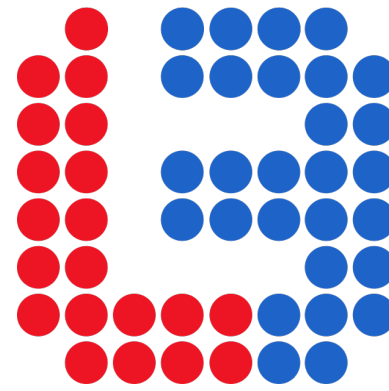


ACTER 1.5

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)

Automatic Term Extraction

list of all candidate terms

candidate terms in context per file

multi_sample_file.tmx ▾

Highlighted text

Zo is er artikel 314 van het Strafwetboek dat de belemmering van de vrijheid van opbod of van inschrijving bij toewijzingen van de eigendom, van het vruchtgebruik of van de huur van roerende of onroerende zaken van een onderneming, van een levering, ... bestraft met een gevangenisstraf van 15 dagen tot zes maanden en met een geldboete van 100 euro tot 3000 euro.

De fiscale wetgeving voorziet er bovendien in dat elk corruptie-element dat op basis van het Strafwetboek kan worden vervolgd, zowel voor een rechtspersoon als voor een natuurlijke persoon, niet aftrekbaar is van de belastbare basis.

Thus Article 314 of the Criminal Code punishes restriction of the freedom of bidding or of registration for transfers of ownership, usufruct or leasing of movable or immovable property of a company, of a supply etc. with 15 days to six months imprisonment and a fine of 100 to 3000 euros.

The tax legislation also specifies that each count of corruption, liable to prosecution under the Criminal Code, both for a natural and for a legal person, can not be deducted from the basis of tax assessment.

candidate term ▲

mitral regurgitation

mitral stenosis

blood pressure

BET

functional mitral stenosis

ischemic mitral regurgitation

diastolic mitral valve tethering

anterior leaflet

papillary muscles

apex

edge-to-edge anastomosis

end-diastolic volume index

LV ejection fraction

LV end-diastolic volume

Machine Learning for ATE

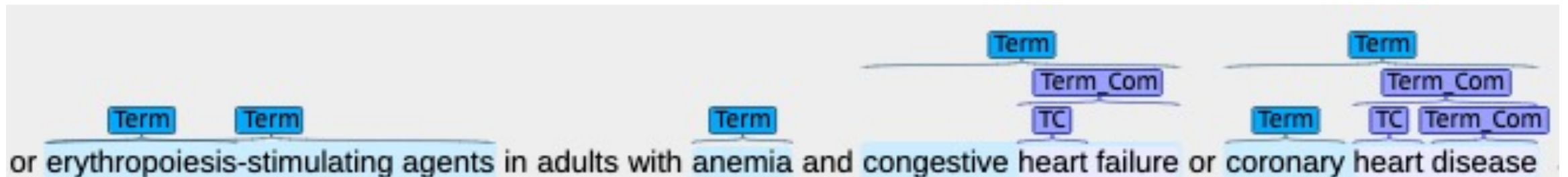
Recent evolution from mainly rule-based methodologies, to more (supervised) machine learning



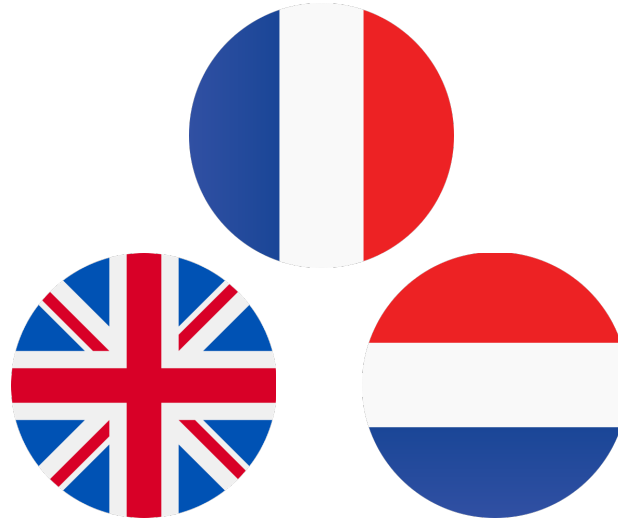
ACTER

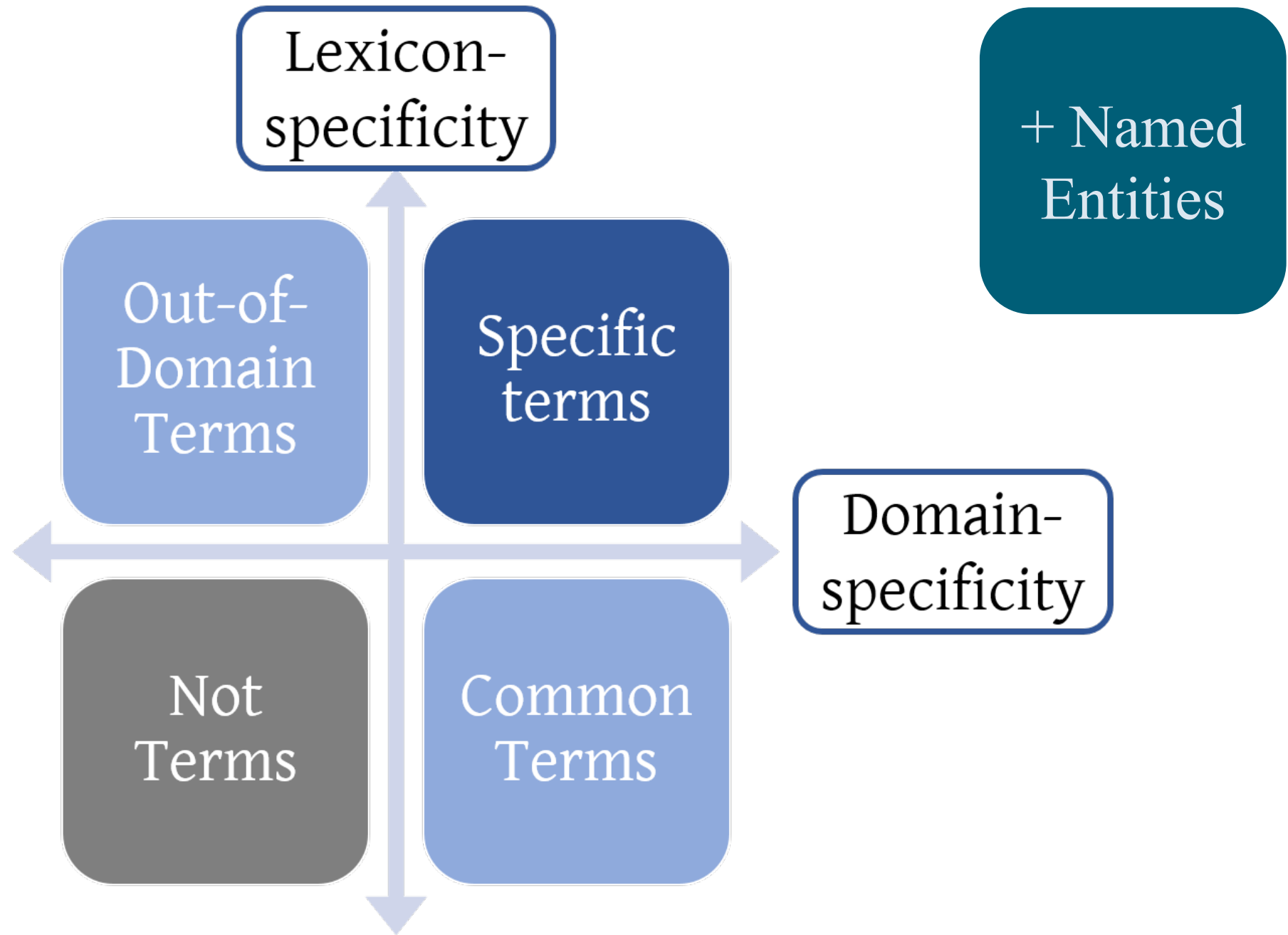
= **A**nnotated **C**orpora for **T**erm **E**xtraction **R**esearch

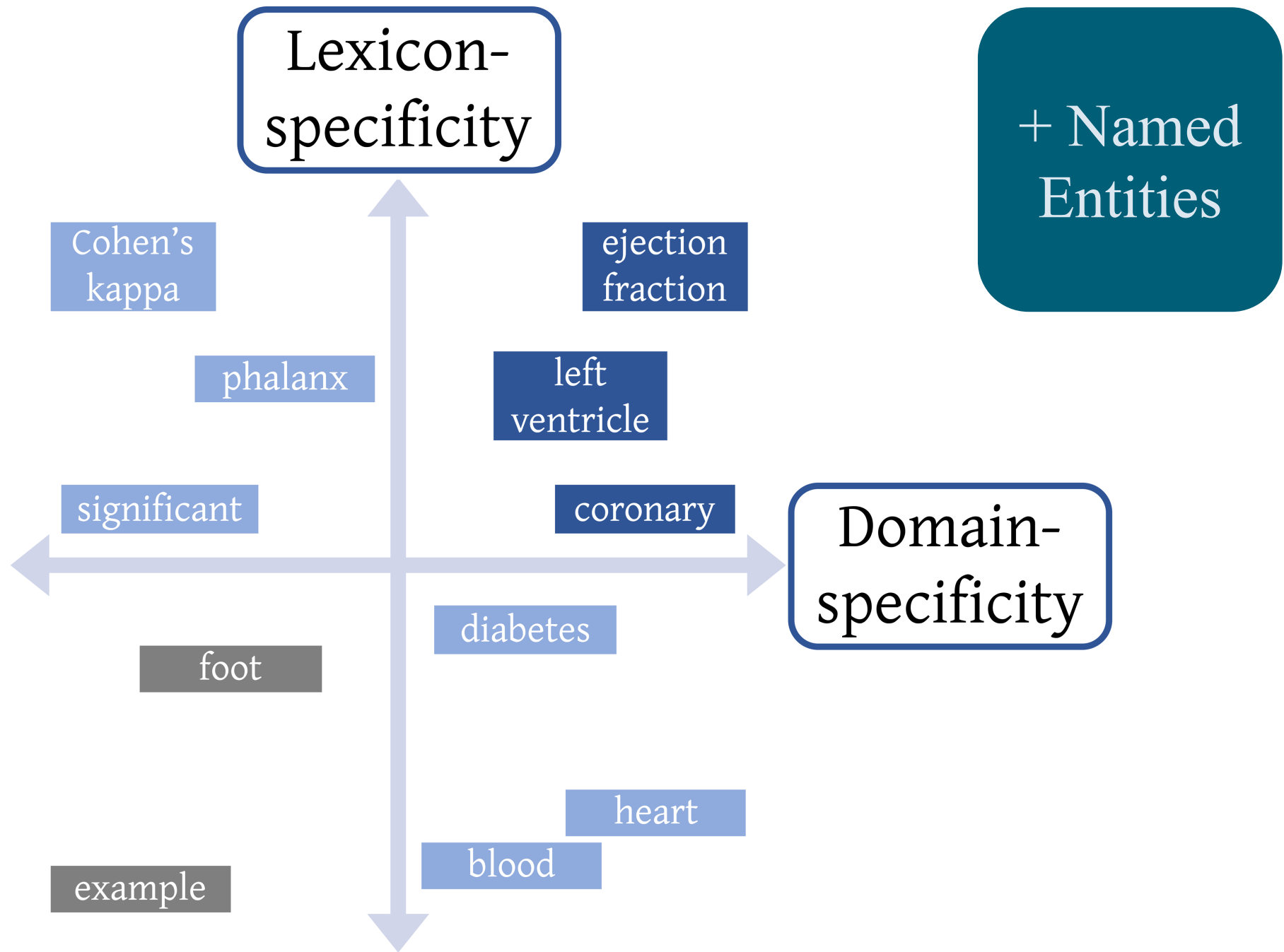
- Annotate based on human intuition, not computer capabilities
- Annotation guidelines, validated through IAA studies



4 domains, 3 languages







Lexicon-specificity

+ Named Entities

Cohen's kappa

ejection fraction

phalanx

left ventricle

significant

coronary

Domain-specificity

foot

diabetes

example

blood

heart

BRAT Rapid Annotation Tool

Comparison of predictors of heart failure-related hospitalization or death in patients with versus without

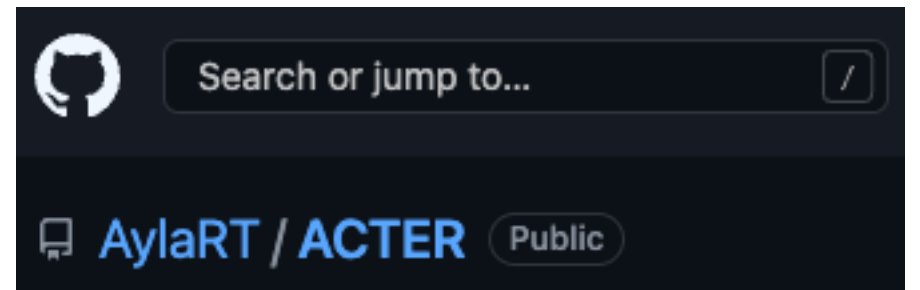
preserved left ventricular ejection fraction.

Heart failure with preserved ejection fraction (HFpEF) is recognized as a major cause of

cardiovascular morbidity and mortality. An ability to identify patients with HFpEF who are at increased

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 0
patients B
with 0
heart B
failure I
? 0
```

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

Rigouts Terry, 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%



D-Terminer

Upload corpus



Extract terms



View monolingual results

View bilingual results

About the demo

View term extraction results

- L1 term extraction: [nl - iob - corp-equi-htfl-wind - specific-common-ood-ne](#)
 - L2 term extraction: [en - iob - corp-equi-htfl-wind - specific-common-ood-ne](#)
- Export**

list of all candidate terms

candidate terms in context per file

multi_sample_file.tmx

Highlighted text

Het Belgische Strafwetboek bevat twee hoofdstukken die van belang zijn in de strijd tegen corruptie : de artikelen 246 e.v. van het Strafwetboek betreffende de publieke omkoping en de artikelen 504bis en ter voor wat betreft de private omkoping .

Straffen kunnen gaan van zes maanden tot drie jaar gevangenisstraf en mits

The Belgian Criminal Code contains two chapters of importance in combating corruption . They are Criminal Code Articles 246 ff. , which concern public bribery , and 504bis and ter on private bribery .

Penalties range from six months to three years imprisonment . If there are

2 ways to view results:

1. List of all candidate terms

List of all unique candidate terms extracted from the entire corpus, presented as a table. For each candidate term in the source language, one or more candidate terms in the target language are suggested as equivalents. Click on the plus sign next to the first (most probable) equivalent to see other options.

The scores are ways to calculate how probable the equivalence between source and target term is. They can be used to sort the results.

2. Candidate terms in context per file

Thank you! Questions?

ayla.rigoutsterryn@kuleuven.be