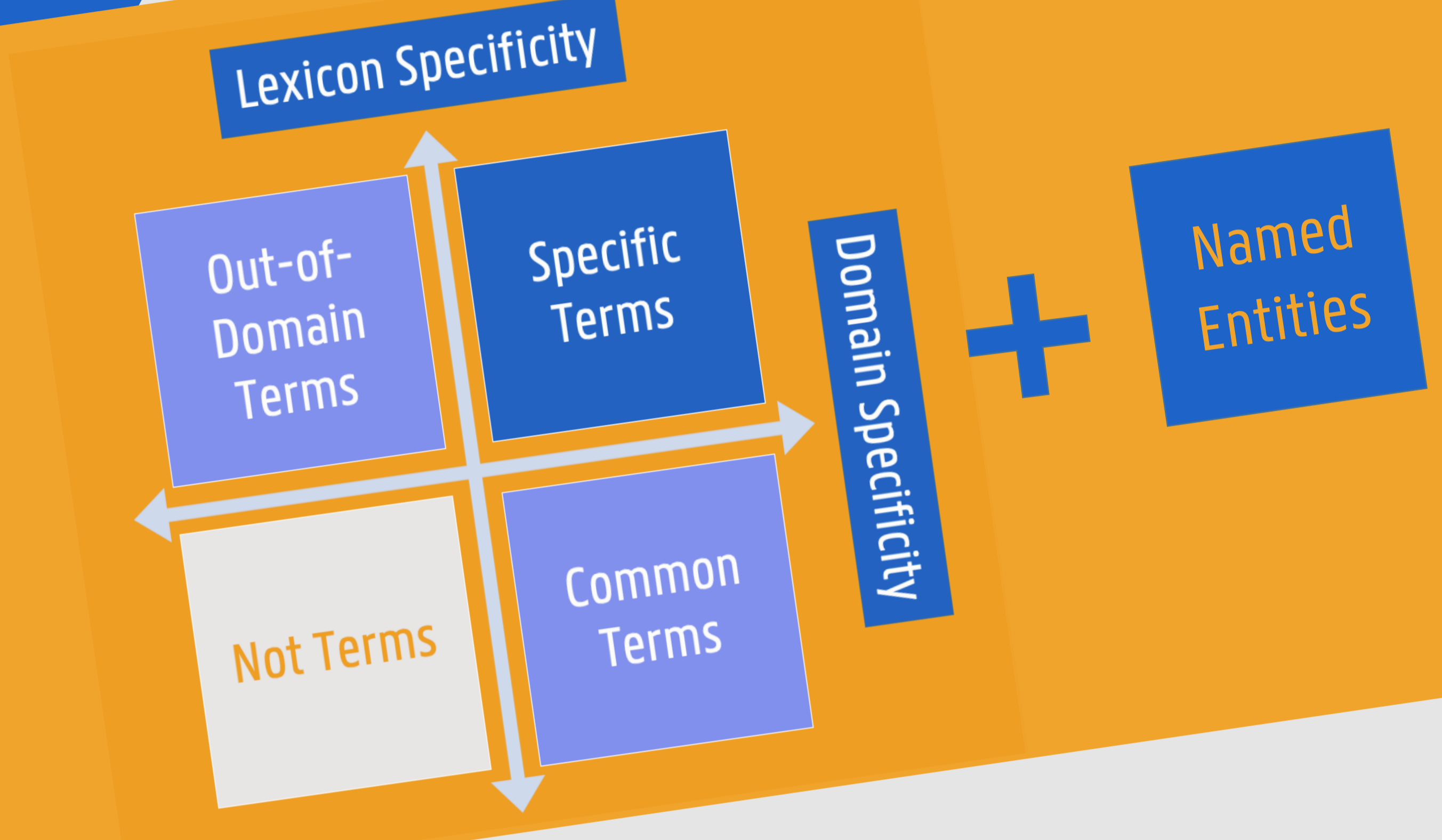


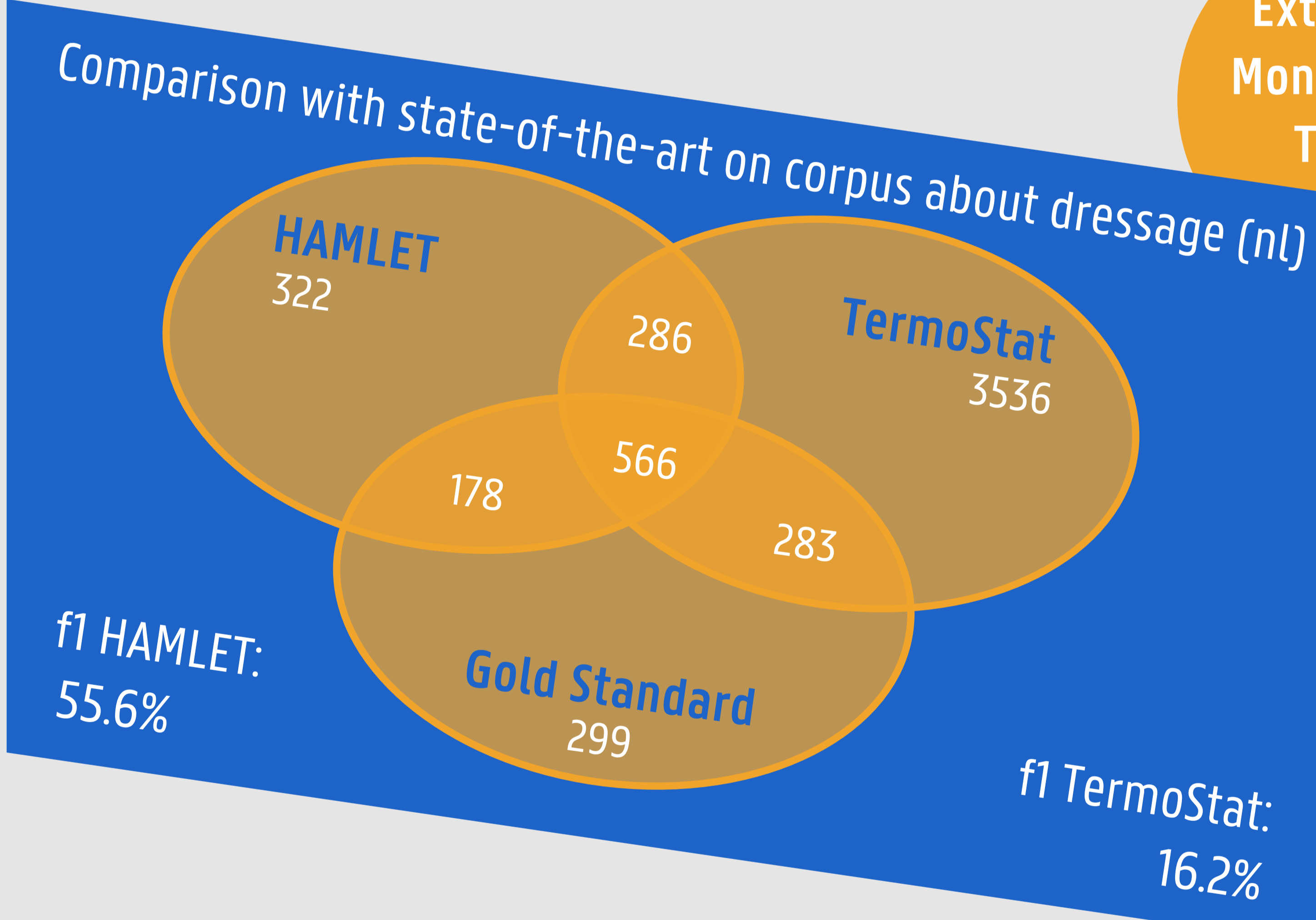
AUTOMATIC TERM EXTRACTION FROM COMPARABLE CORPORA

Data Collection & Annotation

- 1 comparable corpus
- 3 parallel corpora
- 3 languages (Dutch, English, French)
- 4 domains (corruption, dressage, heart failure, wind energy)
- 4 labels (Specific, Common, and OOD Terms & Named Entities)
- 534,559 tokens of specialised text annotated
- 110,444 monolingual annotations (terms & named entities)
- 11,312 cross-lingual annotations (of equivalents)



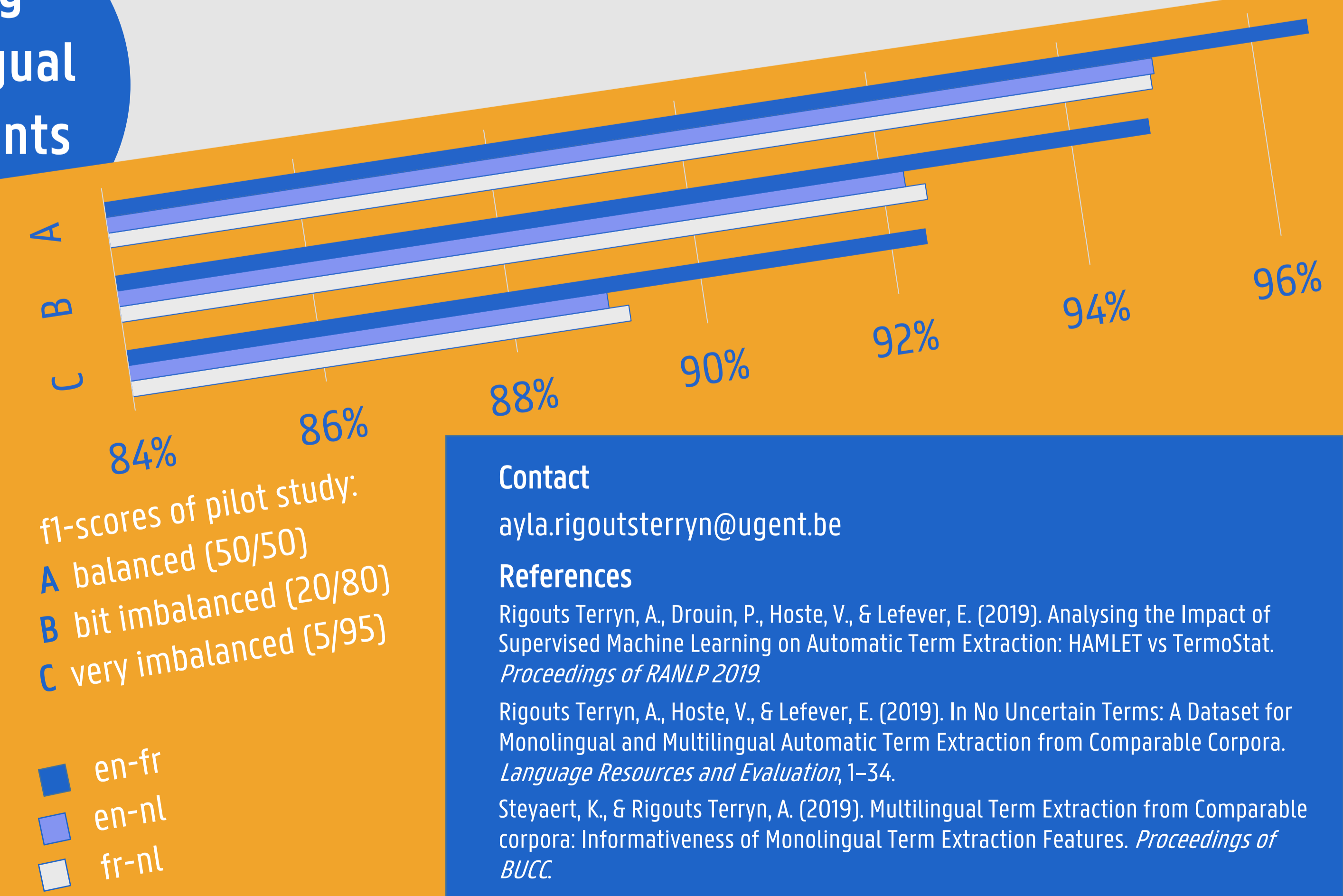
Extracting Monolingual Terms



- HAMLET**
- Hybrid > combination of linguistic and statistical features
- Adaptable > to domains, languages and term types
- Machine Learning approach to supervised binary random forest classifier
- Extract > identify in specialized corpora
- Terminology > specialised, domain-specific linguistic units
- Compares favourably against non-ML approaches

Linking Multilingual Equivalents

- ### Ongoing & Future Work
- ✓ annotate multilingual training & test data
 - ✓ pilot study with binary classifier and existing features
 - add distributional & character features
 - implement entire pipeline for multilingual term extraction from comparable corpora
 - calibrate different components for optimal interaction
 - evaluation and validation



Contact
ayla.rigoutsterryn@ugent.be

References
Rigouts Terryn, A., Drouin, P., Hoste, V., & Lefever, E. (2019). Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat. *Proceedings of RANLP 2019*.
Rigouts Terryn, A., Hoste, V., & Lefever, E. (2019). In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, 1-34.
Steyaert, K., & Rigouts Terryn, A. (2019). Multilingual Term Extraction from Comparable corpora: Informativeness of Monolingual Term Extraction Features. *Proceedings of BUCC*.