The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Begoña Altuna
Fondazione Bruno Kessler
University of the Basque Country
altuna@fbk.eu

Alberto Lavelli
Fondazione Bruno Kessler
lavelli@fbk.eu

Manuela Speranza
Fondazione Bruno Kessler
manspera@fbk.eu

Roberto Zanoli
Fondazione Bruno Kessler
zanoli@fbk.eu

Abstract

English. We present the European Clinical Case Corpus (E3C) project, aimed at collecting and annotating a large corpus of clinical cases in five European languages (Italian, English, French, Spanish, and Basque). Project results include: (i) a freely available collection of multilingual clinical cases; and (ii) a two-level annotation scheme based on temporal relations (derived from THYME), whose purpose is to allow the construction of clinical timelines, and taxonomy relations based on medical taxonomies, to be used for semantic reasoning over clinical cases.

1 Introduction

Identifying clinically relevant events and anchoring them to a chronology is very important in clinical information processing, as the ability to access an ordered sequence of events can help to understand the evolution of clinical conditions in patients. However, although interest in information extraction from clinical narratives has increased in recent decades, attention has been focused on clinical entity extraction and classification (Schulz et al., 2020; Grabar et al., 2019; Dreisbach et al., 2019; Luo et al., 2017) rather than on temporal information. If temporal information is extracted from clinical free text, it can be added to structured data collections, e.g. MIMIC III (Johnson et al., 2016), to train clinical prediction systems. Despite some effort on the organization of clinical narratives processing challenges, e.g. CLEF eHealth (Kelly et al., 2019), few shared training and test data sets have been created, and thus developing tools for this task is still difficult.

In fact, the amount of freely available annotated corpora for any of the clinical information extraction tasks has not grown at the same rate as interest in the field, mainly due to patient privacy and data protection issues. In addition, most datasets consist of English texts, which makes research focus on that language.

In an attempt to overcome these problems, we present the European Clinical Case Corpus (E3C)[1], a project aimed at offering a freely available multilingual corpus of semantically annotated clinical narratives. The project will build a 5-language (Italian, English, French, Spanish, and Basque) clinical narrative corpus to allow for the linguistic analysis, benchmarking, and training of information extraction systems. We build upon available resources and collect new data when necessary, with the goal to harmonize current annotations, introduce new annotation layers, and provide baselines for information extraction tasks.

We foresee two types of annotations: (i) temporal information and factuality: events (including attributes expressing factuality-related information), time expressions, and temporal relations according to the THYME standard; and (ii) clinical entities: pathologies, symptoms, procedures, body parts, etc., according to standard clinical taxonomies (e.g. SNOMED-CT[2] (Donnelly, 2006) and ICD-10[3] (WHO, 2015)).

The E3C corpus is organized into three layers, with different purposes:

Layer 1: about 25K tokens per language of clinical narratives with full manual or manually

[1] E3C is a one-year pilot project, started in July 2020. The E3C website is available at https://e3c.fbk.eu. The project is funded by the European Language Grid (ELG), an initiative aimed at developing a cloud platform that provides access to Language Technologies (i.e. running tools and services, data sets and resources) for all European languages.

[2] http://www.snomed.org/

[3] https://icd.who.int/browse10/2019/en

checked annotation of clinical entities, temporal information and factuality, for benchmarking and linguistic analysis.

Layer 2: 50-100K tokens per language of clinical narratives with automatic annotation of clinical entities and manual check of a small sample (about 10%) of this annotation.

Layer 3: about 1M tokens per language of non-annotated medical documents (not necessarily clinical narratives) to be exploited by semi-supervised approaches.

In this paper we present our data collection effort, focused on clinical cases (Section 3), and we describe our annotation scheme (Section 4).

2 Clinical Cases

A clinical case is a statement of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. We focus on clinical cases because they are often de-identified, overcoming privacy issues, and are rich in clinical entities as well as temporal information, which is almost absent in other clinical documents (e.g., radiological reports).

> A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk. Continuous superficial ulcers and spontaneous bleeding were observed under colonoscopy. Subsequent gastroscopy revealed mucosa with diffuse edema, ulcers, errhysis, and granular and friable changes in the stomach and duodenal bulb, which were similar to the appearance of the rectum. After ruling out other possibilities according to a series of examinations, a diagnosis of GDUC was considered. The patient hesitated about intravenous corticosteroids, so he received a standardized treatment with pentasa of 3.2 g/d. After 0.5 mo of treatment, the patient's symptoms achieved complete remission. Follow-up endoscopy and imaging findings showed no evidence of recurrence for 26 mo.

Here we present a sample case extracted from our collection. It is about a patient presenting gastric symptoms (mucopurulent bloody stool and epigastric persistent colic pain), who is finally diagnosed with gastroduodenitis associated with ulcerative colitis (GDUC). To reach the diagnosis, two consecutive medical tests (colonoscopy and gastroscopy) were performed. Treatment (treatment with pentasa of 3.2 g/d), outcome (complete remission) and follow-up (no evidence of recurrence) are also present in the text. Symptoms, tests, observations, treatments and diseases are relevant events for the history of a patient, and it is relevant to place them in chronological order, so as to understand the evolution of the health situation of the patient. For example, we know that the symptoms started 2 weeks prior to the hospital visit, that the colonoscopy was performed before the gastroscopy, that the treatment lasted for half a month and that the patient had no recurrence in the following 26 months.

Since precision in symptom description and diagnosis is utterly important in the clinical field, the clinical findings, body structures, medicines, etc., have to be uniquely identified. This can be done through international coding standards, which allow to assign a unique code to every clinically relevant element in the text.

3 Data Collection

When building the E3C corpus, a big concern has been ensuring its reusability and shareability, which forced us to use anonymised and freely redistributable clinical cases. We deal with three types of clinical narratives: discharge summaries, clinical cases published in journals, and clinical cases from medical training resources. The clinical cases in the E3C corpus contain narratives such as the excerpt presented here.

> 2020-09-01. The patient enters the ER due to abdominal pains. He reports chest pain 5 days ago.

The state of the data collection efforts for the five languages addressed by the project vary depending on their online presence and the number of publications available. For Spanish, a large dataset of clinical narratives and other clinical text collections already exist; for English and French, a significant amount of published material is publicly available. Corpus collection for Italian and Basque, on the other hand, has been more demanding, as we have had to manually extract clinical cases from a number of different sources.

This is shown by the data in Table 1, where we report statistics about the clinical cases col-

Language	Clinical cases	Tokens	Tok./doc
Italian	1,323	73K	55.1
English	9,533	928K	97.2
French	1615	548K	339.1
Spanish	1,400	531K	379.27
Basque	122	26K	214.2

Table 1: Statistics on the clinical cases collected for each language.

Language	Tokens	L1 (25K)	L2 (50K)	L3 (1M)
Italian	13.2M	100%	96%	100%
English	9.7M	100%	100%	100%
French	13.7M	100%	100%	100%
Spanish	1.1M	100%	100%	100%
Basque	74K	100%	2.27%	4.76%

Table 2: Statistics on the layer coverage for each language.

lected so far for each language[4]. The collection of clinical cases has been completed for all languages with respect to Layer 1 and for most languages with respect to Layer 2. Layer 3 of English, French and Spanish is also totally or partially filled with clinical cases.

Italian. The clinical cases come from two main sources, either cases described in public examinations (*test di abilitazione* and *test di specializzazione*) (1276 cases, 56,496 tokens) or clinical cases presented in clinical journals distributed under CC licenses (47 cases, 16,412 tokens). Apart from the clinical cases, we have also collected 8,087 patient information leaflets for medicines (13M tokens).

English. The dataset for English consists of 63,515 abstracts extracted from PubMed with the 'clinical case' query (9.7M tokens). From those, we identified automatically 9,533 clinical case descriptions (928,554 tokens). We first downloaded all abstracts through the PubMed API and then selected only those coming from CC-licensed journals, in order to ensure their redistribution.

French. We used the same strategy to build the French corpus. We downloaded the abstracts from PubMed and selected those from CC-licensed journals. In total, we obtained almost 12,000 abstracts (around 1,5M tokens) out of which we have automatically recognised 199 clinical case descriptions (21,485 tokens). In addition, we have also automatically extracted 1416 clinical cases (547,644 tokens) from CC-BY licensed medical journals. Apart from those, we have also collected circa 8,000 patient information leaflets for medicines (13M tokens).

Spanish. The SPACCC corpus (Intxaurrondo et al., 2018) contains 1000 clinical cases (350,761 tokens) extracted from SciELO[5] and distributed under a CC license. We have also collected an additional dataset of clinical cases extracted from SciELO (400 documents, 180,216 tokens). In addition, two datasets that contain sentences extracted from clinical cases have been added to our corpus: NUBes (518,068 tokens) and IULA+ (38,208 tokens) (Lima López et al., 2020).

Basque. The Basque dataset consists of model discharge summaries (43 documents, 14,239 tokens), clinical cases presented in teaching materials (16 cases, 3,116 tokens), journals and clinical symposia (63 cases, 8,781 tokens) and a dataset of Wikipedia articles on the biomedical domain (47,613 tokens) used in other NLP tasks[6]. Some of the clinical cases are under a CC license, while explicit authorization from the owners has been obtained for the rest.

Taking into account those numbers and the types of documents we have collected for each language, we can say that we have been able to collect enough data to complete Layer 1 in all the languages. For Layer 2, instead, we have only been able to collect enough clinical cases for English, French and Spanish. Reaching the million tokens in Layer 3 is not as complicated as it may seem, as the documents in it do not necessarily need to be clinical cases, although not as many data is available for Basque. The total amount of collected tokens and the layer coverage for each language can be seen in Table 2.

Corpus collection is in a very advanced stage, but new data will be added in the near future. The whole E3C corpus, including core metadata (i.e. language, source, date, length, etc.), will be made available.

3.1 Data Protection in the E3C Corpus

As mentioned, there are two main types of documents in the E3C corpus: clinical narratives and descriptive clinical documents. The latter and

[4]For English and French, the numbers are approximate.

[5]Scientific Electronic Library Online `http://www.scielo.org`

[6]`http://www.statmt.org/wmt20/biomedical-translation-task.html`

even some of the clinical cases (the ones that describe model situations) do not contain any personal data and are out of the scope of data protection regulations. Personal data protection issues, instead, regard the reports that have been written after an actual clinical case. These often contain sensitive patient information and it is the researchers' duty to disseminate them respecting data protection rules (e.g. European Union General Data Protection Regulation) and to address other ethical issues such as achieving informed consent from the patients prior to publication.

All the clinical cases in the E3C corpus have been previously published in other sources, and furthermore, they have been published under licenses that allow redistribution. As a consequence, we consider that all data protection and ethical issues were addressed at the time of first publication and that the documents already comply with the patient data protection policies.

While preparing the E3C dataset, we have also contributed to the protection of personal data, only getting the relevant information for our corpus, responding to the principle of data minimization. For example, many clinical case reports provide illustrative images that have not been considered, as image processing is out of the scope of our project.

In addition, we have also contributed to the reduction of patient traceability, as the article publication date (or an approximate one) has been established as the day the clinical case was written.

4 Annotation Scheme

E3C annotation consists of two levels that provide complementary information. On one hand, annotation of temporal information and factuality follows a mostly language-independent annotation scheme consisting of the THYME guidelines and their extensions (described in more detail in (Speranza and Altuna, 2020)). Annotation and classification of clinical entities, on the other hand, is based on two comprehensive medical taxonomies, SNOMED-CT and ICD-10.

The THYME-driven annotation focuses mainly on clinically relevant events and on the temporal relations between them, with the end goal of coding the information needed to build complete timelines, while the taxonomy-driven annotation provides semantic information and domain-specific knowledge. Looking at the sample clinical case in Section 3, the taxonomy-driven annotation might allow one to infer, for instance, that *abdominal pains* in the first sentence and *chest pain* in the last sentence are closely related, as they are siblings in the hierarchy (in fact, they are both children of [pain of truncal structure] in SNOMED-CT). From the THYME-driven annotation, instead, one might infer the chronological order in which the two events happened.

4.1 THYME-driven Annotation

THYME offers guidelines for the annotation of clinically relevant events, time expressions and the relations between them.

Events are all actions, states, and circumstances that are relevant to the clinical history of a patient (for example, we have pathologies and symptoms such as *pain*, but also more general events such as *enters*, *reports*, and *continue*). The annotation of events also includes a number of attributes, some of which focus on factuality-related information (the contextual modality attribute, for instance, is used to mark non-factual, either generic or hypothetical, events).

Time expressions are all references to time, such as dates (both absolute like *2020-09-01* and relative like *5 days ago*), intervals (*last three days*), etc.

THYME also provides guidelines for the annotation of relations between events and/or time expressions. By expressing precedence, overlap, containment, initiation or ending between two events and/or time expressions, **TLINKs** allow for chronologically ordering them. **ALINKs** are relations that link aspectual events, i.e. events indicating a specific phase (beginning, end, continuation, etc.) of an event, to the event itself.

To obtain annotations that will allow more descriptive timelines, we have expanded the THYME annotation scheme.

Anatomical parts are not annotated in THYME even if noun phrases whose head is a body part can be clinically very relevant (as in *He had a swollen eye*). To annotate them, we have created the new **BODY PART** tag. In addition, a new **ACTOR** tag is used to mark the actors (patients, health professionals, etc.) mentioned in the narratives. Finally, **RML** is a tag we have created to mark test results, results of laboratory analyses, formulaic measurements, and measure values (which are not marked in THYME), as we think that they offer relevant insights into the health status of a patient.

Table 3 represents the annotated version of the clinical case in Section 3. The first column contains the original text (one token per line). The second column shows the span of the THYME-driven annotated elements (specifically, examples of time expressions, actors, events, and body parts) in the IOB2 format, where B-LABEL marks the first token of an element of type LABEL, I-LABEL is used for the subsequent tokens (if any), and O is used for tokens that do not belong to an annotated element. The last two columns represent the taxonomy-driven annotation (see below).

4.2 Taxonomy-driven Annotation

Clinical coding is widely spread in clinical practice; either doctors add the codes for findings, procedures, treatments, etc. to the patients' clinical histories, or large amounts of raw clinical data are automatically coded for the development of clinical prediction systems. The coded concepts are hierarchically classified in taxonomies such as SNOMED-CT and ICD-10.

SNOMED-CT is considered to be the most comprehensive clinical healthcare taxonomy, and is available for most of the languages of the E3C project, i.e. English, French, Spanish, and Basque. There is a validated SNOMED-CT version for the first three languages, while for Basque a partial version has been used (Perez de Viñaspre and Oronoz, 2015). SNOMED-CT offers 19 main categories (and a wide set of subcategories) that range from clinical findings and body structures to social contexts. On the other hand, ICD-10 (International Classification of Diseases, 10th revision) is a classification of diagnoses and procedures. The diseases are classified in 22 categories.

Taxonomy-driven annotation consists of marking in the texts all mentions of clinical entities and mapping them to a code from both international standards.

Table 3 represents the annotated version of the clinical case in Section 3. The third and forth columns show the span of the annotated clinical entities in the IOB2 format, with respect to SNOMED-CT and ICD-10 respectively.

The taxonomy-driven annotation is based, for each concept, on the specific linguistic realization that is coded in the taxonomy, whereas in texts we can find a number of different textual realizations of the same concept. Variability may relate to the alternation between singular and plural and between similar prepositions, or to the presence/omission of a preposition or article. In E3C we have devised a set of rules to account for the variability of linguistic expressions. For instance, looking at the excerpt in Section 3, the textual realization *abdominal pains* is associated with the singular SNOMED-CT concept [abdominal pain]. In addition, if overlapping portions of text match different concepts, we select the most specific one; for instance, [chest pain] is preferred over [pain].

The E3C guidelines for taxonomy-driven annotation are based on both the ShARe (Elhadad et al., 2012) and the ASSESS CT annotation guidelines[7] (Miñarro-Giménez et al., 2018).

4.3 Language-dependent Decisions

Semantic annotation of the E3C corpus is largely language-independent. However, as we are dealing with morpho-syntactically diverse languages, we have added additional annotation guidelines for each language. These guidelines respond mainly to the annotation of the extent of the temporal and clinical entities, since their semantic features are not altered by the morpho-syntactic features.

Both the THYME-driven and the taxonomy-driven annotation schemes were originally developed for English, a language whose morphology is not particularly rich compared to the other languages of the E3C corpus (especially the Basque language). For all these, it was therefore necessary to define language specific guidelines handling the annotation of semantically complex tokens resulting from the combination of different elements (e.g., a preposition and an article)[8].

In the case of romance languages (Italian, French and Spanish), we have taken decisions on the annotation of preposition+article contractions. The article may be part of the extent of time expressions, RML, actors and body parts, whereas the preposition should not be included. When a contraction is present, though, we have decided to capture it inside the extent (1–3).

(1) [Nel condotto uditivo esterno] si evidenziava una lesione. (*[In the external ear canal] an injury was observed.*)

[7]The ASSESS CT annotation guidelines can be found at `https://user.medunigraz.at/jose.minarro-gimenez/docs/assessct/AnnotationGuidelines.pdf`

[8]It is to be remembered that the annotations in the E3C corpus are performed at token-level.

	THYME	Taxonomy	
		SNOMED-CT	ICD-10
2020-09-01	B-TIMEX3	0	O
The	B-ACTOR	O	O
patient	I-ACTOR	O	O
enters	B-EVENT	O	O
the	O	O	O
ER	O	O	O
due	O	O	O
to	O	O	O
abdominal	O	B-ENTITY	B-ENTITY
pains	B-EVENT	I-ENTITY	I-ENTITY
.	O	O	O
He	B-ACTOR	O	O
reports	B-EVENT	O	O
chest	B-BPART	B-ENTITY	B-ENTITY
pain	B-EVENT	I-ENTITY	I-ENTITY
5	B-TIMEX3	O	O
days	I-TIMEX3	O	O
ago	I-TIMEX3	O	O
.	O	O	O

Table 3: Annotation of the excerpt in Section 3 in IOB2 format.

(2) Nous recommandons un suivi [des malades guéris du COVID-19]. (*We recommend a follow up [of the patients cured from Covid-19].*)

(3) El drenaje [del flanco izquierdo] se retiró [al día 16]. (*The drainage [of the left side] was withdrawn [at day 16].*)

Basque, on the other hand, is a highly agglutinative language in which information expressed by prepositions in Indo-european languages is expressed by postpositions. Most of those postpositions appear attached to the nouns, adjectives, verbs and adverbs they refer to, while there is also a small set of free postpositions. The attached postpositions are taken inside the extent of the tags in E3C (4), while the free postpositions are left unmarked (5).

(4) [Larriagotzeetan] infekzio estreptokozikoa izana zuen. (*[In the worsenings] s/he had also had streptococcal infections.*)

(5) Tik motoreak zeuzkan, [bizpahiru urtez] **geroztik**. (*S/he had motor tics,* ***after*** *[two-three years].*)

4.4 Discussion

The two annotation levels can be mapped to address specific tasks, or to develop applications that need to exploit both. Within the E3C project, we are exploring the main issues that emerge when trying to exploit the two annotation levels at the same time. Our future aim within the project is to select a specific task and implement a mapping tailored to that task.

The main mapping issue is determined by non-matching annotated spans. Given that more specific (typically longer) taxonomy concepts are preferred to more generic ones, and that in THYME only the syntactic head of events is marked, in many cases the span of the concept is longer than the span of the event. Compare, for example, the SNOMED-CT concept associated with *abdominal pains* and the THYME event *pains* in Table 3.

More interestingly, in some cases, we can have two separate THYME annotations within the span of a single taxonomic concept. Back to our example, the SNOMED-CT concept [chest pain] overlaps with the two separate THYME annotations *pain* and *chest*.

Another issue is the inevitably different classification criteria in medical taxonomies and THYME. For instance, only a minimal part of what is marked as an event in THYME is a child concept of [event] in SNOMED-ct (e.g., *abuse* and *death*); in most cases what is marked as an event in THYME belongs to a different subpart of the SNOMED-CT hierarchy (for instance, *pain* is part of the [finding] subhierarchy, not of [event]).

5 Conclusions and Future Work

We presented the E3C project, which aims to become a reference European corpus of annotated clinical cases. We focused on two initial achievements: (i) a freely available collection of clinical cases in five languages; and (ii) a comprehensive annotation scheme based both on temporal information and on medical taxonomies.

Our next steps include the extensive manual annotation of the clinical cases in all five languages, and the definition of tasks and baselines on top of the annotated data, taking advantage of neural models derived from training data. More specifically, we plan to target the automatic construction of clinical timelines and question answering over clinical cases.

Acknowledgements

This work has been partially funded by the ELG project (EU grant no. 825627) and by the Basque Government post-doctoral grant POS_2019_1_0030.

References

Kevin Donnelly. 2006. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. In Lodewijk Bos, Laura Roa, Kanagasingam Yogesan, Brian O'Connell, Andy Marsh, and Bernd Blobel, editors, *Medical and Care Compunetics 3*, volume 121 of *Studies in Health Technology and Informatics*, chapter 31, pages 279–290. IOS Press, Amsterdam The Netherlands.

Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, 125:37–46.

Noémie Elhadad, Guergana Savova, Wendy Chapman, Glenn Zaramba, David Harris, and Amy Vogel. 2012. ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes. Technical report, Columbia University.

Natalia Grabar, Cyril Grouin, Thierry Hamon, and Vincent Claveau. 2019. Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019. In *Actes du Défi Fouille de Textes 2019*, pages 7–16, Toulouse, France. Actes DEFT 2019.

Ander Intxaurrondo, Montserrat Marimón, Aitor González-Agirre, José Antonio López-Martín, Heidy Rodríguez, Jesús Santamaría, Marta Villegas, and Martin Krallinger. 2018. Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 280–289, Seville, Spain. Spanish Society for Natural Language Processing.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Liadh Kelly, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 322–339, Cham. Springer International Publishing.

Salvador Lima López, Naiara Pérez, Montse Cuadros, and German Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France, May. European Language Resources Association.

Yuan Luo, William K. Thompson, Timothy M. Herr, Zexian Zeng, Mark A. Berendsen, Siddhartha R. Jonnalagadda, Matthew B. Carson, and Justin Starren. 2017. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety*, 40:1075–1089.

José Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. 2018. Qualitative analysis of manual annotations of clinical text with SNOMED CT. *PLoS ONE*, 13(12).

Olatz Perez de Viñaspre and Maite Oronoz. 2015. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC medical informatics and decision making*, 15 Suppl 2.

Sarah Schulz, Jurica Ševa, Samuel Rodríguez, Malte Ostendorff, and Georg Rehm. 2020. Named Entities in Medical Case Reports: Corpus and Experiments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4495–4500, Marseille, France. European Language Resources Association.

Manuela Speranza and Begoña Altuna. 2020. E3C annotation guidelines. Technical report, Fondazione Bruno Kessler.

World Health Organization WHO. 2015. *International statistical classification of diseases and related health problems*. World Health Organization, 10th revision, fifth edition, 2016 edition.