ACTER 1.5

Annotated Corpora for Term Extraction Research

- Ayla Rigouts Terryn, Veronique Hoste, Els Lefever

- CLARIN, Prague, 10/10/2022

language and translation technology team

UNIVERSITEIT GENT

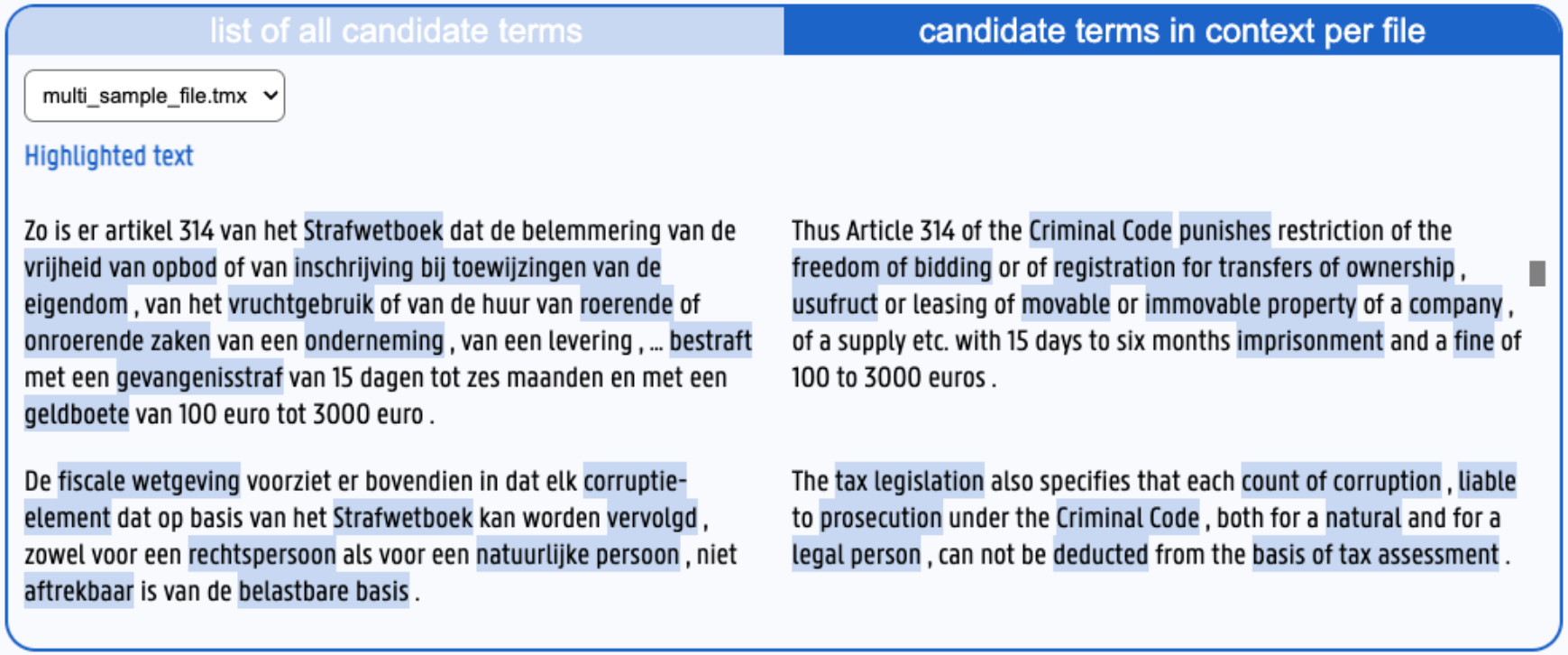
Terminology

= specialised vocabulary that expresses domain-specific concepts through single- or multi-word terms

translators spend about 20% to 60% of their work on “terminology activities” (Champagne, 2004)

Guy Champagne. 2004. The Economic Value of Terminology: An Exploratory Study. Technical report, submitted to the Translation Bureau of Canada.

Automatic Term Extraction



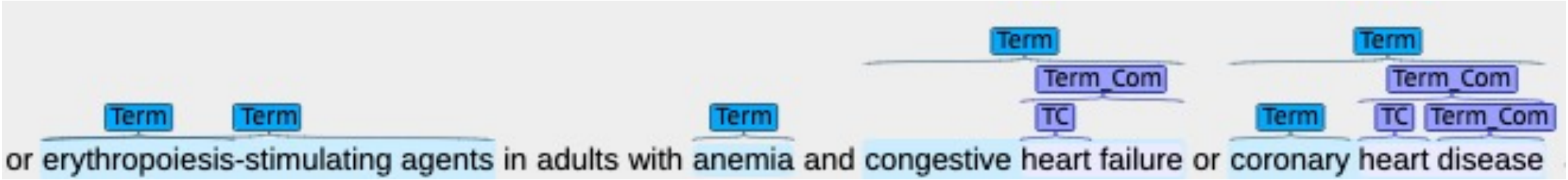
Machine Learning for ATE

Recent evolution from mainly rule-based methodologies, to more (supervised) machine learning

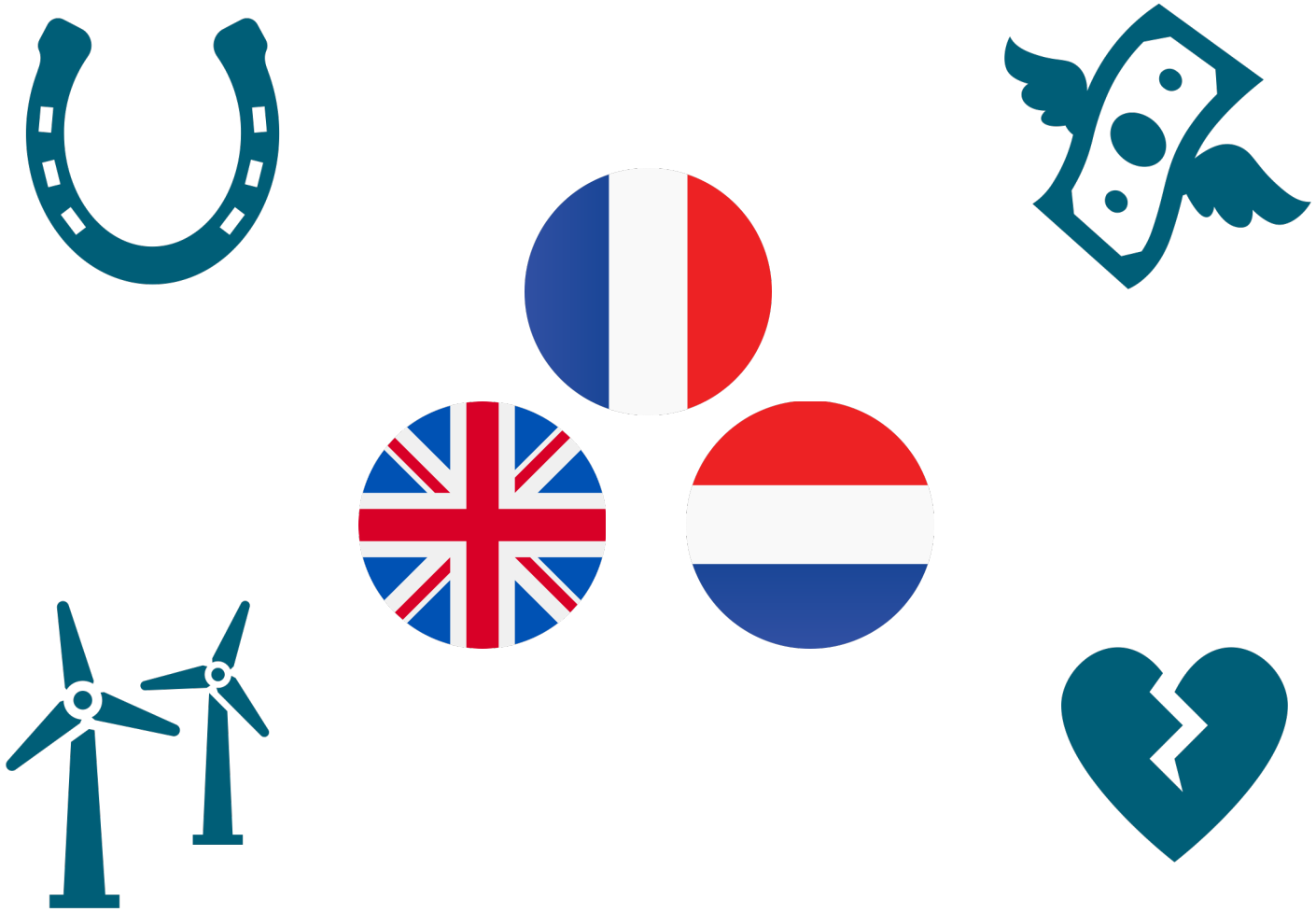
ACTER

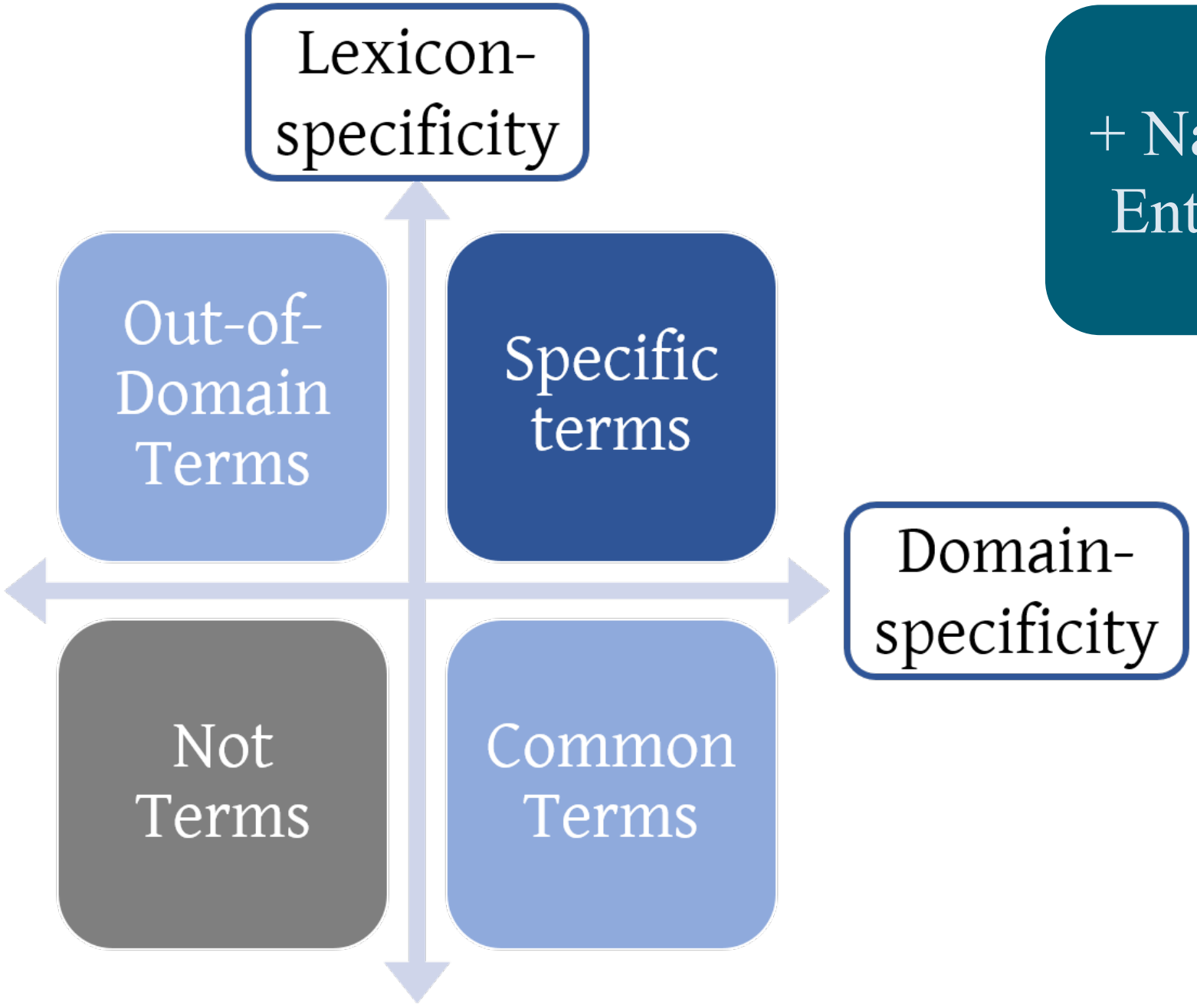
= **A**nnotated **C**orpora for **T**erm **E**xtraction **R**esearch

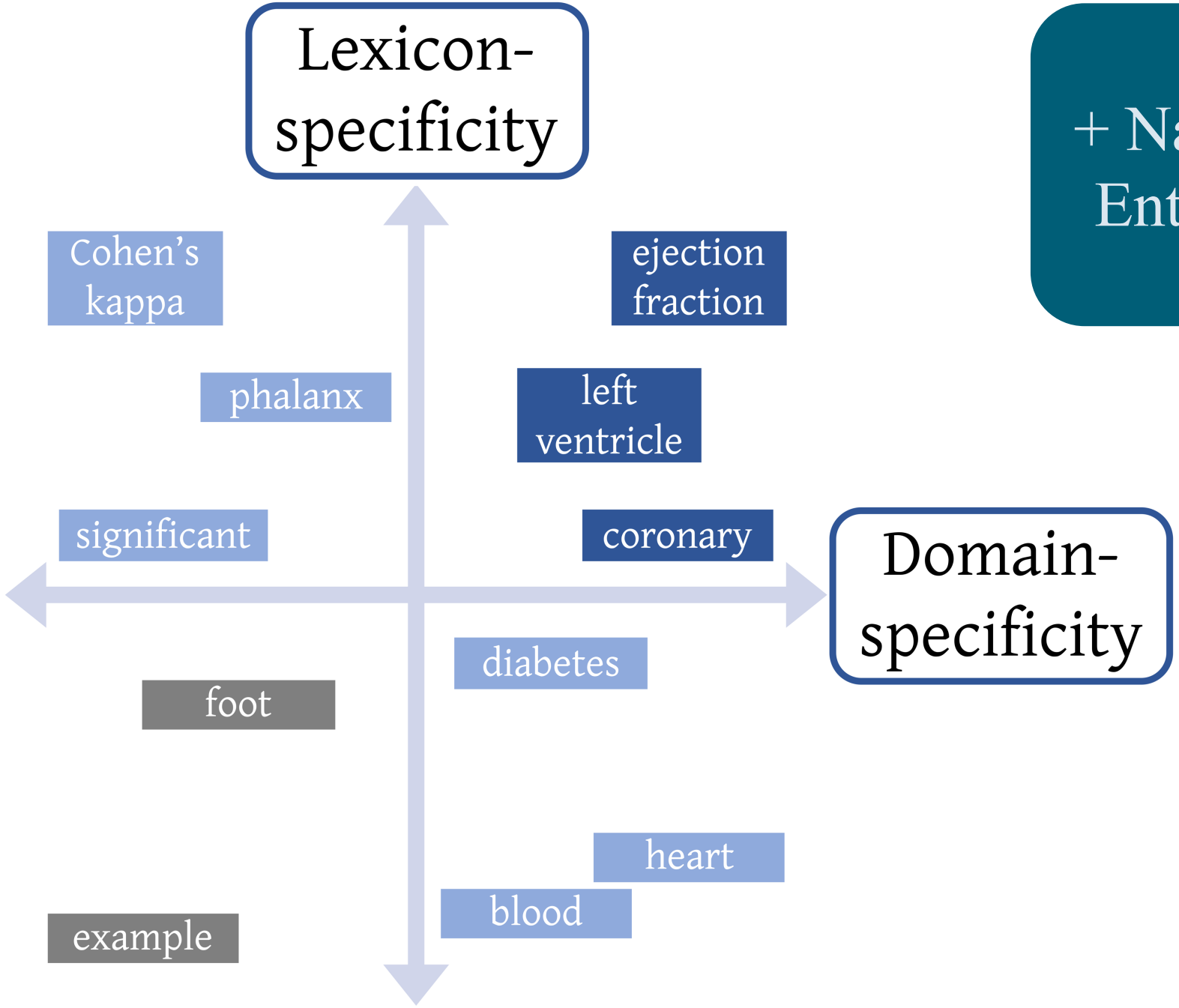
- Annotate based on human intuition, not computer capabilities
- Annotation guidelines, validated through IAA studies



4 domains, 3 languages

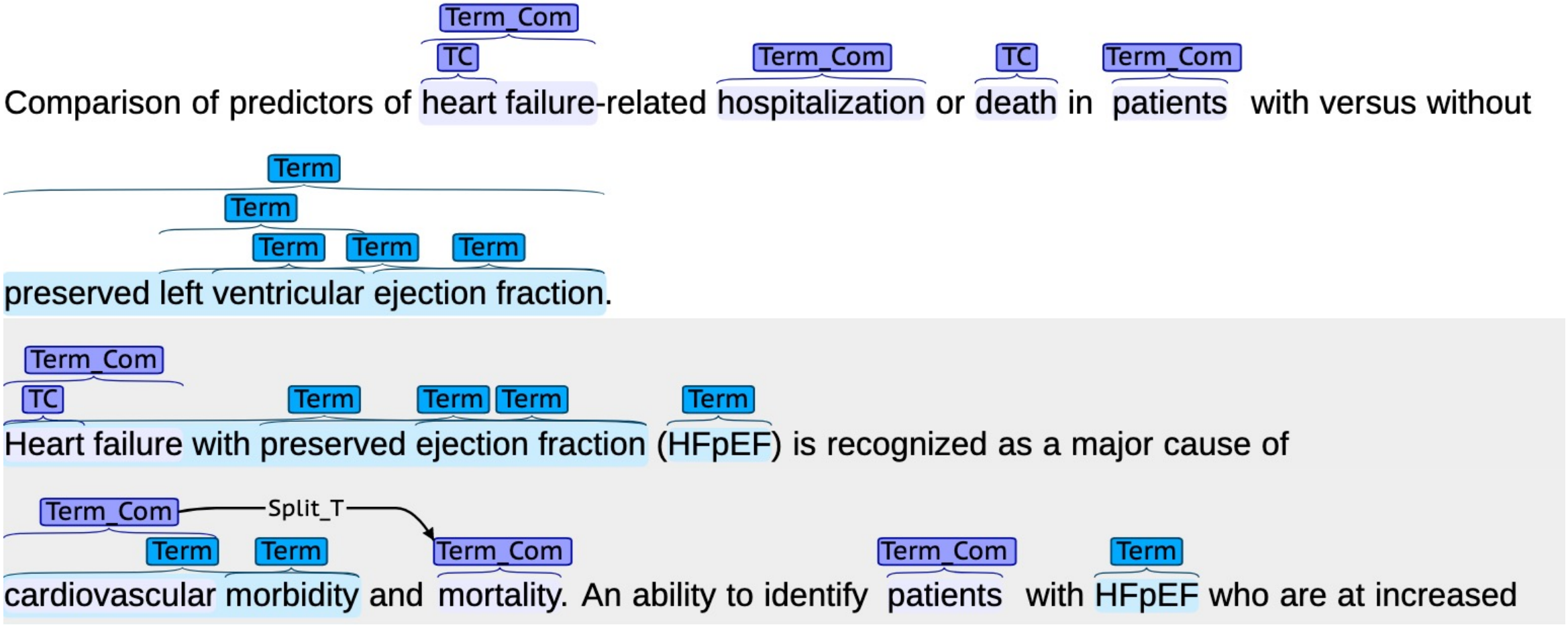



Lexicon-specificity
Out-of-Domain Terms
Specific terms
Not Terms
Common Terms
Domain-specificity
+ Named Entities



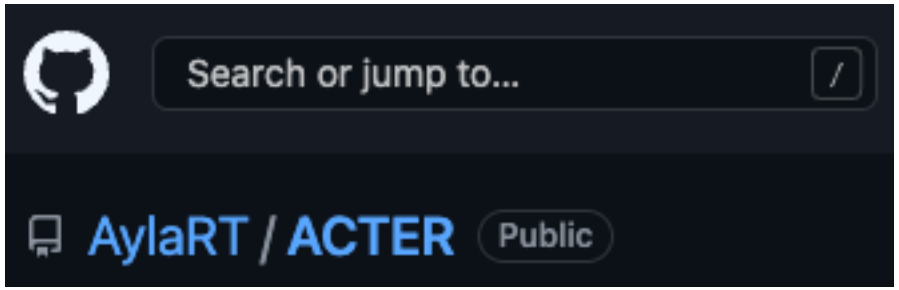
+ Named Entities

BRAT Rapid Annotation Tool



Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.-D., & Tsujii, J. (2011). BioNLP Shared Task 2011: Supporting Resources. *Proceedings of BioNLP Shared Task 2011 Workshop*, 112–120.

ACTER



- ACTER 1.0: first version
- ACTER 1.1: TermEval shared task @CompuTerm2020
- ACTER 1.2: incl. test set for TermEval + labels
- ACTER 1.3: general improvements + Github repo
- ACTER 1.4: better normalisation
- **ACTER 1.5: improve normalisation, documentation + add sequential annotations**

Improve annotations

ACTER

Unique annotation lists (incl. and excl. NEs), tokenised or not, with labels, as tsv

```
6mwt    Specific_Term
8-fluo-camp Specific_Term
aa  Named_Entity
aac Specific_Term
aas Named_Entity
abcc2   Specific_Term
abcd classification Specific_Term
abdominal   Common_Term
abdominal aortic constriction   Specific_Term
ablation    Specific_Term
absorptiometry  Specific_Term
```

ACTER

Sequential annotations (incl. and excl. NEs), IOB or IO, as tsv

```
Remote  B
ischemic    I
conditioning    I
for O
patients    B
with    O
heart   B
failure I
?   O
```

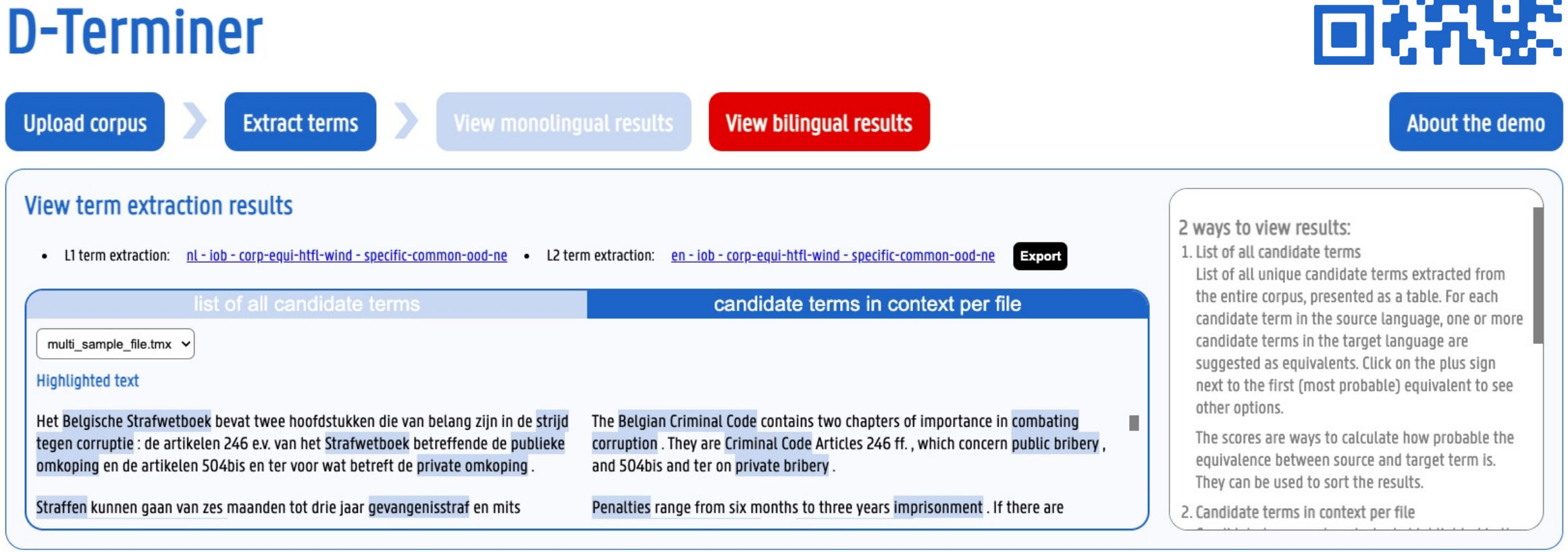
- Only longest possible span annotated
- IOB: Inside, Outside, Boundary
- Or IO (binary)
- Correspondence with original annotations analysed and documented:

  Rigouts Terryn, A., Hoste, V., & Lefever, E. (2022). Tagging Terms in Text: A Supervised Sequential Labelling Approach to Automatic Term Extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(1). https://doi.org/10.1075/term.21010.rig

corpus		corpus counts			annotation counts				
domain	language	docs	sentences	tokens	total	Specific	Common	OOD	NE
corruption	en	12	2,002	50,845	1172	278	641	6	247
	fr	12	1,977	59,13	1207	298	675	5	229
	nl	12	1,988	52,245	1287	308	726	6	247
dressage	en	34	3,09	58,203	1561	769	309	68	415
	fr	78	2,809	61,061	1176	697	234	26	219
	nl	65	3,669	56,45	1541	1020	329	41	151
heart failure	en	190	2,432	55,467	2556	1864	316	157	219
	fr	210	2,177	55,027	2357	1671	486	57	143
	nl	174	2,88	54,966	2215	1535	447	65	168
wind energy	en	5	6,638	57,766	1529	784	295	13	437
	fr	2	4,77	64,989	967	443	308	21	195
	nl	8	3,356	55,328	1229	571	338	21	299
TOTAL		**802**	**37,788**	**681,477**	**18,797**	**10,238**	**5,104**	**486**	**2,969**

corpus		Proportion of sequential labels		
domain	language	I	O	B
corruption	en	81%	12%	7%
	fr	83%	10%	8%
	nl	83%	11%	6%
dressage	en	80%	16%	4%
	fr	83%	13%	4%
	nl	79%	18%	3%
heart failure	en	73%	18%	1\%
	fr	79%	13%	8%
	nl	81%	16%	3%
wind energy	en	82%	10%	8%
	fr	83%	9%	7%
	nl	89%	9%	2%
TOTAL		**81%**	**13%**	**6%**

https://lt3.ugent.be/dterminer/



GHENT UNIVERSITY

language and translation technology team

Thank you! Questions?

ayla.rigoutsterryn@kuleuven.be