terminologies in contrast to reference terminologies. Three terminology settings (SNOMED CT against an alternative, UMLS-derived hybrid terminology and a local terminology collection), are analysed under user interface terminology aspects. It investigated the coverage of such interface terms in these three terminology scenarios against clinical text samples in six languages, using natural language processing and manual annotations. It furthermore proposes a method on semi-automated creation of user interface terms and describes its implementation, ongoing maintenance and formative evaluation.

Keywords

SNOMED CT, UMLS, Terminology coverage, User interface terminology, Electronic Health Records, Semantic Interoperability, Natural language processing, Terminology acquisition, eHealth in Europe

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.
Table of Content

Revision History, Status, Abstract, Keywords, Statement of Originality .................................2
1  Executive Summary  ..................................................................................................................4
2  Introduction ..............................................................................................................................5
   2.1  About this document ...........................................................................................................5
   2.2  Goals and objectives of ASSESS CT ................................................................................5
   2.3  ASSESS CT Workpackage 2 .............................................................................................5
   2.4  Terminology ecosystems ..................................................................................................7
   2.5  Terminology localisation .................................................................................................8
3  Natural Language Processing .....................................................................................................9
   3.1  Terminologies ....................................................................................................................9
   3.2  Terminology settings .........................................................................................................10
       3.2.1  SCT_ONLY ................................................................................................................10
       3.2.2  UMLS_EXT .............................................................................................................10
       3.2.3  LOCAL ..................................................................................................................10
   3.3  Test Corpus .......................................................................................................................10
   3.4  Methods for NLP-based annotation ................................................................................11
4  Results of NLP-based Annotation ..........................................................................................14
   4.1  Human Annotator Agreement Evaluation Results as Reference ..................................14
   4.2  NLP Pipeline Evaluation ................................................................................................15
5  Interpretation of NLP experiment ..........................................................................................18
6  Term coverage in manual annotation experiments ...............................................................19
   6.1.1  Term coverage .............................................................................................................19
7  Methods for semi-automated acquisition of SNOMED CT user interface terms ..............20
   7.1  Previous work .................................................................................................................20
   7.2  Term generation principles .............................................................................................20
   7.3  Terminology localization workflow .................................................................................21
   7.4  Implementation ................................................................................................................22
8  Results of user interface term acquisition ..........................................................................24
9  Discussion of proposed user interface terminology creation method ..............................25
10 Annex - Example translations (ASSESS CT method vs. Google translate) of English SNOMED CT Descriptions into German ..............................................................26
1 Executive Summary

Workpackage 2 of ASSESS CT targets empirical evidence for the fitness for purpose of SNOMED CT, compared to other terminology scenarios. In a variety of experiments using terminology settings reflecting the ASSESS CT scenarios, aspects of terminology use have been studied, and results have been collected and analysed. Results related to terminology coverage and inter-annotator agreement have been reported in the ASSESS CT deliverable D2.3.

Deliverable D2.4 introduces the notion of user interface terminologies in contrast to reference terminologies. Whereas reference terminologies introduce concepts, standardised names and definitions, user interface terminologies bridge between the language expressions used in clinical practice (in different languages and contexts) on the one hand and reference terminologies on the other hand.

Three terminology settings (SNOMED CT only against an alternative, UMLS-derived hybrid terminology and a local German terminology collection), are analysed under user interface terminology aspects. The coverage of user interface terms provided by these three terminology scenarios against clinical text samples in six languages was measured, both using natural language processing technology and manual annotations.

The UMLS based scenario showed a generally better term coverage, compared with the SNOMED CT only scenario, with the exception of Swedish (which is known to be poorly represented in UMLS). In English, the term coverage difference between SNOMED CT and the UMLS scenario is not significant. The much lower SNOMED CT term coverage compared to English (with the same concept coverage, cf. D4.3) shows the importance of user interface terms: the English SNOMED CT includes numerous synonyms whereas the Swedish version has only one term per concept.

The NLP results show no consistent picture across the language scenarios under consideration. Thus, no conclusion can be made that would favour either scenario. Whilst the evaluation values are generally poor compared to other NLP evaluations published (involving a “real” gold standard instead of partly contradicting annotations from two annotators, cf. D4.3), the results are more convincing when compared to the human interrater agreement. In this sense, NLP reaches 68% of the human performance in the UMLS_ONLY scenario, 79% in the SNOMED CT scenario, and as much as 90% in the German local terminology scenario, averaged over all languages. Especially the latter result shows how close NLP results can come to manual annotation results. This is especially noteworthy because standard NLP pipelines were used without special tuning to specific language or sublanguage peculiarities.

The field of user interface terminologies was further explored by an iterative semi-automated approach for creation of a German interface terminology linked to UMLS, already showing good results with a moderate input in human labour, compared to the cost of full professional translation. It appears, however, that the creation and maintenance of user interface terminologies cannot depend on a top-down approach only. The generated terms need to be constantly checked against the terminology, and gaps need to be filled. User-, language-, and domain-specific user interface terms have to be accounted for. This suggests a distributed, bottom-up development of user interface terminologies.

Summing up, two main messages can be distilled out of this deliverable, regarding overall recommendations for SNOMED CT as a core reference terminology. Firstly, the need to emphasize user interface terminology aspects, not primarily as parts of localised SNOMED versions but rather as separate artefacts, and secondly, more emphasis to be put on the analysis of clinical free text, the role of user interface terminologies in this use case, and the development of suited NLP methods and tools.
2 Introduction

2.1 About this document

This document completes the interim deliverable (D2.2) of the ASSESS CT project and presents all work done in ASSESS CT on language aspects of clinical terminologies. Entitled "Use of terminologies for representing structured and unstructured clinical content" it is partly based on content of deliverable D2.3 "Multilingual and multidisciplinary study of terminology coverage and quality". As both topics are closely intertwined, D2.4 should be read as an extension of D2.3, but for reporting purposes both documents need to be separate. This explains some repetitions of D2.3 content.

D2.4 presents findings of a Natural Language Processing (NLP) experiment. Here the view on representation is broadened, regarding the investigation of an alternative to manual assignment of terminology codes to unstructured or semi-structured clinical content. Secondly, D2.4 presents results on term coverage in manual annotations. Finally, the concept of user interface terminology is introduced and juxtaposed to the concept of reference terminology. It is demonstrated how collections of clinical terms in a new language can be semi-automatically generated and linked to a reference terminology.

2.2 Goals and objectives of ASSESS CT

The goal of ASSESS CT is to improve semantic interoperability of eHealth services in Europe by investigating the fitness of the international clinical terminology SNOMED CT, the world's most comprehensive multilingual healthcare terminology, as a potential standard for EU-wide medical documentation. SNOMED CT claims to provide codes for comprehensively representing the content of health records.

As health care systems are organized nationally, the EU has not taken any steps so far towards the adoption of a standardized health terminology. Up to now, SNOMED CT has been introduced in roughly half of the EU member states. However, as the mobility of EU citizens increases and national boundaries lose their importance in a more and more internationalized market for health care services, the question of interoperability of health care data gains importance at a European level. The ASSESS CT consortium has addressed this challenge by investigating the current use of SNOMED CT, analysing reasons for adoption / non-adoption, and identifying success factors, strengths and weaknesses related to SNOMED CT and to other alternative terminologies.

The SNOMED CT adoption is scrutinized against two alternative scenarios, viz. (i) to abstain from actions at the EU level, and (ii) to devise an EU-wide semantic interoperability framework alternative without SNOMED CT. These scenarios were addressed in WP2 through three different terminology settings: SNOMED CT only (SCT_ONLY), a UMLS-derived alternative terminology set (UMLS_EXT, a UMLS subset extended with localised versions of international medical terminologies), and a German only terminology setting (LOCAL) corresponding to a scenario where each country maintains their own terminology without or with minimal EU level coordination.

2.3 ASSESS CT Workpackage 2

The ASSESS CT Workpackage 2 is conducting comparative studies, all of which attempt the following two questions:

- How well does SNOMED CT address selected use cases, compared to an alternative setting, which uses a mix of existing terminologies without SNOMED CT, adapted to the needs of EU member states?
• How well does SNOMED CT address selected use cases, compared to the current state of affairs, i.e. sticking to the terminologies already in used across EU member states?

All use cases are committed to the overall goal of semantic interoperability, assuming its positive impact on patient safety and health service cost-effectiveness. ASSESS CT assumes that this requires standardization of meaning across national and linguistic boundaries. As a result, semantic artefacts are required that introduce language-independent meaningful units (concepts) in a precision and granularity sufficient for clinical documentation purposes across clinical disciplines and specialties. These concepts should ultimately be unambiguous by means of formal or textual definitions, as well as via self-explaining labels aka fully specified names.

The importance of building repositories of words, multiword expressions and short forms that reflect the users' jargon, dependent on languages, dialects, user groups, institutions and professional disciplines has often been neglected. So-called user interface terms are often added as synonyms to clinical terminologies, but do not reach an ideal coverage for their use in machine processing of text or the support of retrieval in terminologies needed for browsing or structured data entry.

This topic will be repeatedly addressed in this document, as it refers to each of the three use cases Workpackage 2 has addressed:

• **Use of SNOMED CT vs. other terminologies for machine annotation of clinical text in different languages.** Natural language continues to be the main carrier of clinical information, in findings reports, clinical evolutions, discharge summaries, orders, prescriptions and problem lists. Electronic Health Record (EHR) systems have substituted paper charts, but often with no change of content structure. Therefore, human written text now exists in computer system, often with minimal document structure, exhibiting idiosyncrasies like highly compacted language, often with idiosyncratic and error-laden terms and passages. New medical terms and abbreviations are constantly being generated and adopted. This is an ongoing challenge for natural language processing (NLP) approaches. Although this discipline has made important advances regarding tools and techniques to analyse human language and map it to controlled terminologies or ontologies, the bottleneck of lexical coverage continues. This use case uses an off-the-shelf text processing pipeline tailored to several languages.

• **Use of SNOMED CT vs. other terminologies for manually annotating clinical texts in different languages.** This is mainly justified by the fact that natural language documents contain the terms clinicians use in their daily practice. The more easily these terms can be linked to concepts in a terminology, the higher is its quality. This depends on two aspects, viz. (i) the granularity of content provided by the concepts in the terminology (concept coverage) and (ii) the wealth of clinically relevant synonyms or entry terms in the terminology (term coverage). Another quality criterion is inter-annotator agreement: the more the annotation results coincide between two or more annotators, the more precisely defined and/or self-explaining is the terminology.

• **Use of SNOMED CT vs. other terminologies for providing textual values for structured data entry forms.** Despite the predominance of text, structured data entry is increasingly important in clinical documentation, especially for clinical research, quality control, disease registries, and billing. The structuring of clinical information is provided by binding the meaning of the data elements of information models to external terminologies and by constraining value sets for coded data elements. Users are often reluctant to structured data entry, especially in those environments where the standard entry of clinical content is done by dictating free text. Structured data entry requires a good representation of the terms clinicians actually use, and large value sets require advance navigation and retrieval support to be efficiently used.
This is, again a use case for user interface terminologies supported by text processing tools.

All use cases provide indicators for SNOMED CT’s technical fitness for use. As technical fitness for use is a prerequisite for clinical fitness for use, and samples of clinical data have been used for the studies, clinical fitness for use can be indirectly assessed. The evidence created by studies proposed in WP2 is assumed to disseminate knowledge about the current state of SNOMED CT, in order to inform policy dialogues and strategic planning processes that are necessary to set the course for EU-wide clinical reference terminologies on the one hand, and to give guidance for the collaborative creation of user interface terminologies and their linkage to user reference terminologies.

2.4 Terminology ecosystems

The apparent heterogeneity of different terminology types, especially regarding the focus of representation and the intended context of use has been intensively discussed in ASSESS CT, which has come up with the following delineation of user interface terminologies from other terminologies, which is also the basis of the strategic recommendations set forth in D4.3 and D4.4. This typology uses the metaphor of an ecosystem, which is populated by different organisms, which exhibit multiple interactions and dependencies in order to present the best possible adaptation to a given environment. A terminology ecosystem can be considered to consist of these three categories of terminologies (cf. Fig. 1):

- Reference terminologies describe the meaning of terms of a domain, together with the properties of the objects that these terms denote, in a neutral sense, i.e. uncommitted to any specific purpose. Representational units of reference terminologies are commonly called "concepts". The meaning of concepts should be the same across languages. It is given by textual definitions, formal definitions and/or maximally unambiguous terms / labels in different languages, such as SNOMED CT FSNs. Reference terminologies are not however expected to cover the complete scope of the language of the end users. Thus, reference terminologies mainly focus on ontological aspects. A core reference terminology (the discussed role of SNOMED CT, as the leading issue in ASSESS CT) is a terminology that plays a pivotal role within a terminology ecosystem.

- For completeness (not being focus of this document), aggregation terminologies (classifications like ICD-10) are systems of non-overlapping classes in single hierarchies, used for data aggregation and ordering. They are typically used for statistics and for reimbursement. Aggregation terminologies, again, are not however expected to cover the complete scope of end user language.

- User interface terminologies\(^1\) are collections of all terms that are used in written and oral communication within a group of users, for example in a data entry form in a healthcare IT system. This requires that entries in user interface terminologies need to be described not only in terms of the natural language they belong to, but also by dialect, time, clinical specialty and professional group. User interface terms tend to be ambiguous, acquiring their semantics by linkage to reference terminologies. User interface terms constitute value sets for data entry as well as dictionary underlying human language processing systems.

---

\(^1\) Following the notion of interface terminology by Rosenbloom ST et al. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277-88. Epub 2006 Feb 24. Due to the different facets of the term "interface" (user interfaces vs. machine interfaces) ASSESS CT has coined the term "user interface terminology"
2.5 Terminology localisation

Terminology translation issues have frequently been addressed with regard to SNOMED CT. The IHTSDO publishes SNOMED CT English and Spanish only. For localization in other languages, the IHTSDO has published translation principles\(^2\). A number of translations of SNOMED CT, which are managed by member countries have been created according to these principles. Access to these translations can be retrieved through the relevant National Release Center, some of them are also accessible via IHTSDO's SNOMED CT browser. With the growth of IHTSDO membership, non-native English speaking countries are joining. It is therefore anticipated that the desire for localisations will increase. Translation / localization issues are often perceived as a barrier to SNOMED CT adoption. However, the need of a complete translation of SNOMED CT has been controversially discussed. The application of IHTSDO's translation guidelines focuses on the translation of fully specified names (FSNs). The creation of fully specified terms follows strict principles, in order to guarantee that they are unambiguous and as self-explaining as possible. This reduces the usefulness of FSNs as user interface terms\(^3\), e.g. for value set creation and natural language processing. Therefore, the question arises whether the introduction of SNOMED CT requires, apart from the costly translation of the terminology according to the guidelines, the creation and maintenance of a broad range of user interface terms, which could be perceived as an additional road blocker on the route to SNOMED CT adoption. It is therefore proposed to separate the roles of reference terminology and user interface terminology as shown in Fig.1. The provision of a broad range of close-to-user terms would therefore not be the primary goal for SNOMED CT, which should focus on its role as a reference terminology, to which a collection of user-, language-, and domain-specific user interface terminologies refer.

---


3 Natural Language Processing

A natural language processing (NLP) pipeline is built to extract specific information units from unstructured texts. In ASSESS CT, the task was to identify terminology codes by mapping excerpts from clinical narratives to the three terminology settings. The underlying text analytics pipeline extracts information units that belong to several semantic categories, such as findings, disorders, procedures, drugs, etc., and their relations from unstructured text. It consists of several NLP components that can be individually adapted to different use cases to process text written in different languages. These text analysis components, so called Analysis Engines (AEs), stuck together in the Apache UIMA\(^4\) (Unstructured Information Management Architecture) framework building an overall solution for different use cases. The UIMA framework provides all necessary methods to create custom annotators and to put them together as an aggregated analysis engine (AAE) consisting of a sequence of multiple AEs. Each of the AEs takes as input the output (annotations) of previous AEs and generates new annotations, building up structured information. Annotations between AEs are exchanged using an object named Common Analysis System (CAS). The CAS is UIMA’s object-based data structure that allows memory-based storage and exchange of annotations with respect to pre-defined type systems of hierarchically organized annotations.

3.1 Terminologies

In order to respond to the ASSESS CT requirements, viz. comparing SNOMED CT to alternative terminology options, the three custom terminology settings were used. They were filtered by selected UMLS Semantic Groups\(^5\). These groups constitute pairwise disjoint divisions of all Metathesaurus. Via SNOMED CT – UMLS mappings, the same semantic groups are also used to partition SNOMED CT. Table 1 gives an overview of the terminology sizes (number of concepts and number of synonyms) used in the experiments.

### Table 1. Terminology content for each language that were used in the annotation experiments (SNOMED CT translation into German is an unofficial version that was produced for internal use only)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Language</th>
<th># Concepts</th>
<th># Entry terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>German*</td>
<td>256,256</td>
<td>645,159</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>English</td>
<td>267,092</td>
<td>699,062</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>69,977</td>
<td>70,217</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>72,368</td>
<td>72,550</td>
</tr>
<tr>
<td></td>
<td>Swedish</td>
<td>254,382</td>
<td>254,749</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>85,268</td>
<td>206,890</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1,977,431</td>
<td>3,820,180</td>
</tr>
<tr>
<td>UMLS_EXT</td>
<td>Finnish</td>
<td>31,706</td>
<td>40,728</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>179,860</td>
<td>265,715</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>169,799</td>
<td>325,214</td>
</tr>
<tr>
<td></td>
<td>Swedish</td>
<td>32,002</td>
<td>36,314</td>
</tr>
<tr>
<td>LOCAL</td>
<td>German</td>
<td>36,481</td>
<td>161,103</td>
</tr>
</tbody>
</table>

\(^4\) http://uima.apache.org/
\(^5\) ANAT = Anatomy, CHEM = Chemicals & Drugs, CONC = Concepts & Ideas, DEVI = Devices, DISO = Disorders, GENE = Genes & Molecular Sequences, LIVB = Living Beings, OBJC = Objects, PROC = Procedures
3.2 Terminology settings

3.2.1 SCT_ONLY

SNOMED CT, international version (English) July 2015, Swedish version, as well as Dutch and French fragments provided by the Belgian government, where parts of SNOMED CT are currently being localised, were included. Only concepts from selected UMLS semantic groups were used, in order to exclude terminology content that is outside of our clinical domain scope. We also excluded the SNOMED CT “Situation” hierarchy, which provides a mechanism to embed terminology codes in epistemic, temporal and other contexts. The reason for the exclusion is that we consider the representation of context as belonging to information models, not terminologies, with the “Situation” hierarchy constituting an information model inside SNOMED CT. In order to guarantee comparability with the other scenarios, the SNOMED CT compositional syntax is not used; composition is simply expressed by unordered groupings of codes. This terminology setting is named SCT_ONLY.

3.2.2 UMLS_EXT

In order to address an alternative terminology setting, an alternative hybrid terminology was created. Starting point was the UMLS Metathesaurus, a repository of over 160 biomedical terminologies in different languages, with all of their content linked to unique identifiers (CUIs). Criteria for inclusion in the ASSESS CT alternative setting were sources that are actively updated. From these, we excluded: (i) sources the use of which makes only sense in an U.S. context, such as U.S: drugs, (ii) sources in languages other than English, Dutch, French, German, Swedish, Finnish, (iii) earlier SNOMED and Read code versions, (iv) sources out of scope regarding our data (nursing, dentistry). This terminology setting is termed UMLS_EXT.

After delineating the scope of terminologies in UMLS_EXT, available localised terminology content available for other languages (but not in UMLS) was linked to their English language homologues, in order to optimise this scenario for non-English use cases.

The English version is the largest with about two million concepts and four million terms. In comparison, the Swedish subset is small with 32,000 concepts, mostly represented by one term only. The Dutch and French subsets of UMLS_EXT have about 175,000 concepts, with 284,000 terms (French) and 342,000 (Dutch). The German subset has about 85,000 concepts and 228,000 terms.

3.2.3 LOCAL

A set of terminologies for a strictly local setting was compiled for German, which included terminologies only available for German without mappings to UMLS or other terminologies, e.g. a German procedure terminology. This terminology setting is termed LOCAL. It included the German versions of ICD-10, ATC, MeSH, together with the German drug vocabulary ABDAMED and the German procedure terminology OPS301.

3.3 Test Corpus

For both the manual and machine annotation tasks, a multilingual corpus was necessary. To this end, clinical texts in six languages (Dutch, English, French, Finnish, German, and Swedish) were collected. The acquisition of corpora was done in a way supposed to approximate representativeness in terms of clinical domains, document sections, and document types. Finally, 60 document snippets (400 – 600 characters), 10 for each language were selected. Each snippet was translated into all other languages by professional translators. After translation, numerous corrections were added by the consortium members. In addition, country-specific drug names were substituted by the names of their active
ingredients, whenever possible. The output was a parallel corpus consisting of 60 text snippets per language. Since the corpus consisted of samples from each language in equal terms regarding both raw translation and subsequent translation checking by medical terminology specialists for all six languages, language specific bias could be minimised. From these 60 text samples, a subset of 20 was annotated by each annotator. This subset served as the reference corpus for our evaluation studies in which we compared the annotations (unique concept identifiers on the document level) between:

1. Annotator 1 vs. annotator 2 and vice versa (and resulting average values)
2. NLP pipeline vs. annotator 1, NLP pipeline vs. annotator 2 (and resulting average values)

### 3.4 Methods for NLP-based annotation

For ASSESS CT, a simple pipeline has been set up that allows for generalizability since it does not include peculiar features (such as ellipsis resolution or resolution of enumerations, or more sophisticated language-specific customizations. Rather than this, the focus has been set on observing differences in the annotation quality when different terminologies come into play (rather than evaluating the underlying NLP system itself). It included the following components:

**Sentence Detection**: Due to the ambiguity of punctuation, the problem of sentence boundary recognition is known to be a non-trivial task. For instance, a full stop might appear in arbitrary contexts such as inside abbreviations or as a decimal point, resulting in different meanings. The sentence detector used for ASSESS CT relies on machine learning methods provided by the OpenNLP framework.

**Tokenizer**: Sequences of characters are split into individual tokens. Tokens do not only describe single words but also punctuation marks such as periods and commas. Just like in the case of the sentence recognition problem, words itself may contain numbers (“B2B”), additional characters (“Dr.”), or dashes (“e-mail”), making tokenization a non-trivial task. The pipeline uses a rule-based approach which can be applied to different languages in different domains. The tokenizer takes sentences annotated by the sentence detector as input.

**Stemming**: A stemmer reduces each token to its word stem, e.g. “mobiles” and “mobility” are reduced to the stem “mobil”. Reducing the number of inflections by mapping terms to their stem improves the concept mapping process and reduces the number of features in the case of statistical methods. The stemmer is rule-based and is available for a number of languages.

**Information extraction**: With the annotations generated by the previous AEs the concept mapper component is able to create mappings between concepts of a terminology and free-text phrases in various morphological layers. For ASSESS CT, the token and the stem representations are used for the mapping procedure. Obviously, the background terminology essentially determines the overall quality of the mapping, as evidenced by the following text examples that correspond to the concepts “Metastasis to lung” (285604008 in SNOMED CT) or “Lung metastasis” (C0153676 in the UMLS), in English and German:

- … lung metastasis …
- … pulmonal metastasis …
- … lung filia …
- … pulmonal filia …

---

7 https://opennlp.apache.org/
8 http://snowball.tartarus.org/
• ... pulmonal metastases ...
• ... pulmonal relapse of a metastasis ...
• ... pulmonal filiae ...
• ... pulmonary metastases ...
• ... pulmonary relapse of a metastasis ...
• ... metastases in the lung ...
• ... Lungenmetastasen ...
• ... Lungenfiliae ...
• ... pulmonale Metastasierung ...

Figure 2 shows concept annotations, i.e. sequences of words that map to terminological entries in a reference system (here: UML\_EXT).

Figure 2. Example of concept annotations by mapping text fragments to terminologies 
(here: UML\_EXT)

Figure 2 shows an example of the result of the manual annotation for the text fragment “Breast left surgery area soft and good around the scar. No surface irregularities” using SNOMED CT and the extended UMLS as reference terminologies.

Table 2. Sample of manual semantic annotations for the scenarios SCT\_ONLY and UMLS\_EXT, respectively. Repeated codes mark annotations in which more than one token corresponds to a code

<table>
<thead>
<tr>
<th>English Tokens</th>
<th>SNOMED ID</th>
<th>UMLS CUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>290061004</td>
<td>C0567478</td>
</tr>
<tr>
<td>left</td>
<td>7771000</td>
<td>C1282914</td>
</tr>
<tr>
<td>surgery</td>
<td>422411000</td>
<td>C0334150</td>
</tr>
<tr>
<td>area</td>
<td>422411000</td>
<td>C0334150</td>
</tr>
<tr>
<td>soft</td>
<td>290061004</td>
<td>C0567478</td>
</tr>
<tr>
<td>and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>20572008</td>
<td>C0205170</td>
</tr>
<tr>
<td>around</td>
<td>355648006</td>
<td>C1282914</td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scar</td>
<td>422411000</td>
<td>C0334150</td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No surface</td>
<td>410679008</td>
<td>C0205148</td>
</tr>
<tr>
<td>irregularities</td>
<td>49608001</td>
<td>C0205271</td>
</tr>
</tbody>
</table>

For the evaluation of the agreement of the human annotators, as well as for the text mining pipeline against humans, we counted human annotations found also by NLP (true positive= TP), human annotations not found by NLP (false negative= FN) and NLP annotations not
corresponding to any human annotation (false positives = FP) by a human annotator (1), respectively the text mining pipeline (2). The following standard metrics have been calculated based on counting all unique concepts identifiers per document and averaging the values for an overall view:

- Recall = \( \frac{TP}{(TP+FN)} \)
- Precision = \( \frac{TP}{TP+FP} \)
- F-score = \( 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \)
4 Results of NLP-based Annotation

4.1 Human Annotator Agreement Evaluation Results as Reference

We calculated true positive annotations, false negative ones, and false positive ones by comparing the annotations made by annotator 1 to those made by annotator 2, vice versa, and computing average values. Table 3 and Figure 3 show the resulting values in terms of F1-score values (for better comparability to NLP results). The LOCAL setting is only available for German.

Table 3. Average agreement values by comparing annotations from annotator 1 to those of annotator 2, and vice versa.

<table>
<thead>
<tr>
<th>Language</th>
<th>Scenario</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOCAL</td>
<td>0.61</td>
</tr>
<tr>
<td>German</td>
<td>UMLS_EXT</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>LOCAL</td>
<td>n/a</td>
</tr>
<tr>
<td>English</td>
<td>UMLS_EXT</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>LOCAL</td>
<td>n/a</td>
</tr>
<tr>
<td>Finnish</td>
<td>UMLS_EXT</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>LOCAL</td>
<td>n/a</td>
</tr>
<tr>
<td>French</td>
<td>UMLS_EXT</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>LOCAL</td>
<td>n/a</td>
</tr>
<tr>
<td>Dutch</td>
<td>UMLS_EXT</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>LOCAL</td>
<td>n/a</td>
</tr>
<tr>
<td>Swedish</td>
<td>UMLS_EXT</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Figure 3. Average agreement (measured with F$_1$-score) between annotator 1 and annotator 2 for the different evaluation scenarios.

In all language scenarios, the agreement for UMLS_EXT is higher than for SNOMED CT. An obvious reason might be the lack of synonyms in SNOMED CT (compared to the number of entry terms in UMLS_EXT) which complicates the identification of matching concepts by using terminology search tools.

4.2 NLP Pipeline Evaluation

The NLP pipeline described in this document was set up to process the test corpus twice (for comparing annotation results with those of annotator 1 and 2). Table 4, Figure 4 and Figure 5 (precision and recall) show the resulting average values. Figure 6 depicts F$_1$-score values of the NLP experiments compared to the agreement of the human annotators.
### Table 4. Average precision, recall and $F_1$-score values for the language and terminology scenarios considered.

<table>
<thead>
<tr>
<th>Language</th>
<th>Scenario</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>LOCAL</td>
<td>0.45</td>
<td>0.70</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>UMLS_EXT</td>
<td>0.51</td>
<td>0.62</td>
<td>0.56</td>
</tr>
<tr>
<td>English</td>
<td>UMLS_EXT</td>
<td>0.49</td>
<td>0.26</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.44</td>
<td>0.30</td>
<td>0.36</td>
</tr>
<tr>
<td>Finnish</td>
<td>UMLS_EXT</td>
<td>0.17</td>
<td>0.47</td>
<td>0.24</td>
</tr>
<tr>
<td>French</td>
<td>UMLS_EXT</td>
<td>0.51</td>
<td>0.41</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.43</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td>Dutch</td>
<td>UMLS_EXT</td>
<td>0.15</td>
<td>0.24</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.19</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>Swedish</td>
<td>UMLS_EXT</td>
<td>0.39</td>
<td>0.56</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>SNOMED CT</td>
<td>0.32</td>
<td>0.31</td>
<td>0.32</td>
</tr>
</tbody>
</table>

![NLP-Results: Recall](image)

**Figure 4.** Recall values for NLP experiments.
Figure 5. Precision values for NLP experiments.

Figure 6. F₁-score values of NLP experiments compared to the agreement values of human annotators
5 Interpretation of NLP experiment

The results show no consistent picture across the language scenarios under consideration. NLP performs better in the SNOMED CT scenario for English and Dutch, whilst UMLS EXT works better for French and Swedish. Here, two factors have to be taken into account:

- NLP crucially depends on the coverage of the lexicon. Only in its English version, SNOMED CT offers synonyms. The coverage of the UMLS Metathesaurus for synonyms is good for English, French and German, but poor for the other languages.
- Agreement is higher for scenarios where the concept coverage of the terminology (number of different concepts in terminology) is low.

Whilst the evaluation values are generally poor compared to other NLP evaluations published (involving a “real” gold standard instead of partly contradicting annotations from two annotators), the results are more convincing when compared to the human interrater agreement. In this sense, NLP reaches 68% of the human performance in the UMLS ONLY scenario, 79% in the SNOMED CT scenario as well as 90% in the abstain scenario, averaged over all languages. Interestingly, the annotation agreement between NLP and both annotators in the French SNOMED CT (and also the Finnish UMLS_EXT) setting (F1-score of 39%) is more consistent than the agreement between these human annotators themselves (F1-score of 30%).

As the results of the NLP experiment are still preliminary, the comparative analysis is not conclusive. For example, we do not yet know what content coverage can be obtained with NLP techniques. Consequently, this chapter is mostly a comparison of the free text annotation experiment and the terminology binding experiment.
6 Term coverage in manual annotation experiments

How well different terminology scenarios qualify as providers of user interface terms was studied, as a side issue, in the manual annotation experiments described in D2.3. The term coverage results are presented in this document, in order to complete the picture regarding user interface terms, also to better interpret the results from the NLP annotation experiment.

6.1.1 Term coverage

The results of term coverage calculation – as a measure of the suitability of a terminology setting as a source of user interface terms – are given by Table 5 and plotted in Figure 7. Term coverage is calculated for English, Swedish, Dutch, French, Finnish and German languages according to the terminology settings available: SCT_ONLY, UMLS_EXT and LOCAL.

Table 5: Average term coverage values for SCT_ONLY, UMLS_EXT and LOCAL settings and for English, Swedish, Dutch, French, Finnish and German. The confidence interval (CI) is provided with 95% significance.

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

Figure 7: Term coverage for “SCT_ONLY” and “UMLS_EXT” settings and for English, Swedish, Dutch and French. The table shows the average term coverage and the confidence interval with 95%.
7 Methods for semi-automated acquisition of SNOMED CT user interface terms

7.1 Previous work

ASSESS CT task 2.2 built on previous work done within the EU SEMCARE\(^9\) project, following up a study on the semi-automated creation of German user interface terms linked to SNOMED CT. The German language was chosen as the example of the language with the most speakers EU-wide, for which no official SNOMED CT translation exists. Within ASSESS CT, this work was embedded into the new concept of User Interface Terminology, for which it constitutes an ongoing engineering endeavour. Due to the short period of ASSESS CT funding, a final evaluation of this work is not yet possible.

7.2 Term generation principles

The semi-automated term generation methods employed capitalized on three factors:

- **Redundancy of SNOMED CT content.** Many SNOMED CT terms are rather long and can be divided into sub-terms. For example, the sub-term "magnetic resonance imaging" is included in a total of 627 SNOMED CT descriptions such as "Magnetic resonance imaging for simultaneous modulated accelerated radiation therapy planning". Similarly, the term "second degree burn" occurs in 166 terms. This shows an enormous potential for rationalization of localization and translation processes if such sub-terms are identified.

- **Web-based machine translation.** At least for languages with a high representation in the Web and for which many parallel or closely related corpora are available, the quality of machine translations of medical terms has reached a satisfactory level, as a controlled study has shown\(^10\). However, rare terms often remain untranslated, which shows gaps that need to be filled by other technologies. The availability of machine translation results significantly speeds up the translation process, especially when the result is not expected to be a canonical professional SNOMED CT translation according to IHTSDO guidelines, but rather a resource that fulfills the requirements of an user interface terminology.

- **Term similarities.** Medical terms that are formed out of Greek or Latin stems are very similar across languages, e.g., "cryoglobulinemia" vs. "Kryoglobulinämie"; "cardioesophageal" vs. "kardioösophageal", or "chemothalamectomy" vs. "Chemothalamektomie". Based on this observation, a methodology was described which acquires substring substitution rules for single terminology terms\(^11\). Character translation rules were automatically acquired from pairs of English words and their automated translations to German. Using a training set with single words extracted from SNOMED CT as input, a list of 268 translation rules was obtained. The evaluation of these rules improved the translation of 60% of words compared to Google Translate and 55% of translated words that exactly match the right translations. On a subset of words where machine translation had failed, the method improved translation in 56% of cases, with 27% exactly matching the gold standard.

\(^9\) SEMCARE - D4.3 Enriched multilingual biomedical terminology


7.3 Terminology localization workflow

The workflow of terminology translation is described by the following steps, executed by humans (H) and machine processes (M).

1. (M) Single out terms that are identical across languages (e.g. Latin organism names). The hypothesis is that all multiword terms that are identical in the English and Spanish versions do not need to be translated.

2. (M,H) Chunking long terms into constituting fragments (e.g. Burn injury | with | deep necrosis). This requires a human-created set of chunking rules.

3. (M) Substitution of digits by a single character "°". Digits are later re-introduced from the source, which however requires that no inversions occur in the translation process.

4. (M) Creation of word n-gram lists (n = 1 to 6), out of the chunks, ranked by decreasing frequency.

5. (M) Removal of all n-1 grams that fully depend on n-grams (e.g., "collision on takeoff", which has the same incidence as "aircraft collision on takeoff").

6. (M) Submission of frequent n-grams to a web-translation engine, removal of non-translated n-grams for n > 1; identification of non-translated tokens.


8. (M) Application of validated character translation rules on untranslated tokens. Validation of translation candidates (one or more per token) against domain corpus.

9. (M) Import of external translation lists, e.g. Latin anatomy terms.

10. (H) Check and correct multiword n-grams following a list ordered by decreasing frequency. Cut-off point dependent on available resources.

11. (H) Check and correct token translation following a list ordered by decreasing frequency. Cutoff point dependent on available resources. Subtasks:
    - (H) normalize adjectives (male, neuter, singular, positive);
    - (H) identify part of speech (noun, adjective, article, preposition with case);
    - (H) identify gender of nouns;
    - (H) correct misspellings;
    - (H) add common synonyms.

12. (H) formulate re-writing rules (gender, number, case agreement).

13. (M) re-assemble terms.

14. (M) apply re-writing rules.

15. (M) re-introduce digits (same order as in original term).

16. (H) validation of sample. Iterative sample validations lead to iterative cleansing and refinement steps, starting with 10.

The workflow is visualized in Fig. 8
7.4 Implementation

The process was implemented in an MS Access database (Master ACCDB file that imports nine tables from separate ACCDB files). The master file also includes the processing routines in VBA. One ETL-like routine controls the whole workflow from reading the source SNOMED CT description file until the production of the output terminology.

The content of two tables is edited by the user, viz. (1) the language-specific term chunking rules and (2) the n-gram translations (decreasing frequency), mostly originating from machine translated terms, with the following tags, specific to German noun / adjective inflection issues:

- |JJ| Adjective in singular neuter indefinite form;
- |NN|N, |NN|M, |NN|F, |NN|P Noun neuter singular, male singular, female singular, plural;
- |DET|D Definite article, |DET|I Indefinite article (both singular, nominative);
- |PP|G, |PP|D, |PP|A, |PP|DA Preposition that requires Genitive, Dative, Accusative, or either of the latter ones.

A sample is presented in Table 6.

The following rules apply:

- Chunking rules introduce the "\" character as chunk delimiter, the "°" character as a substitute for digits, and the "," character as a decimal separator.
- Chunking must be done in a way that cross-chunk grammatical phenomena are avoided, therefore, e.g., delimiters are put left of prepositions. Chunking rules should not introduce new characters besides the abovementioned ones.
- Only tokens in noun phrases are tagged (here singular neuter indefinite form are required for adjectives), prepositional phrases (already inflected) are not tagged.
- In Latin terms only the leftmost one is tagged with a noun tag containing the gender of the complete term.
- Up to five translations can be provided.
- For single words a translation is required, for phrases with two or more words translations are optional (existing translations can even be removed to avoid garden path effects). Whereas the chunking rules can be adapted to several languages by editing the database table, the word formation rules specific to German are hardwired in the code.
- Overly ambiguous tokens are avoided, especially acronyms and abbreviations. They are, however allowed in multiword terms as long as their meaning is unambiguous.

<table>
<thead>
<tr>
<th>Original n-gram</th>
<th>Count</th>
<th>Translation 1</th>
<th>Translation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary malignant neoplasm</td>
<td>1129</td>
<td>sekundäres</td>
<td>JJ malignes</td>
</tr>
<tr>
<td>assessment scale</td>
<td>1126</td>
<td>Bewertungsskala</td>
<td>NN</td>
</tr>
<tr>
<td>ovary</td>
<td>1125</td>
<td>Eierstock</td>
<td>NN</td>
</tr>
<tr>
<td>primary malignant</td>
<td>1124</td>
<td>primäres</td>
<td>JJ malignes</td>
</tr>
<tr>
<td>Gas</td>
<td>1123</td>
<td>Gas</td>
<td>NN</td>
</tr>
<tr>
<td>ventricle</td>
<td>1121</td>
<td>Kammer</td>
<td>NN</td>
</tr>
<tr>
<td>score</td>
<td>1116</td>
<td>Score</td>
<td>NN</td>
</tr>
<tr>
<td>primary malignant neoplasm</td>
<td>1115</td>
<td>primär bösartiges</td>
<td>JJ Neubildung</td>
</tr>
<tr>
<td>combined</td>
<td>1114</td>
<td>kombiniertes</td>
<td>JJ</td>
</tr>
<tr>
<td>of hip</td>
<td>1113</td>
<td>der Hüfte</td>
<td></td>
</tr>
<tr>
<td>Aneurysm</td>
<td>1113</td>
<td>Aneurysma</td>
<td>NN</td>
</tr>
<tr>
<td>esophagus</td>
<td>1111</td>
<td>Speiseröhre</td>
<td>NN</td>
</tr>
<tr>
<td>bile</td>
<td>1109</td>
<td>Galle</td>
<td>NN</td>
</tr>
<tr>
<td>streptococcus</td>
<td>1109</td>
<td>Streptokokkus</td>
<td>NN</td>
</tr>
<tr>
<td>Visual</td>
<td>1105</td>
<td>visuelles</td>
<td>JJ</td>
</tr>
<tr>
<td>atrioventricular</td>
<td>1101</td>
<td>atrioventrikuläres</td>
<td>JJ</td>
</tr>
<tr>
<td>resection</td>
<td>1099</td>
<td>Resektion</td>
<td>NN</td>
</tr>
<tr>
<td>ring</td>
<td>1091</td>
<td>Ring</td>
<td>NN</td>
</tr>
<tr>
<td>mouth</td>
<td>1086</td>
<td>Mund</td>
<td>NN</td>
</tr>
</tbody>
</table>
8 Results of user interface term acquisition

Raw translations were generated using Google Translate and character translation rules. Whereas all tokens were machine-translated, n-grams for n > 1 were only machine translated if they occurred more than 20 times. The following steps were manually done by a domain expert:

- Validation of all token translations, correction of adjective endings, addition of tags;
- Validation of frequent n-gram translations, removal of infrequent ones that appeared regularly so that they could be reconstructed;
- Addition of new n-gram translations for known homonyms, which would produce wrong translations out of context, such as "delivery", or verb phrases, e.g., beginning with "does" or "does not";
- Iterative analyses of sample translations for error tracking.

The generation procedure generated 1,013,157 German terms out of 519,281 descriptions (terms) from the international versions (only preferred terms and synonyms, no FSNs).

The manual input for content improvement was considerable, roughly estimated as two person months. A sample of 100 terms translated following the ASSESS-CT approach, were compared with the translation using Google Translate by a domain expert. In 17 cases the Google translation was considered better, in 26 cases the ASSESS-CT translation was better, and in 57 cases both translations were comparable.

Using a 5-point Likert scale (5 best, 1 worst), linguistic correctness of the ASSESS CT translation achieved an average rating of 4.61 against 4.45 of the Google translation.

Content fidelity (comprehensibility) was rated 4.52 (ASSESS CT) vs. 4.46% (Google Translate). The differences are statistically not significant.

The random sample is provided in Annex A.

For estimating the costs of this approach, efforts done within the SEMCARE project, ASSESS CT project, as well as adjustments for local purposes at the MUG project partners' site need to be taken into account. A rough estimation amounts to 5 person months, including structure, content and tool production, done by an experienced medical terminologist.
9 Discussion of proposed user interface terminology creation method

The analysis of this sample can be summarised as follows:

- The rule-based ASSESS CT user interface terminology creation approach results in less grammatical errors (wrong number, case);
- Both approaches showed a lower precision in long terms, where Google tends to produce highly ungrammatical and non-understandable results;
- Results point to several technical weaknesses of the ASSESS CT generation approach; the most noteworthy is the problem of translating phrases that contain the preposition "of". So far, they are split into separate chunks (e.g. |Neoplasm| of body | of pancreas), which obviates the generation of common single-noun compounds like "Neoplasie des Pankreaskörpers").
- Both scenarios show poor results for verb phrases.

As important differences between the two approaches must be highlighted

- The ASSESS CT generation process is fully controlled by the user. Errors can be easily fixed.
- Whereas the Google translator produces only one output term per input term, the ASSESS CT generator also generates synonyms. This may result in 10 term variants or more in the case of long source terms.
- Many of these term variants are understandable but uncommon. This does not constitute a problem if this resource is used for NLP, where the size of the lexicon is a negligible factor.
- If a user interface terminology is created as a source of value sets, uncommon term variants should be filtered out. This can be done against a corpus that represents the vocabulary of a domain.
- The approach is not suited to produce FSNs in the target language, since it does not specify preference criteria, nor does it incorporate IHTSDO’s translation guidelines.
- The surprisingly good outcome of Web-based machine translation should be capitalised on. At least for the use in NLP approaches, the advantage may overcome potential negative effects due to bad translations.
- As proposed in SEMCARE D4.3, a broader range of possible (public domain) sources could be integrated into the token and n-gram translation process.

The main outcome of this study is that the proposed iterative semi-automated approach for creation of a SNOMED CT user interface terminology yields good results given a moderate input in human labour (compared to the cost of full professional translation). It also shows the surprising improvement in quality of Web translation engines regarding medical terminology. They can and should be used as an additional source for candidate user interface terms.

However, the creation and maintenance of user interface terminologies cannot depend on a top-down approach only. The generated terms need to be constantly checked against the terminology, and gaps need to be filled. Local terminologies will require cooperation with the user interface terminology users (and users of tools that are powered by the user interface terminology), e.g. by crowdsourcing approaches as proposed in SEMCARE D4.3.
### 10 Annex - Example translations (ASSESS CT method vs. Google translate) of English SNOMED CT Descriptions into German

<table>
<thead>
<tr>
<th>SNOMED CT code</th>
<th>English description</th>
<th>German description (ASSESS CT methodology)</th>
<th>German description (Google translate)</th>
<th>Linguistic correctness 1 min 5 max</th>
<th>Content fidelity 1 min 5 max</th>
<th>Linguistic correctness 1 min 5 max</th>
<th>Content fidelity 1 min 5 max</th>
</tr>
</thead>
<tbody>
<tr>
<td>7099003</td>
<td>Attachment plaque of desmocytoma or hemidesmocytoma</td>
<td>Inserierte-Plaque des Desmokromes oder Hemidesmokrom</td>
<td>Befestigte Plaque von desmocytome oder Hemidesmocytome</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>703529003</td>
<td>Meningeal outflow obstruction</td>
<td>mütterliche Zufriedenheit mit Arbeits-Analysen</td>
<td>Mütterliche Zufriedenheit mit der Arbeit Analysen</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>399017005</td>
<td>Maternal satisfaction with labor analgesia</td>
<td>mütterliche Zufriedenheit mit Arbeit Analysen</td>
<td>mütterliche Zufriedenheit mit Arbeit Analysen</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15300006</td>
<td>Intralobular bile duct obstruction</td>
<td>intralobulärer Ductus cholecodochus</td>
<td>intralobulärer Ductus cholecodochus</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>305180001</td>
<td>Oral form zinc</td>
<td>Oralform Zink</td>
<td>Oralform Zink</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>95044003</td>
<td>Neoplasm of uncertain behavior of short bone of upper limb</td>
<td>Neoplasie zwischen dem Ols breve der oberen Extremität</td>
<td>Neoplasie zwischen dem Ols breve der oberen Extremität</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>396537005</td>
<td>Oxidative megadose injection of vestibul for vestibular distortion</td>
<td>Injektion des Vestibul zur Störung</td>
<td>Injektion des Vestibul zur Störung</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>73819002</td>
<td>Congestion</td>
<td>Ohnmacht</td>
<td>Ohnmacht</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>86175003</td>
<td>368255001</td>
<td>422153005</td>
<td>399751008</td>
<td>359828005</td>
<td>422230003</td>
<td>461619006</td>
<td>396006002</td>
</tr>
</tbody>
</table>

- **ASSESS CT – D2.4**
- **commercial in confidence**
- **Page 26 of 28**
**Jewellery ring**

Schmuck-Ring

5

5

Selektive Herzkathesterisierung des linken Korons für Verlagerung

5

5

Selektive Herzkathesterisierung des linken Herzens für Verlagerung

5

5

Northern Floor Milie

4

5

Northern Floor Milie

4

5

Nordpolische Milie

4

5

Nordpolische Milie

4

5

**Skin of penis**

Haut des Beckens

5

5

Haut des Beckens

5

5

Haut des Beckens

5

5

Haut des Beckens

5

5

**Reconstruction of external ear with distal flap**

Rekonstruktion des Ohres mit distaler Lappen

4

5

Rekonstruktion des Ohres mit distaler Lappen

4

5

Rekonstruktion des Ohres mit distaler Lappen

4

5

Rekonstruktion des Ohres mit distaler Lappen

4

5

**Microcalcifications in tumour and nonneoplastic tissue present**

Mikrokalk in Tumor und nichtneoplastisches Gewebe vorhanden

5

5

Mikrokalk in Tumor und nichtneoplastisches Gewebe vorhanden

5

5

Mikrokalk in Tumor und nichtneoplastisches Gewebe vorhanden

5

5

Mikrokalk in Tumor und nichtneoplastisches Gewebe vorhanden

5

5

**Excision of vertical hernia**

Exzision des vertikalen Hernien

5

5

Exzision des vertikalen Hernien

5

5

Exzision des vertikalen Hernien

5

5

Exzision des vertikalen Hernien

5

5

Exzision des vertikalen Hernien

5

5

Exzision des vertikalen Hernien

5

5
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>pT4a: Tumour invades through cortical bone, into deep (extrinsic) muscle of tongue (genioglossus, hyoglossus, palatoglossus, and styloglossus), maxillary sinus, or skin of face (oral cavity)</th>
<th>pT4a: Geschwulst invadiert durch kortikalen Knochen, in eine tiefe (äußere) Zungenmuskel (M. genioglossus, hyoglossus, palatoglossus, und styloglossus), oder Gesichtsschädel (Cavitas oris)</th>
<th>pT4a: Tumor infiltrates through cortical bone, into deep (extrinsic) muscle of tongue (genioglossus, hyoglossus, palatoglossus, and styloglossus), maxillary sinus, or skin of face (oral cavity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>396579004</td>
<td>pT4a: Tumour invades through cortical bone, into deep (extrinsic) muscle of tongue (genioglossus, hyoglossus, palatoglossus, and styloglossus), maxillary sinus, or skin of face (oral cavity)</td>
<td>pT4a: Geschwulst invadiert durch kortikalen Knochen, in eine tiefe (äußere) Zungenmuskel (M. genioglossus, hyoglossus, palatoglossus, und styloglossus), oder Gesichtsschädel (Cavitas oris)</td>
<td>pT4a: Tumor infiltrates through cortical bone, into deep (extrinsic) muscle of tongue (genioglossus, hyoglossus, palatoglossus, and styloglossus), maxillary sinus, or skin of face (oral cavity)</td>
<td></td>
</tr>
<tr>
<td>396579004</td>
<td>Repair AND revision of stoma of esophagus</td>
<td>Reparatur und Revision eines Stomas des Ösophagus</td>
<td>Reparatur und Revision von Stoma der Speiseröhre</td>
<td></td>
</tr>
<tr>
<td>396579004</td>
<td>Skin of elbow</td>
<td>Haut des Ellenbogens</td>
<td>Haut der Ellenbogen</td>
<td></td>
</tr>
<tr>
<td>396579004</td>
<td>Thoracic spinal ganglion</td>
<td>Brustganglion sensorium nervi spinalis</td>
<td>BWS-Ganglion</td>
<td></td>
</tr>
<tr>
<td>396579004</td>
<td>Fatty tissue of breast</td>
<td>Fettgewebe der Brust</td>
<td>Fettgewebe der Brust</td>
<td></td>
</tr>
</tbody>
</table>

*commercial in confidence*