E3C: European Clinical Case Corpus

Bernardo Magnini, Begoña Altuna, Alberto Lavelli,
Manuela Speranza and Roberto Zanoli

{magnini;altuna;lavelli;manspera;zanoli}@fbk.eu
Fondazione Bruno Kessler (FBK) - Trento, Italy

European Clinical Case Corpus

- **EUROPEAN**: E3C consists of clinical cases in five European languages: Italian, English, Spanish, French, and Basque.

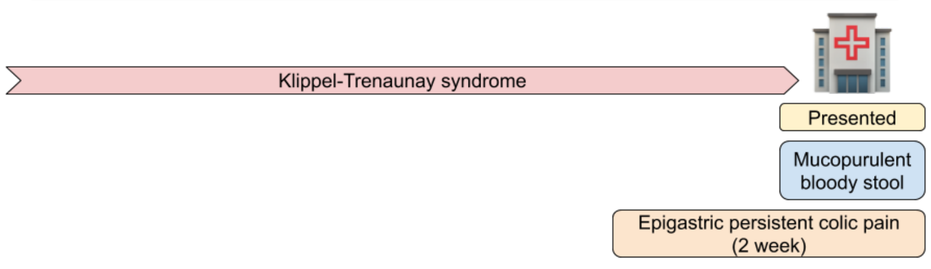
Hello! Ciao! ¡Hola! Salut! Kaixo!

- **CLINICAL CASE**: Clinical cases are statements of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the situation of a patient. They are rich in clinical entities as well as temporal information.
- **CORPUS**: E3C will be annotated with clinical entities (e.g. symptoms and pathologies), temporal information, and factuality. It will be widely available and re-usable as we built it on top of resources that are distributed under public copyright licenses.

Sample clinical case and annotation

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.

	THYME	Taxonomy	
		SNOMED-CT	ICD-10
A	B-ACTOR	O	O
25-year-old	I-ACTOR	O	O
man	I-ACTOR	O	O
with	O	O	O
a	O	O	O
history	O	O	O
of	O	O	O
Klippel-Trenaunay	O	B-ENTITY	B-ENTITY
syndrome	B-EVENT	I-ENTITY	I-ENTITY
presented	B-EVENT	O	O
to	O	O	O
the	O	O	O
hospital	O	O	O
with	O	O	O
mucopurulent	O	B-ENTITY	O
bloody	O	B-ENTITY	I-ENTITY
stool	B-EVENT	I-ENTITY	I-ENTITY
and	0	O	O
epigastric	O	B-ENTITY	B-ENTITY
persistent	O	I-ENTITY	I-ENTITY
colic	O	I-ENTITY	I-ENTITY
pain	B-EVENT	I-ENTITY	I-ENTITY
for	O	O	O
2	B-TIMEX3	O	O
wk	I-TIMEX3	O	O
.	O	O	O



Motivation

- Due to patient data protection issues, the availability of datasets of clinical cases is very limited.
- Most of the available datasets are in English.

Goals

- Stimulating researchers to work on clinical data on a shared benchmark in a multilingual setting.
- Fostering task and technology development, including applications for metadata tagging and to support clinical predictions.
- Boosting the clinical NLP scene for languages other than English.

Data collection

Data collection process:

1. **Source** identification (PubMed, journals, teaching materials, medical examinations, etc.) and **license** verification (CC-BY(-NC)(-SA) licenses).
2. Clinical case **extraction** (identification of the text section that describes the clinical case)
3. Clinical case **metadata** collection

Clinical cases for each language:

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

Statistics on the layer coverage 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%

A two-level annotation scheme

- **temporal information** and **factuality** (based on THYME [1])
  - EVENT: events
  - TIMEX3: time expressions
  - TLINK: temporal relations
  - ALINK: aspectual relations
  - RML: results, measurements and lab test results
  - ACTOR: actors
  - BODYPART: body parts
- **clinical entity taxonomies** (based on ShARe [2] and ASSESS-CT [3])

SNOMED CT The global language of healthcare ICD-10

Outcomes

- A freely available collection of multilingual clinical cases
  - **Layer 1: about 25K tokens per language** of clinical narratives with full manual or manually 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.
- Baseline systems for automatic annotation of temporal and factuality information and clinical entities

References

[1] William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014.

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

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