The Quaero French medical corpus: a resource for medical entity recognition and normalization

Aurélie Névéol [1] Cyril Grouin [1] Jeremy Leixa [2]
Sophie Rosset [1] Pierre Zweigenbaum [1]

[1]LIMSI–CNRS [2]ELDA

31st May 2014

Context and objectives

Observations — biomedical corpora are useful

- **Vast amount of information** in the biomedical domain available as **free texts**
- Text processing tools and methods require **training corpora**
- **No/few corpora available** for French

Objective — development of French corpus

- Development of an **extensive corpus** of biomedical documents
- Annotation at **mention** and **concept** levels based upon automatic pre-annotation
- Annotated corpus freely available to the scientific community

Corpora

Types — different genres of biomedical documents

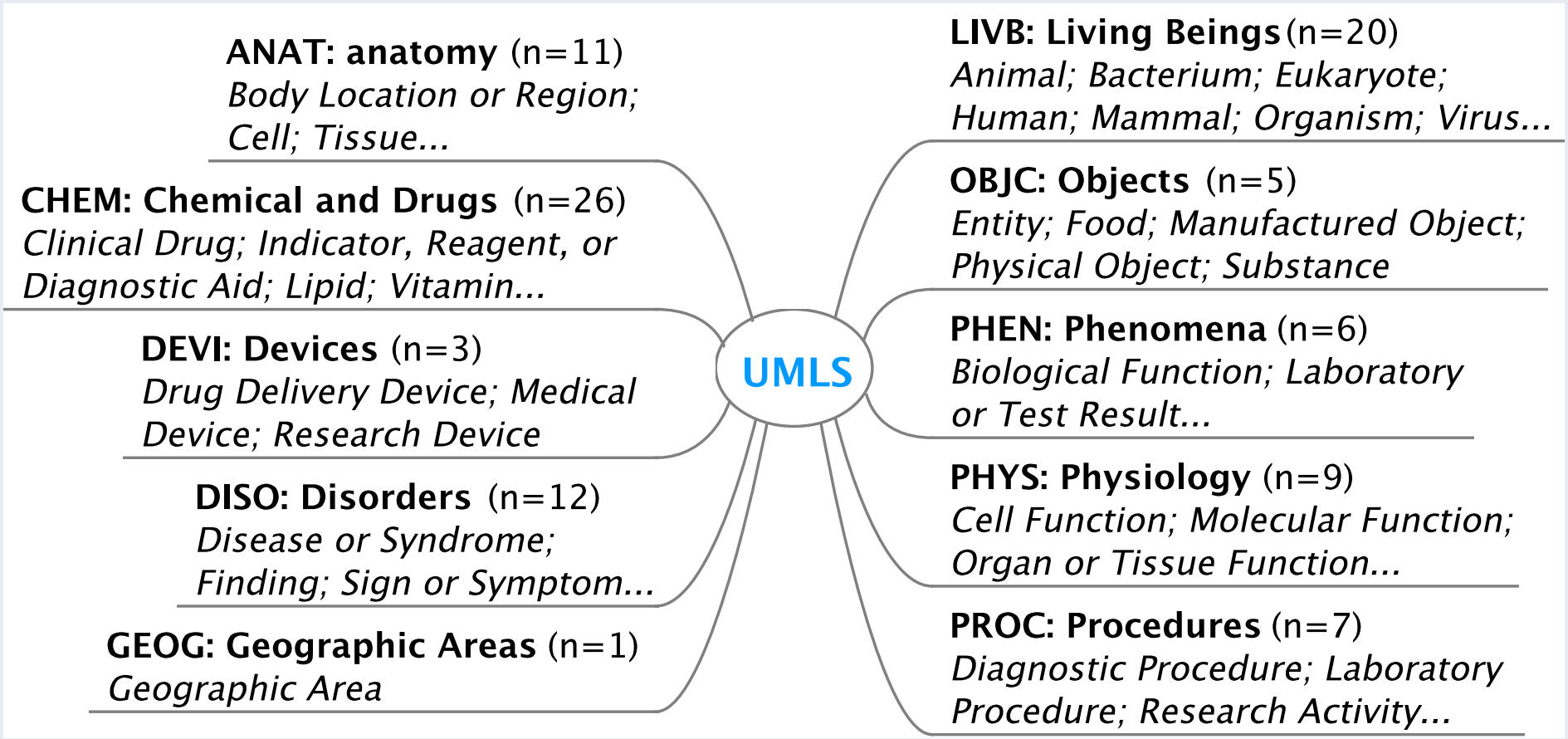
- **EMEA** (European Medicines Agency), information on marketed drugs: 13 documents
- **MEDLINE**: 2,500 titles of research papers
- **EPO** (European Patent Office): 25 full patents containing keyword *"maladie"* (disease) or *"médicament "* (drug)

Choices — ensure compatibility and coverage

- Sources recently used in cross-lingual medical NER challenge: **CLEF-ER 2013**
- Documents available in at least one language other than French
  - EMEA: several european languages
  - MEDLINE: English
  - EPO: English & German

Annotation schema

UMLS Semantic Groups & Semantic Types



Principles

Comprehensive annotations

- Annotate all relevant Semantic Groups
  - *prévention des* PHENOMENON C0034897 [DISORDER C2825055 *récidives*]

    recurrence prevention
- Annotate all relevant concepts within the same Semantic Group
  - *patients* [DISORDER C0564408 DISORDER C0338831 *maniaques*]

    obsessive patients
- Annotate all entity mentions, including embedded mentions
  - [DISORDER C0027051 *infarctus du* ANATOMY C0027061 *myocarde*]

    myocardial infarction
- Annotate discontinuous entities separately

  [DISORDER C0678236 DISORDER C0008679 *maladies*] *rares* *et* [DISORDER C0008679 *chroniques*]

  chronic and rare diseases

Principles

Annotation process

For each entity mention:

- does a **corresponding concept** exist in the UMLS Metathesaurus?
- if found, is the associated **category** listed in the **annotation manual**?
- if found, report the **associated CUI**;
- create a **complete annotation** for the entity mention:
  - all relevant semantic groups
  - all relevant CUIs
  - all embedded entities
  - all discontinuous entities

Pre-annotated vs. annotated excerpt

Document	*Facteurs de croissance et cancers intestinaux.* "Growth factors and intestinal cancers."
Pre-annotated document	*Facteurs de croissance et* <DISO CUI="C0346627"> *cancers intestinaux.* </DISO>
Annotated document	<CHEM CUI="C0018284"> *Facteurs de* <PHYS CUI="C18270"> *croissance* </PHYS> </CHEM> *et* <DISO CUI="C0346627"> <DISO CUI="C0027651"> *cancers* </DISO> <ANAT CUI="C0021853"> *intestinaux.* </ANAT> </DISO>

Corpus statistics

Overview of the Quaero medical corpus

	EMEA	MEDLINE	EPO	All
Tokens	58,874	23,647	20,537	103,057
Pre-annotations				
Entities (all)	7,280	1,692	1,662	10,634
Entities (unique)	1,672	1,194	305	3,009
CUIs (all)	12,098	3,207	2,211	17,516
CUIs (unique)	1,653	1,879	378	3,325
Final corpus				
Entities (all)	12,761	8,781	4,865	26,407
Entities (unique)	2,839	5,600	960	8,460
CUIs (all)	12,647	8,767	4,867	26,281
CUIs (unique)	1,807	4,156	759	5,796

Pre-annotation benefit

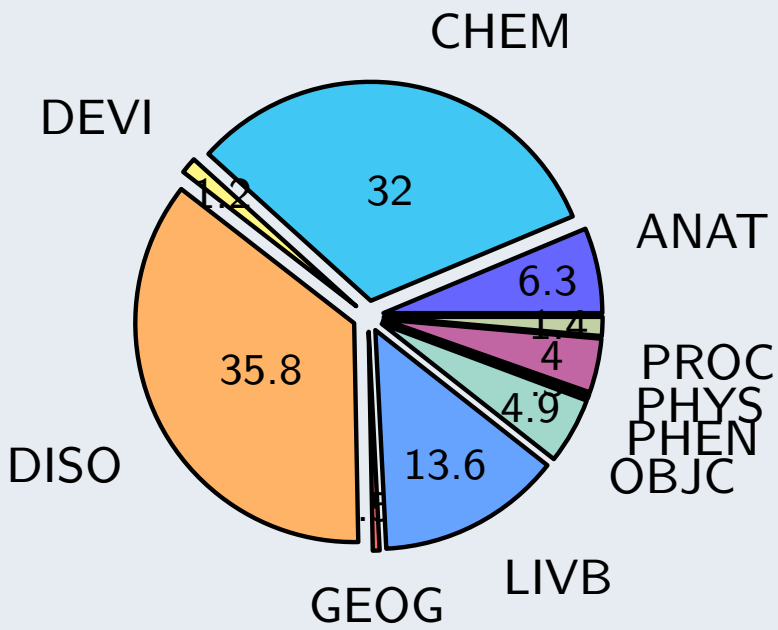
Pre-annotation yielded high precision and medium-to-low recall

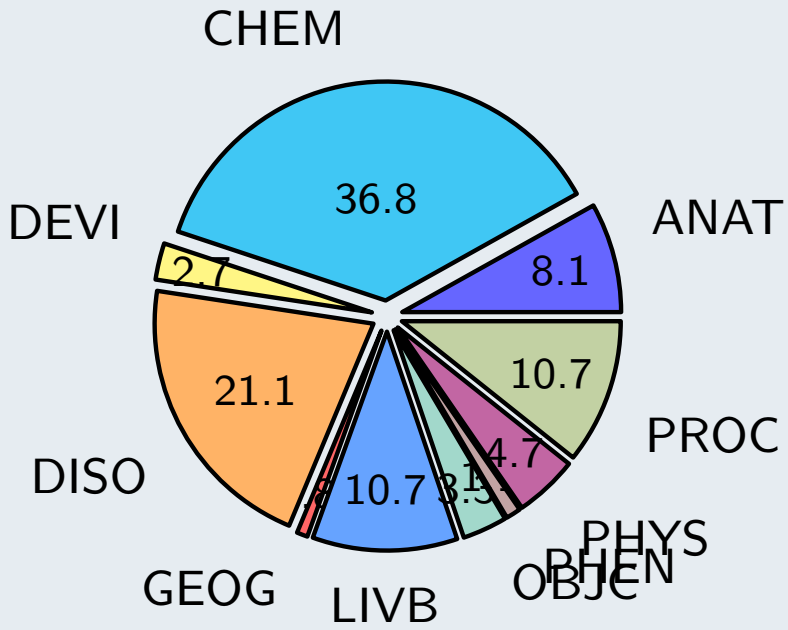
	EMEA	MEDLINE	EPO
Kappa	0.361	0.142	0.187
F-measure	0.595	0.303	0.428
Precision	0.831	0.937	0.841
Recall	0.463	0.181	0.287
Correct (TP) #	5,906	1,585	1,398
Inserted (FN) #	6,261	7,123	3,394
Deleted (FP) #	610	38	192
Substituted #	588	69	72

Table: From pre-annotated to final corpus

Annotation coverage (%)

EMEA corpus (7,280 entities) chemical and disorder annotations





Pre-annotated corpus

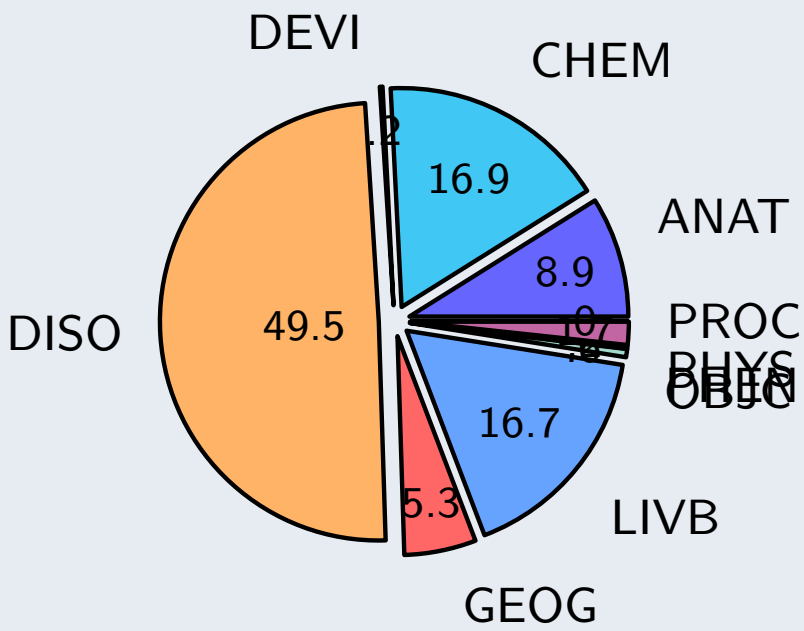
- high number of CHEM and DISO

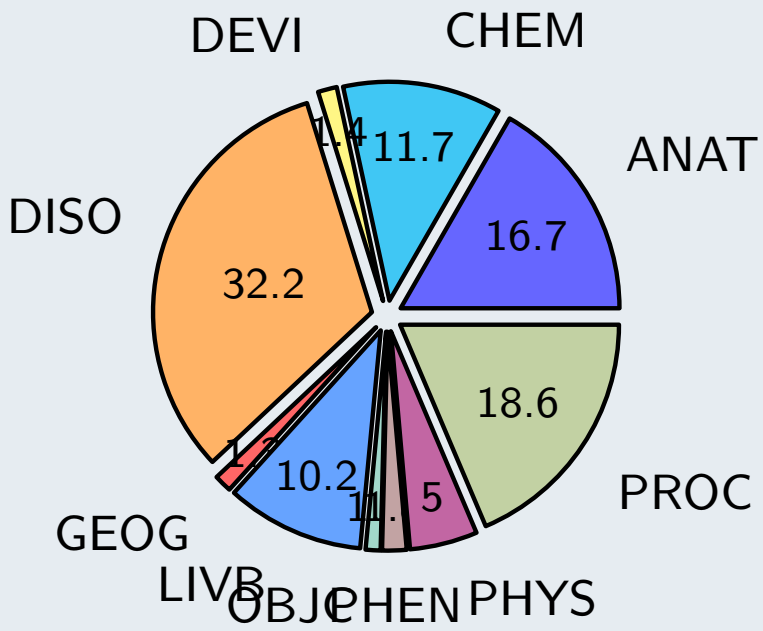
Final corpus

- high number of CHEM
- DISO and LIVB decreased

Annotation coverage (%)

MEDLINE corpus (1,692 entities) — disorder annotations





Pre-annotated corpus

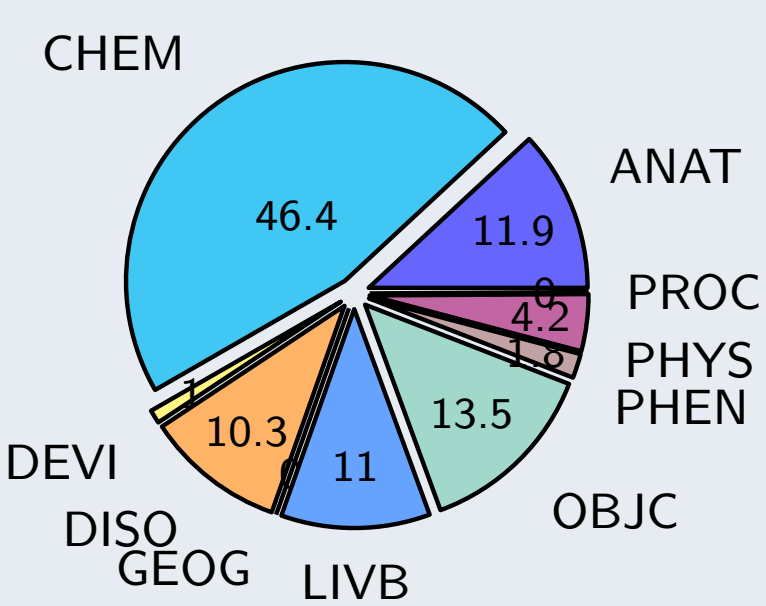
- huge number of DISO

Final corpus

- high number of DISO
- CHEM and LIVB decreased
- PROC and ANAT increased

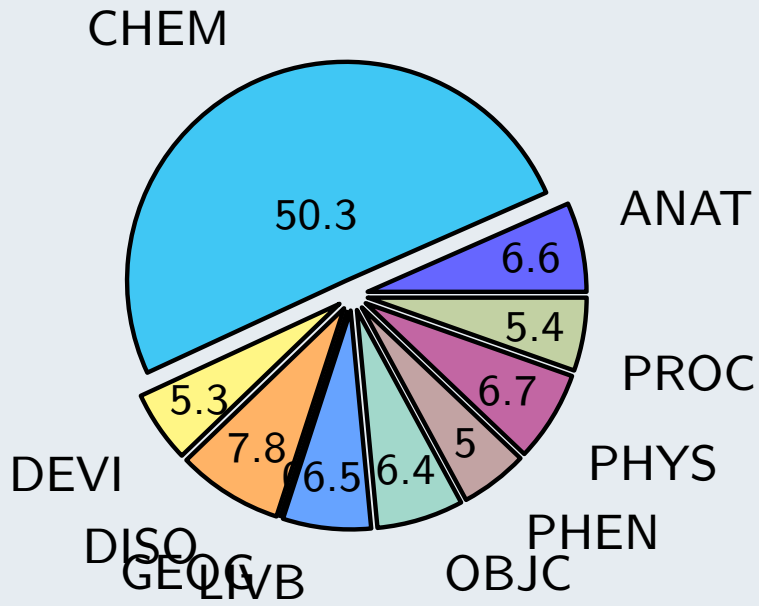
Annotation coverage (%)

EPO corpus (1,662 entities) — chemical annotations



Pre-annotated corpus

- huge number of CHEM



Final corpus

- huge number of CHEM
- well balanced for other categories

Inter-annotator agreement

MEDLINE

- IAA computed on three random samples of 100 pre-annotated titles at three different time points in the annotation process
  - **Set 1**: novice annotation vs. expert annotation
  - **Sets 2 and 3**: novice annotation vs. revised by the expert
- Agreement (F-measure) based on exact match:

	Set 1	Set 2	Set 3
Entities	0.77	0.92	0.90
CUIs	0.66	0.91	0.91

- Set 1: IAA helped **identify differences** in guideline interpretation $\rightarrow$ consensus decisions were applied to the remainder of the corpus
- Sets 2 and 3: IAA shows that differences were **resolved suitably** and **high quality annotations** were produced

What now?

- Corpus preparation for release: **BioC format**
- Material for biomedical **entity recognition** and **normalization** for French
- Material for a **shared task** in French (annotations similar to CLEF ER Challenge)

Thank you!

Acknowledgements

ELDA annotators

Funding: Quaero

Funding: Cabernet