**ACTER Annotated Corpora for Term Extraction Research, version 1.5**

ACTER is a manually annotated dataset for term extraction, covering 3 languages (English, French, and Dutch), and 4 domains (corruption, dressage, heart failure, and wind energy).

**Readme structure:**

## 1. General

- **Creator**: Ayla Rigouts Terryn
- **Association**: LT3 Language and Translation Technology Team, Ghent University
- **Date of creation version 1.0**: 17/12/2019
- **Date of creation current version 1.5**: 08/04/2022
- **Last updated**: 08/04/2022
- **Contact**: ayla.rigoutsterryn@ugent.be
- **Context**: Ayla Rigouts Terryn's PhD project + first TermEval shared task (CompuTerm2020)
- **PhD**: D-Termine: Data-driven Term Extraction Methodologies Investigated http://hdl.handle.net/1854/LU-8709150
- **Shared Task**: see https://termeval.ugent.be; workshop proceedings with overview paper at https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/COMPUTERM2020book.pdf)
- **Annotation Guidelines**: http://hdl.handle.net/1854/LU-8503113
- **Source**: https://github.com/AylaRT/ACTER
- **License**: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) (https://creativecommons.org/licenses/by-nc-sa/4.0/)
- **Reference**: Please cite the following Open Access paper if you use this dataset https://doi.org/10.1007/s10579-019-09453-9
    - Authors: Ayla Rigouts Terryn, Véronique Hoste, Els Lefever
    - Title: In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora
    - Date of online publication: 26 March 2019
    - Date of print publication: 2020 (Volume 54, Issue 2, pages 385-418)
    - Journal: Language Resources and Evaluation (LRE)
    - Publisher: Springer
- **Demo**: Online term extraction demo based on dataset: D-Terminer https://lt3.ugent.be/dterminer

## 2. Abbreviations

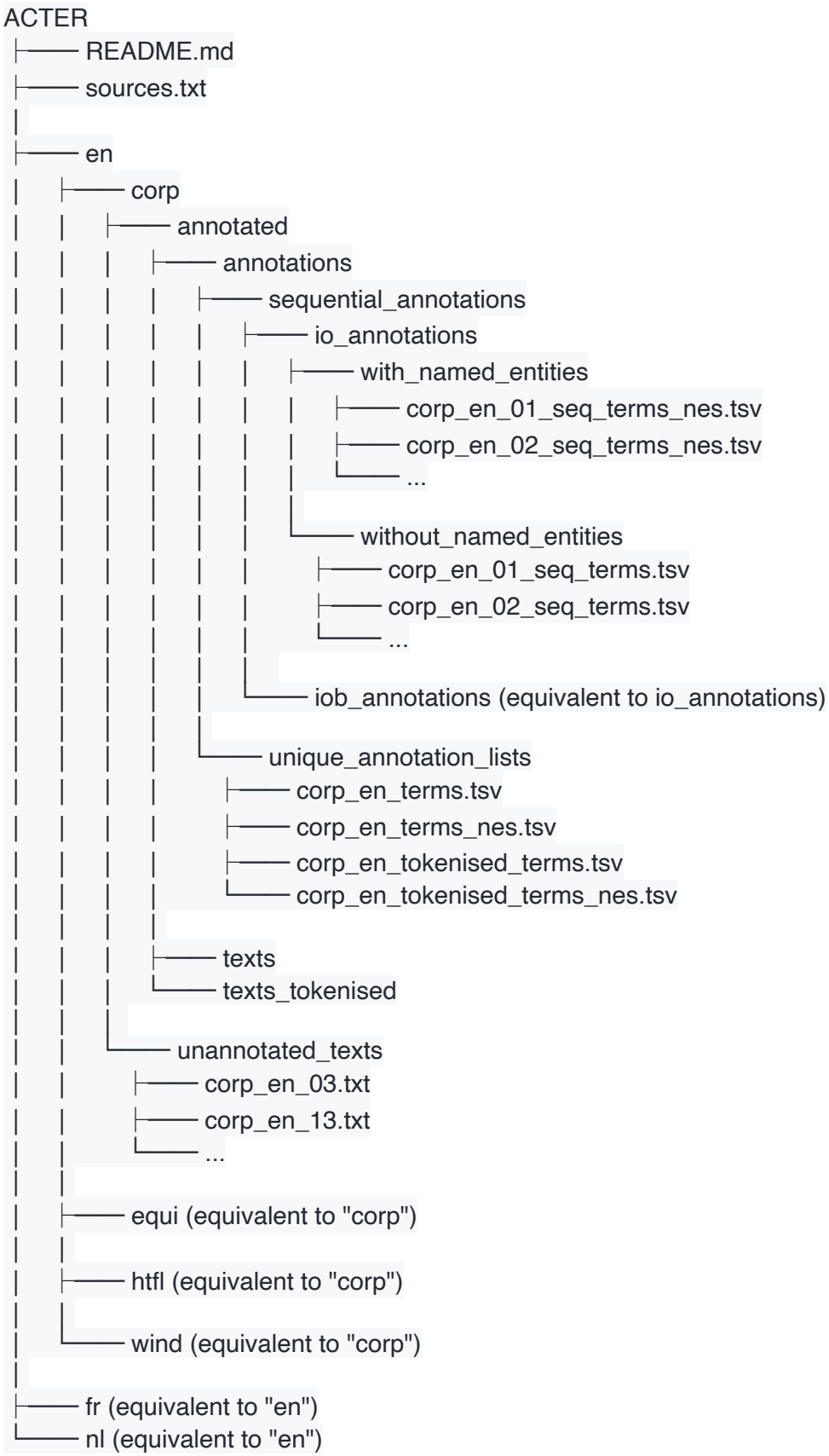
**Languages and domains**:

- "en" = English
- "fr" = French
- "nl" = Dutch
- "corp" = corruption
- "equi" = equitation (dressage)
- "htfl" = heart failure
- "wind" = wind energy
- "cor" = parallel part of corruption corpus; completely unannotated

**Annotation labels**:

- "Spec" or "Specific": Specific Terms
- "Com" or "Common": Common Terms
- "OOD": Out-of-Domain Terms
- "NE(s)": Named Entities

### 3. Data Structure




- **README.md, sources.txt**
  At the first level, there are two files with information about the dataset: the current README.md file and sources.txt, which mentions the sources of all texts in the dataset.

- **languages** and language/**domains**
  At the first level, there is also one directory per language with an identical structure of subdirectories and files for each language. At the second level, there are four directories, i.e., one per domain, each with an identical structure of subdirectories and files. The corpora in each domain are comparable per language (i.e., similar size, topic, style). Only the corruption (corp) corpus is parallel, i.e., translations.

- language/domain/**unannotated_texts**
  Per domain, there are annotated and unannotated texts. For the unannotated texts, only the original (normalised) texts themselves are offered as .txt-files.

- language/domain/**annotated**
  For the annotated texts, many types of information are available, ordered in subdirectories.

- language/domain/annotated/**annotations**
  The annotations can be found here, ordered in subdirectories for different formats of the data.

- language/domain/annotated/**texts** and language/domain/annotated/**texts_tokenised**
  The texts of the annotated corpora can be found here, with the original (normalised) texts and the (normalised) tokenised texts in different directories. The texts were tokenised with LeTs PreProcess*, with one sentence per line and spaces between all tokens.

    - van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The Multilingual LT3 Linguistic Preprocessing Toolkit. Computational Linguistics in the Netherlands Journal, 3, 103–120.)
- language/domain/annotated/annotations/**sequential_annotations**
  Sequential annotations always have one token per line, followed by a tab and a sequential label (more info in next section). There are empty lines between sentences.

    - .../**io(b)_annotations**: one directory per annotation scheme (IO versus IOB)
    - ../io(b)_annotations/**with(out)_named_entities**: per annotation scheme, one directory for data including and excluding Named Entities.
- language/domain/annotated/annotations/**unique_annotation_lists**
  Lists of all unique annotations (lowercased, unlemmatised) for the entire corpus (langauge-domain), with one annotation per line, followed by a tab and its label (Specific_Term, Common_Term, OOD_Term, or Named Entity).

    - **domain_language_terms.tsv**: original annotations as they occur in the untokenised texts, including only term annotations (Specific_Term, Common_Term, OOD_Term), no Named Entities.
    - **domain_language_terms_nes.tsv**: same, but including Named Entities.
    - **domain_language_tokenised_terms.tsv**: original annotations mapped to tokens, including only those annotations that align exactly with token boundaries at least once in the corpus; including only term annotations (Specific_Term, Common_Term, OOD_Term), no Named Entities.
    - **domain_language_tokenised_terms_nes.tsv**: same, but including Named Entities.

## 4. Annotations

### 4.1 General

The annotations are provided in simple UTF-8 encoded plain text files. No lemmatisation was performed.

### 4.2 Sequential annotations

#### 4.2.1 Reference

For an in-depth review of how the sequential labels were obtained and how they relate to the list-versions of the annotations, please check:

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

#### 4.2.2 General

- one token per line, followed by a tab and the IO(B) label
- based on the tokenised version of the corpus (see under language/domain/annotated/texts_tokenised)
- normalised (see further), but with original casing
- in case of nested annotations, the longest possible span is given sequential labels.
    - e.g., "myocyte hypertrophy": if "myocyte", "hypertrophy", and "myocyte hypertrophy" were originally all annotated separately, the sequential labels will be based only on the longest possible annotation, i.e., "myocyte hypertrophy".
- when a token was partially (not completely) annotated, the token is gets a positive (I or B) label (= different strategy for unique annotation lists)
    - e.g., "defibrillator-only therapy": if "defibrillator" was annotated but the complete token ("defibrillator-only") was not, the full token will still get a positive sequential label, but "defibrillator" will only occur in the unique annotations lists if it occurs as a separate token somewhere else in the corpus.
- annotations of parts of terms also get a positive (I or B) label (= different strategy for unique annotation lists)
    - e.g. "left and right ventricular assist devices": "left" is part of the term "left ventricular assist devices", but because the term is split, the full term cannot be annotated with an uninterrupted annotation. "left" will get a positive sequential label, but will not be included as an annotation in the unique annotation lists

#### 4.2.3 IOB versus IO

**IOB** (Inside, Outside, beginning): the first token of any annotation gets labelled "B" and each subsequent token of the same annotation gets labelled "I". Tokens that are not part of any annotation are "O".

**IO** (Inside, Outside): same as IOB but with no distinction between the first and subsequent tokens of an annotation.

**Impact**: binary labelling (IO) is easier to model, so technically gets higher f1-scores, but loses some detail in case of adjacent annotations. For instance, if "diabetic patients" occurs and both "diabetic" and "patients" are annotated separately, but "diabetic patients" is not annotated as a term, then this can be accurately encoded with IOB labels ("diabetic[B] patients[B]"). With the binary IO scheme, this will become "diabetic[I] patients[I]", which would be the same as if "diabetic patients" were annotated, instead of the two separate entities.

For a more detailed analysis of the difference, see the paper cited in 4.2.1.

### 4.3 Unique annotations lists

#### 4.3.1 General

- one annotation per line, tab-separated from its label
- one list per corpus (language-domain), combining all unique annotations (no doubles)
- normalised and lowercased
- in case of annotations with different labels depending on the context, the most frequently assigned label is given.
- only complete and uninterrupted annotations are included (in contrast to the sequential dataset)

#### 4.3.2 Labels

More details on the annotation labels are provided in the main publication accompanying this dataset.

**Overview with examples in the domain of heart failure**:

- **Specific Terms**: are domain-specific and lexicon-specific, i.e., relevant to the domain and known only by domain-experts, not by laypeople.
  - e.g., ejection fraction, ventricular assist device, tachycardia
- **Common Terms**: are domain-specific but not lexicon-specific, i.e., relevant to the domain and known by laypeople
  - heart, patients, quality-of-life
- **Out-of-Domain Terms**: are not domain-specific, but they are lexicon-specific, i.e, not directly relevant to the domain, but not generally known by laypeople
  - e.g., confidence interval, p-value, structured-telephone-support
- **Named Entities**: are proper names of people, places, organisations, brands, etc.
  - e.g., MEDLINE, HeartMate, New York

#### 4.3.3 Tokenised annotations

Tokenised annotations have a space between each token and are mostly identical to the original annotations, except that they only include those annotations that can be mapped to complete tokens. When an annotation never aligns with token boundaries, it is not included. The differences are minor (see also 5.3 Number of annotations per corpus), but it is important to mention which of the two versions of the data is used.

## 5. Additional Information

### 5.1 Websites

- For more information about the annotation guidelines, visit: http://hdl.handle.net/1854/LU-8503113
- For more information about the TermEval shared task, visit: https://termeval.ugent.be
- For more information about the CompuTerm workshop, visit: https://sites.google.com/view/computerm2020/
- Online term extraction demo based on dataset: D-Terminer https://lt3.ugent.be/determiner

### 5.2 Publications

- Rigouts Terryn, A., Hoste, V., & Lefever, E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. Proceedings of LREC 2018.
- 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, 54(2), 385–418. https://doi.org/10.1007/s10579-019-09453-9
- Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), 85–94.
- 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

The dataset has been updated since the publication of the former two papers. These papers also discuss aspects of the data which have not been made available yet, such as cross-lingual annotations and information on the span of the annotations.

### 5.3 Number of annotations per corpus

#### 5.3.1 Explanation of differences in numbers

- **Original versus tokenised**: only annotations that can be accurately mapped to token boundaries at least once in the corpus, are included as tokenised annotations.
- **differences per label (with and without NEs)**: the most commonly assigned label is mentioned, so when Named Entities are included or excluded, this can impact the frequencies of the other labels as well, whenever an instance is assigned a Named Entity label in some contexts and a different label in others.

#### 5.3.2 Original annotations, with Named Entities

*path: language/domain/annotated/annotations/unique_annotation_lists/domain_language_terms_nes.tsv*

**18,928 Annotations**

| Domain | Language | Specific Terms | Common Terms | OOD Terms | Named Entities | Total |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| corp | en | 278 | 642 | 6 | 247 | 1173 |
| corp | fr | 298 | 675 | 5 | 229 | 1207 |
| corp | nl | 310 | 730 | 6 | 249 | 1295 |
| equi | en | 777 | 309 | 69 | 420 | 1575 |
| equi | fr | 701 | 234 | 26 | 220 | 1181 |
| equi | nl | 1021 | 330 | 41 | 152 | 1544 |
| htfl | en | 1883 | 319 | 157 | 222 | 2581 |
| htfl | fr | 1684 | 487 | 57 | 146 | 2374 |
| htfl | nl | 1559 | 449 | 66 | 180 | 2254 |
| wind | en | 781 | 296 | 14 | 440 | 1531 |
| wind | fr | 444 | 308 | 21 | 195 | 968 |
| wind | nl | 577 | 342 | 21 | 305 | 1245 |

#### 5.3.3 Original annotations, without Named Entities

*path: language/domain/annotated/annotations/unique_annotation_lists/domain_language_terms.tsv*

#### 15,929 Annotations

| Domain | Language | Specific Terms | Common Terms | OOD Terms | Total |
|---|---|---|---|---|---|
| corp | en | 278 | 643 | 6 | 927 |
| corp | fr | 298 | 676 | 5 | 979 |
| corp | nl | 310 | 731 | 6 | 1047 |
| equi | en | 777 | 309 | 69 | 1155 |
| equi | fr | 701 | 234 | 26 | 961 |
| equi | nl | 1022 | 330 | 41 | 1393 |
| htfl | en | 1884 | 319 | 158 | 2361 |
| htfl | fr | 1684 | 487 | 57 | 2228 |
| htfl | nl | 1559 | 449 | 66 | 2074 |
| wind | en | 781 | 296 | 14 | 1091 |
| wind | fr | 444 | 308 | 21 | 773 |

| | | | | | |
|---|---|---|---|---|---|
| wind | nl | 577 | 342 | 21 | 940 |

#### 5.3.4 Tokenised annotations, with Named Entities

*path: language/domain/annotated/annotations/unique_annotation_lists/domain_language_tokenised_terms_nes.tsv*

**18,797 Annotations**

| Domain | Language | Specific Terms | Common Terms | OOD Terms | Named Entities | Total |
|---|---|---|---|---|---|---|
| corp | en | 278 | 641 | 6 | 247 | 1172 |
| corp | fr | 298 | 675 | 5 | 229 | 1207 |
| corp | nl | 308 | 726 | 6 | 249 | 1287 |
| equi | en | 769 | 309 | 68 | 420 | 1561 |
| equi | fr | 697 | 234 | 26 | 220 | 1176 |
| equi | nl | 1020 | 329 | 41 | 152 | 1541 |
| htfl | en | 1864 | 316 | 157 | 222 | 2556 |
| htfl | fr | 1671 | 486 | 57 | 146 | 2357 |
| htfl | nl | 1535 | 447 | 65 | 180 | 2215 |
| wind | en | 784 | 295 | 13 | 440 | 1529 |
| wind | fr | 443 | 308 | 21 | 195 | 967 |
| wind | nl | 571 | 338 | 21 | 305 | 1229 |
| | | | | | | |
| | | | | | | |

#### 5.3.5 Tokenised annotations, without Named Entities

*path: language/domain/annotated/annotations/unique_annotation_lists/domain_language_tokenised_terms.tsv*

**15,834 Annotations**

| Domain | Language | Specific Terms | Common Terms | OOD Terms | Total |
|---|---|---|---|---|---|
| corp | en | 278 | 642 | 6 | 926 |
| corp | fr | 298 | 676 | 5 | 979 |

| | | | | | |
|---|---|---|---|---|---|
| corp | nl | 308 | 727 | 6 | 1041 |
| equi | en | 769 | 309 | 68 | 1146 |
| equi | fr | 697 | 234 | 26 | 957 |
| equi | nl | 1021 | 329 | 41 | 1391 |
| htfl | en | 1865 | 316 | 158 | 2339 |
| htfl | fr | 1671 | 486 | 57 | 2214 |
| htfl | nl | 1535 | 447 | 65 | 2047 |
| wind | en | 784 | 295 | 13 | 1092 |
| wind | fr | 443 | 308 | 21 | 772 |
| wind | nl | 571 | 338 | 21 | 930 |

### 5.4 Corpus counts (only annotated parts of corpus)

| Domain | Language | # files | # sentences | # tokens (excl. EOS) | # tokens (incl. EOS) |
|---|---|---|---|---|---|
| corp | en | 12 | 2002 | 52,847 | 54,849 |
| corp | fr | 12 | 1977 | 61,107 | 63,084 |
| corp | nl | 12 | 1988 | 54,233 | 56,221 |
| equi | en | 34 | 3090 | 61,293 | 64,383 |
| equi | fr | 78 | 2809 | 63,870 | 66,679 |
| equi | nl | 65 | 3669 | 60,119 | 63,788 |
| htfl | en | 190 | 2432 | 57,899 | 60,331 |
| htfl | fr | 210 | 2177 | 57,204 | 59,381 |
| htfl | nl | 174 | 2880 | 57,846 | 60,726 |
| wind | en | 5 | 6638 | 64,404 | 71,042 |
| wind | fr | 2 | 4770 | 69,759 | 74,529 |
| wind | nl | 8 | 3356 | 58,684 | 62,040 |

### 5.6 Normalisation

The following normalisation procedures are applied to all available versions of the data:

1. Unidecode to avoid encoding issues with the "unicodedata" Python package
   normalised_text = unicodedata.normalize("NFC", text_string_to_normalise)

```
2.
3.  Make sure all dashes and quotes use the same characters
    dashes = ["-", "–", "-"]
4.  double_quotes = ["'", '“', '”', '‘', '„', '‟', '″']
5.  single_quotes = ["'", '`', '´', '’', '‘', '′']
6.
7.  # fix double character quotes
8.  for double_quote in [',,', "''", '``', '“”', '’’']:
9.      if double_quote in text:
10.         text_string_to_normalise = text_string_to_normalise.replace(double_quote, "'')
11.
12. # fix single character dashes and quotes
13. normalised_text = ""
14. for char in text_string_to_normalise:
15.     if char in dashes:
16.         string_normalised += "-"
17.     elif char in double_quotes:
18.         string_normalised += "'
19.     elif char in single_quotes:
20.         string_normalised += "'"
21.     else:
22.         string_normalised += char
23.
24. Replace a specifically accented I which could not be handled well with lowercasing
    normalised_text = text_string_to_normalise.replace("İ", "I")
25.
26. Remove very specific and rare special characters which cause problems with Transformers
    library
    problem_chars = ["⌈", "⌉", "⌊", "⌋", ""]
27. for problem_char in problem_chars:
28.     normalised_text = text_string_to_normalise.replace(problem_char, "")
29.
```

**6. Updates**

**Changes version 1.0 > version 1.1**

- English corpora:
  - corruption
    - Removed 1 NE: 'com(2007) 805 final'
  - wind energy
    - Removed 2 terms: 'variable pitch blades', 'renewable sources'
    - Removed 1 NE: 'skuodas'
- French corpora:
  - corruption:
    - Removed 2 terms: 'indélicat', 'loi relative à la corruption'
  - equitation-dressage
    - Removed 2 terms: 'canons', 'équilibration'

    - wind energy
        - Added 1 term: 'systèmes mutisources-multistockages'
        - Removed 4 terms: 'systèmes mutisources', 'quadrature', 'inductance directe', 'résistance statorique'
        - Removed 98 NEs: 'bar', 'esk', 'akh', 'tht', 'enbw', 'rich', 'kama', 'man', 'sab', 'mer', 'deg', 'mor', 'aba', 'abo', 'ana', 'azm', 'joo', 'jen', 'pri', 'han', 'ree', 'dav', 'cou', 'hol', 'sau', 'lal', 'lei', 'vet', 'pur', 'per', 'her', 'hau', 'ans', 'slo', 'win', 'thi', 'ela', 'stem', 'cer', 'lav', 'ack', 'e.on', 'cim', 'luo', 'wik', 'ds1103', 'fag', 'and', 'alm', 'pan', 'rap', 'ric', 'saa', 'reb', 'bor', 'kin', 'sem', 'ecr', 'fau', 'ukt', 'kun', 'creg', 'sal', 'bou', 'crap', 'mog', 'nget', 'stu', 'sei', 'lec', 'dir', 'nor', 'abb', 'doh', 'rwe', 'mul', 'oud', 'bea', '96/92/ce', 'gar', 'eri', 'cal', 'goi', 'ish', 'fra', 'cra', 'bna', 'ull', 'des', 'ips', 'dro', 'uct', 'mat', 'ds 1104', 'mar', 'svk', 'bla', 'buh'
- Dutch corpora:
    - corruption
        - Added 1 term: 'anticorruptie-eenheid'
        - Removed 4 terms: 'verslagen corruptiebestrijding', 'auditdiensten', 'anticorruptie', 'wet betreffende de omkoping'
    - equitation-dressage
        - Removed 2 terms: 'promotie', 'stuw'
    - wind energy
        - Removed 2 terms: 'windturbines een horizontale as', 'power coefficient'

**Changes version 1.1 > version 1.2**

- Included domain of heart failure (test domain for TermEval shared task)

**Changes version 1.2 > version 1.3**

- corrected wrong sources in htfl_nl
- changed heart failure abbreviation to "htfl" to be consistent with four-letter domain abbreviations
- created Github repository for data + submitted it to CLARIN

**Changes version 1.3 > version 1.4**

- applied limited normalisation on both texts and annotations:
    - unicodedata.normalize("NFC", text)
    - normalising all dashes to "-", all single quotes to "'" and all double quotes to '"'

**Changes version 1.4 > version 1.5**

Not many changes to actual annotations, but major update to how the annotations are presented etc.:

- Removed a few very long Named Entity annotations (from wind-en and from htfl-en; counts updated) over which there was doubt whether it was a real NE.
- Updated normalisation:
    - Replaced "İ" with "I" in the annotations to avoid problems lowercasing (concerns mainly wind_en_01)
    - Removed rare but problematic characters: ["⌈", "⌉", "⌊", "⌋", ""] (not handled well by some transformers)
- Major update of README.md
- Different structure of all data:
    - include sequential annotations
    - include tokenised versions of annotations

**7. Error Reporting**

The ACTER dataset is an ongoing project, so we are always looking to improve the data. Any questions or issues regarding this dataset may be reported via the Github repository at: https://github.com/AylaRT/ACTER and will be addressed asap.

## 8. License

- *License*: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) (https://creativecommons.org/licenses/by-nc-sa/4.0/)
- *Reference*: Please cite the following Open Access paper if you use this dataset for your research (https://doi.org/10.1007/s10579-019-09453-9)
    - Authors: Ayla Rigouts Terryn, Véronique Hoste, Els Lefever
    - Title: In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora
    - Date of online publication: 26 March 2019
    - Date of print publication: 2020 (Volume 54, Issue 2, pages 385-418)
    - Journal: Language Resources and Evaluation (LRE)
    - Publisher: Springer

The data can be freely used and adapted for non-commercial purposes, provided the above-mentioned paper is cited and any changes made to the data are clearly stated.