Title	Overview of manually annotated text corpora
Version	2.5
Author(s)	Darja Fišer, Jakob Lenardič
Date	11-03-2019
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2019-1384

Overview of manually annotated text corpora

Table of contents

1. Introduction

In the following report, we present an overview of manually annotated text corpora that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). Manual corpora are collections of texts containing manually validated or manually assigned linguistic information, such as morphosyntactic tags, lemmas, syntactic parses, named entities etc. These corpora can be used to train new language annotation tools as well as to test the accuracy of existing annotation tools.

The report was conducted in two steps:

(i) by manually searching the VLO and the national consortia with the following keywords: “manual* corpus”, “training corpus”, “gold corpus”, “treebank”, and “manual* anno*”
(ii) with input provided by CLARIN National Coordinators and User Involvement Representatives

The full results are available in a Google Docs Spreadsheet. A total of 73 corpora were originally identified. Information on about half of the identified corpora was provided by User Involvement representatives and National Coordinators, whom we would like to thank for their invaluable input. In this report, we originally included 62 corpora, excluding from the survey speech corpora, which deserve special attention and will be covered in a future survey, and corpora that are listed in various repositories but do not provide downloadable or otherwise accessible data. 8 additional corpora and 4 corpus collections were added after a round of revision based on comments received from CLARIN National Coordinators and User Involvement representatives.

In Section 2, we provide a comprehensive overview of the manually annotated text corpora that are part of the CLARIN infrastructure, describing their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, annotation, license). In Section 3, we list and describe the corpora in detail, classifying them into 6 categories: PoS/MSD tagging, Lemmatisation, Syntactic parsing, Named Entity recognition, Sentiment analysis, and Other.

2. Overview of manually annotated text corpora in the CLARIN infrastructure

There are 74 manually annotated training corpora and corpus collections in the CLARIN infrastructure (for a detailed overview, see Section 3). Section 2.1 lists the resources that cannot yet be found in the VLO, but can be found through a CLARIN node. Section 2.2 provides an overview of the availability of the corpora and corpus collections and Section 2.3 an overview of the relevant metadata.

2.1. Identification

The vast majority of the CLARIN corpora and corpus collections can be found in the VLO (70 out of 74). The following 4 corpora cannot, likely because they are not included in a B-certified CLARIN repository.

[1] NKJP1M
[2] Polish Coreference Corpus
[3] Polish Dependency Bank in Universal Dependency format
[4] Polish Summaries Corpus

2.2. Availability

The following 22 corpora and corpus collections are available for download and through a concordancer:

[1] Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0
[2] Training corpus hr500k 1.0
[3] Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0
[4] Training corpus SETimes.SR 1.0
[5] CMC training corpus Janes-Norm 1.2
[6] CMC training corpus Janes-Tag 2.0
[7] Training corpus ssj500k 2.1
[8] Czech Legal Text Treebank 2.0
[9] Prague Discourse Treebank 2.0
[10] Prague Czech-English Dependency Treebank 2.0 Coref
[11] The Morphologically Annotated Part of BulTreeBank
[12] Lassy Klein-corpus
[13] BNC Sampler
[14] Corpus of morphologically disambiguated Estonian texts
[15] Prague Dependency Treebank 3.5
[16] FicTree 1.0
[17] HamleDT 3.0
[18] Universal Dependencies 2.3
[19] Treebanks of INESS
[20] Facebook Data for Sentiment Analysis
[21] Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1)
[22] The ACL RD-TEX 2.0

The following 52 corpora and corpus collections are available for download only:

[1] MULTEXT-East "1984" annotated corpus 4.0
[2] xLiMe Twitter Corpus XTC 1.0.1
[3] CINTIL-Corpus Internacional do Português
[4] Greek Coreference Corpus
[5] Greek Textual Entailment Corpus
[6] Training corpus jos1M 1.1
[7] TimeML annotated corpus of Estonian newspaper articles
[8] Czech Named Entity Corpus 1.1
[9] Polish Coreference Corpus
[10] SoNaR-1
[11] B4 Heliand
[12] Polish Spatial Texts 1.0
[13] Prague Arabic Dependency Treebank 1.0
[14] Slovak Dependency Treebank
[15] Artificial Treebank with Ellipsis

[16] Estonian Treebank
[17] UD Estonian ver.2.3
[18] Turku Dependency Treebank
[19] Polish Dependency Bank in Universal Dependency format
[20] CINTIL DependencyBank
[21] CINTIL TreeBank
[22] CINTIL-DeepBank
[23] CINTIL-PropBank
[24] Tamil Dependency Treebank v0.1
[25] Twitter sentiment for 15 European languages
[26] Dataset and baseline model of moderated content FRENK-STYRIA-24sata 1.0
[27] Aspect-Term Annotated Customer Reviews in Czech
[28] Grundtvig's Works Corpus
[29] NoReC: The Norwegian Review Corpus
[30] KPWr (Polish Corpus of Wrocław University of Technology) 1.2
[31] NKJP1M
[32] WUT Relations Between Sentences Corpus
[33] Corpus of comma placement Vejica 1.3
[34] Dataset and baseline model of moderated content FRENK-MMC-RTV 1.0
[35] Manually sentiment annotated Slovenian news corpus SentiNews 1.0
[36] Terminology identification dataset KAS-term 1.0
[37] Bilingual terminology extraction dataset KAS-biterm 1.0
[38] Dependency-Annotated Subset of the CREG Corpus
[39] Szeged Corpus 2.0
[40] Lithuanian morphologically annotated corpus - MATAS
[41] Syntactic Reference Corpus of Medieval French
[42] Tübingen Treebank of Written German / Newspaper Corpus (TüBa-D/Z)
[43] Szeged Treebank 2.0
[44] Lithuanian Treebank ALKSNIS
[45] Polish Summaries Corpus
[46] Finnish TreeBank 1
[47] Finnish TreeBank 2
[48] GRUG Parallel Treebank
[49] Speech, Thought and Writing Presentation Corpus
[50] TimeML annotated corpus of Estonian newspaper articles
[51] Estonian Treebank annotated with coreference relations
[52] Semantically disambiguated corpus of Estonian

2.3. Metadata

2.3.1. Languages

The vast majority (64 out of 74) of the corpora and corpus collections are monolingual, for the following 19 languages:

[1] Arabic (1 corpus)
[2] Bulgarian (1 corpus)
[3] Croatian (3 corpora)
[4] Czech (8 corpora)
[5] Danish (1 corpus)
[6] Dutch (2 corpora)
[7] English (3 corpora)
[8] Estonian (7 corpora)
[9] Finnish (3 corpora)
[10] French (1 corpus)
[11] German (3 corpora)
[12] Greek (2 corpora)
[13] Hungarian (2 corpora)
[14] Lithuanian (2 corpora)

[15] Norwegian (1 corpus)
[16] Polish (7 corpora)
[17] Portuguese (5 corpora)
[18] Serbian (2 corpora)
[19] Slovenian (8 corpora)
[20] Tamil (1 corpus)

10 corpora and corpus collections are multilingual. The resource with the most languages is the corpus collections Universal Dependencies 2.3 (75 languages, full list).

2.3.2. Size

Information on size is available for all corpora and corpus collections except for the following:

[1] Grundtvig's Works Corpus

Size overview (for corpora and corpus collections that have information on token number available):

- 39 contain <1 million tokens
- 13 contain between 1 and 10 million tokens
- 3 contains >10 million tokens (NoReC: The Norwegian Review Corpus, Universal Dependencies 2.3, Treebanks of INESS)

2.3.3. Annotation

Information on the manual annotation layers is available for all the corpora and corpus collections. In this report, we take into account the following annotation layers:

- PoS/MSD tagging (13 corpora)
- Lemmatisation (7 corpora)
- Syntactic parsing (32 corpora and 3 corpus collections)
- Named Entity recognition (11 corpora)
- Sentiment mark-up (7 corpora)
- Other (23 corpora and 1 corpus collection)

As Table 1 shows, the most training corpora are available for Part of Speech tagging, syntactic parsing, and Named Entity recognition. Note that many corpora are manually annotated for several layers, so the numbers reported in Table 1 do not match the number of corpora in the infrastructure reported in Section 2.3.1 where the number of corpora is given without the duplicated listings. Please also note that for clearer presentation, Table 1 lists individual corpora only, while all the corpora belonging to the 4 corpus collections – HamleDT 3.0, Treebanks of INESS, Universal Dependencies 2.3 and Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1) are listed in Sections 3.3 and 3.6.

Table 1: The availability of manually annotated corpora in the CLARIN infrastructure according to language and annotation layer.

	PoS/MSD	Lemm.	Syntactic	NER	Sentiment	Other	Σ
Arabic			1				1
Bulgarian	1						1
Croatian	1	2	1	2	1	1	8
Czech			5	1	2	2	10
Danish						1	1
Dutch			2			1	3
English	1					2	3
Estonian	1		3			3	7
Finnish			3				3
French			1				1
German			3				3
Greek						2	2
Hungarian	1		1				2
Lithuanian	1		1				2
Norwegian					1		1
Polish	1		1	2		3	7
Portuguese	1		4	1			6
Serbian	1	1	1	2		1	6
Slovenian	2	3	1	2	1	4	13
Tamil			1				1
Multilingual	2	1	3	1	2	3	12
Σ	13	7	32	11	7	23	93

2.3.4. Licence

Information on license is available for all the corpora and corpus collections except for the following two:

[1] Szeged Corpus 2.0
[2] Szeged Treebank 2.0

52 out of 74 corpora a and corpus collections re available under CC-BY licences, 9 under ELRA licences, 2 under CLARIN ACA, 1 under CLARIN RES, the rest miscellaneous.

3. Manually annotated corpora in the CLARIN infrastructure – detailed presentation

In this section, we present the manually annotated corpora and corpus collections in detail, listing them under the relevant annotation layers. If a corpus is manually annotated for more than one linguistic information, then it is listed under all the relevant sections. For instance, the xLiMe Twitter Corpus XTC 1.0.1 is manually annotated for PoS tags, Named Entities and sentiment, so it is listed under all the three relevant sections. Each section begins with a brief overview of the availability and relevant metadata of the corpora. Note that 4 out of the 74 resources are treated as corpus collections, which were annotated under the following umbrella initiatives: HamleDT 3.0, Treebanks of INESS, Universal Dependencies 2.3, and Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1).

3.1. PoS/MSD tagging

There are 13 corpora with manual PoS/MSD-tagging in the CLARIN infrastructure, presented in detail in Table 2. For reasons of space, we list only those PoS/MSD-tagged corpora that are *not* syntactically parsed in this Section; for a list of corpora with both manual PoS-tagging and syntactic parsing, see Section 3.3.

The following 7 corpora are available for download and through a concordancer:

[1] The Morphologically Annotated Part of BulTreeBank
[2] Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0
[3] Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0
[4] CMC training corpus Janes-Tag 2.0
[5] Training corpus jos1M 1.1
[6] BNC Sampler
[7] Corpus of morphologically disambiguated Estonian texts

The following 6 corpora are available for download only:

[1] MULTEXT-East "1984" annotated corpus 4.0
[2] xLiMe Twitter Corpus XTC 1.0.1
[3] Szeged Corpus 2.0
[4] NKJP1M
[5] CINTIL-Corpus Internacional do Português
[6] Lithuanian morphologically annotated corpus - MATAS

The following 12 languages are represented:

[1] Bulgarian (1 corpus)
[2] Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovenian Serbian (1 corpus)
[3] Croatian (1 corpus)
[4] English (1 corpus)
[5] Estonian (1 corpus)
[6] Hungarian (1 corpus)
[7] Lithuanian (1 corpus)
[8] German, Italian, Spanish (1 corpus)
[9] Polish (1 corpus)
[10] Portuguese (1 corpus)
[11] Serbian (2 corpora)
[12] Slovenian (1 corpus)

The corpora are between 3495 (B4 Heliand) and 2 million (BNC Sampler) tokens in size. Aside from manual PoS/MSD tags, the corpora are also manually annotated for the following 3 layers:

[1] Lemmatisation (6 corpora)
[2] Sentence segmentation (6 corpora)
[3] Named Entity recognition (6 corpora)

Information on licence is missing for the following corpus: Szeged Corpus 2.0.

5 of the corpora are available under CC-BY licences, the rest under miscellaneous licences.

Table 2: Corpora with manual PoS/MSD-tagging[1]

Corpus	Language	Description
MULTEXT-East "1984" annotated corpus 4.0 **Size:** 80,000 sentences, 1 million words **Annotation (purpose):** morphosyntactic tagging, lemmatisation, sentence alignment **Licence:** CC BY-NC-SA 4.0	Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovenian	This corpus contains 11 human translations of George Orwell's *Nineteen Eighty-Four,* as well as the original text. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available for download from the CLARIN.SI repository.
The Morphologically Annotated Part of BulTreeBank **Size:** 214,000 tokens **Annotation (purpose):** morphosyntactic tagging **Licence:** MS-NC-NoReD	Bulgarian	This corpus is available for download from META-SHARE and through the concordancer *Corpuscle.*
Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0 **Size:** 89,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Croatian	This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
BNC Sampler **Size:** 2 million tokens **Annotation (purpose):** PoS tagging **Licence:** BNC Licence	English	The corpus was manually post-edited to correct the PoS tags automatically assigned by CLAWS. The corpus is available for download from the Oxford Text Archive and through the CQPWeb concordancer (free registration required).
Corpus of morphologically disambiguated Estonian texts **Size:** 513,000 Tokens **Annotation (purpose):** morphological disambiguation **Licence:** CLARIN_ACA-NC	Estonian	This corpus contains texts from the 1980s subcorpus of the Corpus of Written Estonian 1890-1990. The corpus is available for download from a dedicated page provided by CELR as well as through the concordancer Korp.
xLiMe Twitter Corpus XTC 1.0.1 **Size:** 364,000 tokens **Annotation (purpose):** PoS tagging, Named Entity recognition, sentiment analysis **Licence:** MIT License	German, Italian, Spanish	This corpus contains Tweets. The corpus is available for download from the CLARIN.SI repository.
Szeged Corpus 2.0 **Size:** 1.5 million tokens **Annotation (purpose)**: morphosyntactic tagging	Hungarian	This corpus is available for download from a dedicated webpage.

[1] The label MSD stands for "morphosyntactic description" and denotes a fine-grained feature-structure based PoS tag, used to account for the rich inflectional paradigms like those in Slavic languages. In this report, we use the descriptor "morphosyntactic tagging" in lieu of "PoS-tagging" whenever we wish to specify that a corpus contains morphological mark-up in addition to the PoS information.

Lithuanian morphologically annotated corpus - MATAS **Size:** 1.6 million words **Annotation (purpose):** morphosyntactic tagging **Licence:** CLARIN ACA	Lithuanian	The corpus contains texts from various domains (documents, fiction, periodicals, scientific texts, wordforms). This corpus is available for download from the CLARIN-LT repository.
NKJP1M **Size:** 1 million tokens **Annotation (purpose):** morphosyntactic tagging **Licence:** GNU GPL 3	Polish	This corpus is a manually annotated subset of the National Corpus of Polish. The corpus is available for download from the Computational Linguistics in Poland website.
CINTIL-Corpus Internacional do Português **Size:** 1 million tokens **Annotation (purpose):** morphosyntactic tagging, Named Entity recognition **Licence:** CLARIN RES	Portuguese	The corpus contains transcriptions of spoken communication as well as written texts from several genres (news, literature, magazines, etc.). The corpus is available for download from the ELRA Catalogue.
Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0 **Size:** 92,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Serbian	This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
CMC training corpus Janes-Tag 2.0 **Size:** 75,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains computer-mediated communication (CMC). The corpus is morphosyntactically tagged following the MULTEXT-East Version 5 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Training corpus jos1M 1.1 **Size:** 1 million words **Annotation (purpose):** morphosyntactic tagging and lemmatisation **Licence:** CC BY-NC 4.0	Slovenian	This corpus contains sampled paragraphs from the Slovenian national corpus FidaPLUS. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available for download from the CLARIN.SI repository.

3.2. Lemmatisation

There are 7 corpora with manual lemmatisation in the CLARIN infrastructure, presented in detail in Table 3. For reasons of space, we list only those lemmatised corpora that are *not* fully syntactically parsed in this Section; for a list of corpora with both manual lemmatisation and syntactic parsing, see Section 3.3.

The following 4 corpora are available for download and through a concordance:

[1] Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0
[2] Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0
[3] CMC training corpus Janes-Tag 2.0
[4] Training corpus ssj500k 2.1

The following 3 corpora are available for download only:

[1] MULTEXT-East "1984" annotated corpus 4.0
[2] Training corpus hr500k 1.0
[3] Training corpus jos1M 1.1

The following 4 languages are represented:

[1] Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovenian Serbian (1 corpus)
[2] Croatian (2 corpora)
[3] Serbian (1 corpus)
[4] Slovenian (3 corpora)

The corpora are between 75,000 (CMC training corpus Janes-Tag 2.0) and 1 million (MULTEXT-East "1984" annotated corpus 4.0, Training corpus jos1M 1.1) tokens in size.

All the corpora with manual lemmatisation are also manually annotated with PoS/MSD-tags, so they are subset of the corpora in section 3.3.

All the corpora are available under CC-BY licences.

Table 3: Corpora with manual lemmatisation

Corpus	Language	Description
MULTEXT-East "1984" annotated corpus 4.0 **Size:** 80,000 sentences, 1 million words **Annotation (purpose):** morphosyntactic tagging, lemmatisation, sentence alignment **Licence:** CC BY-NC-SA 4.0	Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovenian	This corpus contains 11 human translations of George Orwell's *Nineteen Eighty-Four,* as well as the original text. The corpus is available for download from the CLARIN.SI repository.
Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0 **Size:** 89,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Croatian	This corpus contains Tweets. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Training corpus hr500k 1.0 **Size:** 500,000 tokens **Annotation (purpose):** tokenisation,	Croatian	The corpus is available through the concordancers KonText and noSketchEngine

sentence segmentation, morphosyntactic tagging, lemmatisation and Named Entity recognition. Half of the corpus also syntactically parsed **Licence:** CC BY-SA 4.0		and for download from the CLARIN.SI repository.
Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0 **Size:** 92,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Serbian	This corpus contains Tweets. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
CMC training corpus Janes-Tag 2.0 **Size:** 75,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains computer-mediated communication (CMC). The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Training corpus jos1M 1.1 **Size:** 1 million words **Annotation (purpose):** morphosyntactic tagging and lemmatisation **Licence:** CC BY-NC 4.0	Slovenian	This corpus contains sampled paragraphs from the Slovenian national corpus FidaPLUS. The corpus is available for download from the CLARIN.SI repository.
Training corpus ssj500k 2.1 **Size:** 586,000 tokens **Annotation (purpose):** fully – tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. Half of the corpus – syntactic parsing, Named Entity recognition, and verbal multiword expression tagging. Quarter: semantic roles **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains standard Slovenian. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

3.3. Syntactic parsing

There are 35 corpora and corpus collections with manual syntactic parsing in the CLARIN infrastructure, presented in detail in Table 4. In addition, there are 3 large collections of treebanks in the infrastructure (*Treebanks of INESS*, *Universal Dependencies*, and *HamleDT*), which are listed at the end of the Table with links for searching within the individual treebanks.

The following 10 corpora and corpus collections are available for download and through a concordancer:

[1] Training corpus hr500k 1.0
[2] Czech Legal Text Treebank 2.0
[3] Prague Czech-English Dependency Treebank 2.0 Coref
[4] Prague Discourse Treebank 2.0
[5] Training corpus SETimes.SR 1.0
[6] FicTree 1.0
[7] HamleDT 3.0
[8] Treebanks of INESS
[9] Universal Dependencies 2.3
[10] Prague Dependency Treebank 3.5

The following 25 corpora and corpus collections are available for download only:

[1] Prague Arabic Dependency Treebank 1.0
[2] Slovak Dependency Treebank
[3] Artificial Treebank with Ellipsis
[4] Lassy Klein-corpus
[5] Estonian Treebank
[6] UD Estonian ver.2.3
[7] Turku Dependency Treebank
[8] Polish Dependency Bank in Universal Dependency format
[9] CINTIL DependencyBank
[10] CINTIL TreeBank
[11] CINTIL-DeepBank
[12] CINTIL-PropBank
[13] Tamil Dependency Treebank v0.1
[14] SoNaR-1
[15] Syntactic Reference Corpus of Medieval French
[16] B4 Heliand
[17] Dependency-Annotated Subset of the CREG Corpus
[18] Szeged Treebank 2.0
[19] Lithuanian Treebank ALKSNIS
[20] Tübingen Treebank of Written German / Newspaper Corpus (TüBa-D/Z)
[21] TimeML annotated corpus of Estonian newspaper articles
[22] Finnish TreeBank 1
[23] Finnish TreeBank 2
[24] GRUG Parallel Treebank
[25] Training corpus ssj500k 2.1

The individual corpora represent the following 18 languages:

[1] Arabic (1 corpus)
[2] Croatian (1 corpus)
[3] Czech (5 corpora)
[4] Czech, English (1 corpus)
[5] Czech, English, Finnish, Russian, Slovak (1 corpus)
[6] Dutch (2 corpora)
[7] Estonian (3 corpora)
[8] French (1 corpus)
[9] Finnish (3 corpora)
[10] Georgian, Ukranian, Russian, German (1 corpus)
[11] German (3 corpora)
[12] Hungarian (1 corpus)

[13] Lithuanian (1 corpus)
[14] Polish (1 corpus)
[15] Portuguese (4 corpora)
[16] Serbian (1 corpus)
[17] Slovenian (1 corpus)
[18] Tamil (1 corpus)

In addition, the treebank collection HamleDT 3.0 represents 19 languages (full list), the Treebanks of INESS represent 72 languages, and the treebanks in Universal Dependencies 2.3 represent 75 languages (full list).

Size overview (for corpora and corpus collections that have information on token number available):

- 15 contain <1 million tokens;
- 4 contain between 1 and 10 million tokens;
- 1 contains >10 million tokens

In addition to manual syntactic parsing, which characterizes all the corpora in this Section, the following corpora and corpus collections display additional annotation layers:

[1] Training corpus hr500k 1.0 (tokenisation, lemmatisation and Named Entity recognition)
[2] Prague Discourse Treebank 2.0 (mark-up of discourse phenomena enriched by the annotation of secondary connectives)
[3] Artificial Treebank with Ellipsis (mark-up of ellipsis)
[4] CINTIL-DeepBank (Logical Forms in the sense of Formal Semantics)
[5] Training corpus SETimes.SR 1.0 (tokenisation, sentence segmentation, lemmatisation, Named Entity recognition)
[6] Training corpus ssj500k 2.1 (tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation, Named Entity recognition, verbal multiword expression tagging)

Information on the licence is missing for Szeged Treebank 2.0.

The corpora and corpus collections are available under various licences: 11 under CC-BY, 4 under ELRA licences, the rest miscellaneous.

Table 4: Corpora with manual syntactic parsing

Corpus	Language	Description
Prague Arabic Dependency Treebank 1.0 **Size:** 113,500 tokens **Annotation (purpose):** syntactic parsing and morphosyntactic tagging **Licence:** CC BY-NC-SA 3.0	Arabic	The corpus is available for download from the LINDAT repository.
Training corpus hr500k 1.0 **Size:** 500,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, morphosyntactic tagging, lemmatisation and Named Entity recognition. Half of the corpus also syntactically parsed **Licence:** CC BY-SA 4.0	Croatian	The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Czech Legal Text Treebank 2.0 **Size:** 1121 sentences **Annotation (purpose):** syntactic parsing, labelling of semantic entities **Licence:** CC BY-NC-SA 4.0	Czech	This corpus contains legal texts. The corpus is available through the concordance KonText, the PML-TQ tool and for download from the LINDAT repository.
FicTree 1.0	Czech	This corpus contains fictional texts.

Size: 12760 sentences **Annotation (purpose)**: syntactic parsing and morphosyntactic tagging **Licence:** CC BY-NC-SA 4.0		The corpus is available for download from LINDAT and through the concordancer KonText.
Prague Dependency Treebank 3.5 **Size:** 2 million words **Annotation (purpose):** syntactic parsing and morphosyntactic tagging **Licence:** CC BY-NC-SA 4.0	Czech	The corpus is manually annotated at several levels – aside from syntactic parsing and morphological information, it is annotation for sentence information structure, multiword expression, coreference, bridging relations and discourse relations. The corpus is available for download from the LINDAT repository and through the concordancer KonText.
Prague Discourse Treebank 2.0 **Size:** 49,500 sentences **Annotation (purpose):** syntactic parsing, mark-up of discourse phenomena enriched by the annotation of secondary connectives **Licence:** CC-BY	Czech	This corpus is a subset of the Prague Dependency Treebank 3.5. The corpus is available through the PML-TQ tool.
Slovak Dependency Treebank **Size:** 106,000 tokens, 10,600 sentences **Annotation (purpose):** syntactic parsing **Licence:** CC BY-SA 4.0	Czech	The syntactic parsing is modelled after the Prague Dependency Treebank. The corpus is available for download from the LINDAT repository.
Prague Czech-English Dependency Treebank 2.0 Coref **Size:** 49,000 sentences **Annotation (purpose):** syntactic parsing, mark-up of coreference **Licence:** CC-BY-NC-SA + LDC99T42 (restricted use)	Czech, English	This corpus is an extended version of Prague Czech-English Dependency Treebank 2.0, with added mark-up of coreference. The syntactic parsing follows the PDT 2.0 style. The corpus is available for download from the LINDAT repository. The version without coreference annotation is available through the concordancer KonText and the PML-TQ tool.
Artificial Treebank with Ellipsis **Size:** 106,000 tokens, 10,604 sentences **Annotation (purpose):** syntactic parsing, mark-up of elliptical constructions **Licence:** Licence Universal dependencies v2.1	Czech, English, Finnish, Russian, Slovak	The syntactic parsing follows the Universal Dependencies schema. The corpus is available for download from the LINDAT repository.
Lassy Klein-corpus **Size:** 1 million tokens **Annotation (purpose):** PoS tagging, syntactic parsing, lemmatisation **Licence:** Agreement	Dutch	The corpus is available for download from the Dutch Language Institute, and through the online environments PaQu and GrETEL.
SoNaR-1 **Size:** 1 million words	Dutch	This is a manually annotated subset of the much larger (approx. 500 million) word) SoNaR corpus.

Annotation (purpose): PoS tagging, syntactic parsing, semantic role labelling		The corpus is available for download from the Dutch Language Institute.
Estonian Treebank **Size:** 1,000 sentences **Annotation (purpose):** syntactic parsing **Licence:** CLARIN_ACA	Estonian	The corpus contains fictional and newspaper texts. The corpus is available for download from META-SHARE (CELR distribution).
UD Estonian ver.2.3 **Size:** 434,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** CC-BY-SA	Estonian	This corpus contains fictional, newspaper and scientific texts. The syntactic parsing follows the Universal Dependencies schema. The corpus is available for download from META-SHARE (CELR distribution).
TimeML annotated corpus of Estonian newspaper articles **Size:** 22,000 words **Annotation (purpose):** morphosyntactic tagging and syntactic parsing **Licence:** CC-BY-SA	Estonian	This corpus contains newspaper articles. The corpus is available for download from META-SHARE (CELR distribution).
Finnish TreeBank 1 **Size:** 160,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** CC-BY 3.0	Finnish	This corpus contains 19,000 sentences from the Large Grammar of Finnish. The corpus is available for download from the Language Bank of Finland.
Finnish TreeBank 2 **Size:** 160,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** CC-BY 3.0	Finnish	This corpus contains 19,000 sentences from the Large Grammar of Finnish. The corpus is available for download from the Language Bank of Finland.
Turku Dependency Treebank **Size:** 204,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** CC-BY-SA	Finnish	The syntactic parsing follows the Universal Dependencies schema. The corpus is available for download from the Turku BioNLP Group.
Syntactic Reference Corpus of Medieval French **Size:** 245,000 words **Annotation (purpose)**: syntactic parsing **Licence:** CLARIN ACA	French	This corpus contains Old French texts. The corpus is available for download from the IMS CLARIN-D repository.
GRUG Parallel Treebank **Size:** 10,400 sentence pairs **Annotation (purpose)**: syntactic parsing, PoS tagging **Licence:** CC-BY	Georgian, Ukranian, Russian, German	The corpus is syntactically parsed following the TIGER guidelines. The corpus is available for download from a dedicated website provided by the CLARIN-D consortium as well as through TIGERSearch.
B4 Heliand **Size:** 3495 tokens **Annotation (purpose)**: PoS tagging, syntactic parsing	German	This corpus contains historical German texts. The corpus is available for download from the HZSK repository.

Licence: CC-BY		
Dependency-Annotated Subset of the CREG Corpus **Size:** 109 sentences **Annotation (purpose):** PoS tagging, syntactic parsing **Licence:** CLARIN RES	German	This corpus consists of answers to reading comprehension questions written by American college students learning German. The corpus is available for download from the Tübingen CLARIN Repository.
Tübingen Treebank of Written German / Newspaper Corpus (TüBa-D/Z) **Size:** 1.9 million tokens **Annotation (purpose):** syntactic parsing **Licence:** CLARIN RES	German	This corpus contains newspaper articles. The corpus is available for download from the Tübingen CLARIN Repository.
Szeged Treebank 2.0 **Size:** 82,000 sentences **Annotation (purpose):** syntactic parsing	Hungarian	This corpus is available for download from a dedicated webpage.
Lithuanian Treebank ALKSNIS **Size:** 2,355 sentences **Annotation (purpose):** syntactic parsing **Licence:** CLARIN PUB	Lithuanian	Syntactic parsing follows the rules of the Prague Dependency Treebank schema. This corpus is available for download
Polish Dependency Bank in Universal Dependency format **Size:** 22,000 trees, 351,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** CC BY-NC-SA 4.0	Polish	This corpus also contains sentences showing certain problematic syntactic phenomena – sentences with ellipsis, comparative constructions, constructions with the bi-functional subordinating conjunction *jako*, etc. The syntactic parsing follows the Universal Dependencies schema. The corpus is available for download from the Computational Linguistics in Poland website.
CINTIL DependencyBank **Size:** 110,000 tokens **Annotation (purpose):** morphosyntactic tagging and syntactic parsing **Licence:** ELRA END USER	Portuguese	This corpus contains literary and newspaper texts. The corpus is available for download from the ELRA catalogue.
CINTIL TreeBank **Size:** 110,000 tokens **Annotation (purpose):** syntactic parsing **Licence:** ELRA END USER	Portuguese	This corpus contains literary and newspaper texts. The corpus is available for download from the ELRA catalogue.
CINTIL-DeepBank **Size:** 110,000 tokens **Annotation (purpose):** PoS-tagging, syntactic parsing, grammatical functions, logical forms **Licence:** ELRA END USER	Portuguese	This corpus contains literary and newspaper texts. The corpus is available for download from the ELRA catalogue.
CINTIL-PropBank	Portuguese	This corpus contains literary and newspaper texts.

Size: 110,000 tokens **Annotation (purpose):** syntactic parsing and phrase semantic roles **Licence:** ELRA END USER		The corpus is available for download from the ELRA catalogue.
Training corpus SETimes.SR 1.0 **Size:** 87,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, morphosyntactic tagging, lemmatisation, syntactic parsing, and Named Entity recognition **Licence:** CC BY-SA 4.0	Serbian	This corpus contains posts from the Southeast European Times news portal, which is now defunct. The syntactic parsing follows the Universal Dependencies framework. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Training corpus ssj500k 2.1 **Size:** 586,000 tokens **Annotation (purpose):** fully – tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. Half of the corpus – syntactic parsing, Named Entity recognition, and verbal multiword expression tagging. Quarter: semantic roles **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains standard Slovenian. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Tamil Dependency Treebank v0.1 **Size:** 600 sentences **Annotation (purpose):** syntactic parsing and morphosyntactic tagging **Licence:** CC BY-NC-SA 3.0	Tamil	The syntactic parsing follows the rules of the Prague Dependency Treebank. The corpus is available for download from the LINDAT repository.
HamleDT 3.0 **Size:** 19 treebanks **Annotation (purpose):** syntactic parsing and morphosyntactic tagging **Licence:** HamleDT 3.0 Licence Terms	19 languages	This treebank collection is available for download from LINDAT. The treebanks can be individually queried through KonText and the treebank tool PML-TQ. We list them here by language: 1. Arabic (KonText, PML-TQ) 2. Bengali (KonText) 3. Catalan (KonText) 4. Czech (KonText, PML-TQ) 5. Dutch (KonText, PML-TQ) 6. English (KonText) 7. Estonian (KonText, PML-TQ) 8. German (KonText) 9. Greek (KonText) 10. Hindi (KonText) 11. Latin (KonText, PML-TQ) 12. Persian (KonText, PML-TQ) 13. Polish (KonText, PML-TQ) 14. Portuguese (KonText, PML-TQ) 15. Romanian (KonText, PML-TQ) 16. Russian (KonText) 17. Slovenian (KonText, PML-TQ) 18. Spanish (KonText)

		19. Tamil (KonText, PML-TQ)
Treebanks of INESS **Size:** 532 treebanks **Annotation (purpose)**: syntactic parsing **Licence:** CC-BY	71 languages	This is a collection of treebanks made available through the *Infrastructure for the Exploration of Syntax and Semantics* (INESS). The corpora are available for online querying through INESS.
Universal Dependencies 2.3 **Size:** 18 million tokens **Annotation (purpose):** morphosyntactic tagging, syntactic parsing **Licence:** Licence Universal Dependencies v2.3 publicly available	75 languages	This corpus collection contains 126 treebanks. The corpus collection is available for download from the LINDAT repository. The individual treebanks in Universal Dependencies 2.3 can also be queried through the concordancer KonText and the treebank query tool PML-TQ. Below we provide links to these search environments for all the treebanks. For a detailed description of all the treebanks, see the Universal Dependencies project page. 1. UD_Akkadian-PISANDUB - KonText 2. UD_Amharic-ATT - KonText, PML-TQ 3. UD_Armenian-ArmTDP - KonText, PML-TQ 4. UD_Breton-KEB - KonText, PML-TQ 5. UD_Buryat-BDT - KonText, PML-TQ 6. UD_Cantonese-HK - KonText, PML-TQ 7. UD_Chinese-HK - KonText, PML-TQ 8. UD_Chinese-CFL - KonText, PML-TQ 9. UD_Coptic-Scriptorium - KonText, PML-TQ 10. UD_Croatian-SET - KonText, PML-TQ 11. UD_English-ESL - KonText, PML-TQ 12. UD_Faroese-OFT - KonText, PML-TQ 13. UD_Galician-TreeGal - KonText, PML-TQ 14. UD_Hindi_English-HIENCS - KonText 15. UD_Kazakh-KTB 2.2 - KonText, PML-TQ 16. UD_Komi_Zyrian-Lattice - KonText, PML-TQ 17. UD_Komi_Zyrian-IKDP

		- KonText, PML-TQ 18. UD_Kurmanji-MG - KonText, PML-TQ 19. UD_Lithuanian-HSE - KonText, PML-TQ 20. UD_Maltese-MUDT - KonText, PML-TQ 21. UD_Marathi-UFAL - KonText, PML-TQ 22. UD_Naija-NSC - KonText, PML-TQ 23. UD_Persian-Seraji - KonText, PML-TQ 24. UD_Russian-Taiga - KonText, PML-TQ 25. UD_Sanskrit-UFAL - KonText, PML-TQ 26. UD_Serbian-SET - KonText, PML-TQ 27. UD_Slovenian-SST - KonText, PML-TQ 28. UD_Tagalog-TRG - KonText, PML-TQ 29. UD_Telugu-MTG - KonText, PML-TQ 30. UD_Ukrainian-IU - KonText, PML-TQ 31. UD_Upper_Sorbian-UFAL - KonText, PML-TQ 32. UD_Uyghur-UDT - KonText, PML-TQ 33. UD_Warlpiri-UFAL - KonText, PML-TQ 34. UD_Yoruba-YTB - KonText, PML-TQ 35. UD_Afrikaans-AfriBooms - KonText 36. UD_Ancient_Greek-PROIEL - KonText 37. UD_Ancient_Greek-Perseus - KonText, PML-TQ 38. UD_Arabic-PADT - KonText, PML-TQ 39. UD_Arabic-PUD - KonText, PML-TQ 40. UD_Arabic-NYUAD - KonText 41. UD_Bambara-CRB - KonText, PML-TQ 42. UD_Basque-BDT - KonText, PML-TQ 43. UD_Belarusian-HSE - KonText, PML-TQ 44. UD_Bulgarian-BTB - KonText, PML-TQ 45. UD_Catalan-AnCora - KonText, PML-TQ

		46. UD_Chinese-GSD - KonText, PML-TQ 47. UD_Chinese-PUD - KonText, PML-TQ 48. UD_Czech-PDT - KonText, PML-TQ 49. UD_Czech-CAC - KonText, PML-TQ 50. UD_Czech-FicTree - KonText, PML-TQ 51. UD_Czech-PUD - KonText, PML-TQ 52. UD_Czech-CLTT - KonText, PML-TQ 53. UD_Danish-DDT - KonText, PML-TQ 54. UD_Dutch-Alpino - KonText, PML-TQ 55. UD_Dutch-LassySmall - KonText, PML-TQ 56. UD_English-ParTUT - KonText, PML-TQ 57. UD_English-GUM - KonText, PML-TQ 58. UD_English-EWT - KonText, PML-TQ 59. UD_English-PUD - KonText, PML-TQ 60. UD_English-LinES - KonText, PML-TQ 61. UD_Erzya-JR - KonText, PML-TQ 62. UD_Finnish-FTB - KonText, PML-TQ 63. UD_Finnish-TDT - KonText, PML-TQ 64. UD_Finnish-PUD - KonText, PML-TQ 65. UD_French-ParTUT - KonText, PML-TQ 66. UD_French-GSD - KonText, PML-TQ 67. UD_French-Sequoia - KonText, PML-TQ 68. UD_French-Spoken - KonText, PML-TQ 69. UD_French-PUD - KonText, PML-TQ 70. UD_French-FTB - KonText 71. UD_Galician-CTG - KonText, PML-TQ 72. UD_German-GSD - KonText, PML-TQ 73. UD_German-PUD - KonText, PML-TQ 74. UD_Gothic-PROIEL

		- KonText, PML-TQ 75. UD_Greek-GDT - KonText, PML-TQ 76. UD_Hebrew-HTB - KonText, PML-TQ 77. UD_Hindi-HDTB - KonText, PML-TQ 78. UD_Hindi-PUD - KonText, PML-TQ 79. UD_Hungarian-Szeged - KonText, PML-TQ 80. UD_Indonesian-GSD - KonText, PML-TQ 81. UD_Indonesian-PUD - KonText, PML-TQ 82. UD_Irish-IDT - KonText, PML-TQ 83. UD_Italian-ISDT - KonText, PML-TQ 84. UD_Italian-ParTUT - KonText, PML-TQ 85. UD_Italian-PUD - KonText, PML-TQ 86. UD_Japanese-GSD KonText, PML-TQ 87. UD_Japanese-PUD - KonText, PML-TQ 88. UD_Japanese-Modern - KonText, PML-TQ 89. UD_Korean-Kaist - KonText, PML-TQ 90. UD_Korean-GSD - KonText, PML-TQ 91. UD_Korean-PUD - KonText, PML-TQ 92. UD_Latin-PROIEL - KonText, PML-TQ 93. UD_Latin-ITTB - KonText, PML-TQ 94. UD_Latin-Perseus - KonText, PML-TQ 95. UD_Latvian-LVTB - KonText, PML-TQ 96. UD_North_Sami-Giella - KonText, PML-TQ 97. UD_Norwegian-Bokmaal - KonText, PML-TQ 98. UD_Norwegian-Nynorsk - KonText, PML-TQ 99. UD_Norwegian-NynorskLIA - KonText, PML-TQ 100. UD_Old_Church_Slavonic-PROIEL - KonText, PML-TQ 101. UD_Old_French-SRCMF - KonText, PML-TQ 102. UD_Polish-LFG - KonText, PML-TQ

		103. UD_Polish-SZ - KonText, PML-TQ 104. UD_Portuguese-Bosque - KonText, PML-TQ 105. UD_Portuguese-GSD - KonText, PML-TQ 106. UD_Portuguese-PUD - KonText, PML-TQ 107. UD_Romanian-RRT - KonText, PML-TQ 108. UD_Romanian-Nonstandard - KonText, PML-TQ 109. UD_Russian-GSD - KonText, PML-TQ 110. UD_Russian-PUD - KonText, PML-TQ 111. UD_Russian-SynTagRus - KonText, PML-TQ 112. UD_Slovak-SNK - KonText, PML-TQ 113. UD_Slovenian-SSJ - KonText, PML-TQ 114. UD_Spanish-AnCora - KonText, PML-TQ 115. UD_Spanish-GSD - KonText, PML-TQ 116. UD_Spanish-PUD - KonText, PML-TQ 117. UD_Swedish-Talbanken - KonText, PML-TQ 118. UD_Swedish-LinES - KonText, PML-TQ 119. UD_Swedish-PUD - KonText, PML-TQ 120. UD_Swedish_Sign_Language-SSLC - KonText, PML-TQ 121. UD_Tamil-TTB - KonText, PML-TQ 122. UD_Thai-PUD - KonText, PML-TQ 123. UD_Turkish-IMST - KonText, PML-TQ 124. UD_Turkish-PUD - KonText, PML-TQ 125. UD_Urdu-UDTB - KonText, PML-TQ 126. UD_Vietnamese-VTB - KonText, PML-TQ

3.4. Named Entity recognition

There are 11 corpora with manual Named Entity recognition in the CLARIN infrastructure, presented in detail in Table 5.

The following 6 corpora are available for download and through a concordancer:

[1] Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0
[2] Training corpus hr500k 1.0
[3] Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0
[4] Training corpus SETimes.SR 1.0
[5] CMC training corpus Janes-Tag 2.0
[6] Training corpus ssj500k 2.1

The following 5 corpora are available for download only:

[1] Czech Named Entity Corpus 1.1
[2] xLiMe Twitter Corpus XTC 1.0.1
[3] KPWr (Polish Corpus of Wrocław University of Technology) 1.2
[4] Polish Spatial Texts 1.0
[5] CINTIL-Corpus Internacional do Português

The following 8 languages are represented:

[1] Croatian (2 corpora)
[2] Czech (1 corpus)
[3] English (1 corpus)
[4] German, Italian, Spanish (1 corpus)
[5] Polish (2 corpora)
[6] Portuguese (1 corpus)
[7] Serbian (2 corpora)
[8] Slovenian (2 corpora)

The corpora are between 46,000 (Polish Spatial Texts 1.0) and 1 million (CINTIL-Corpus Internacional do Português) tokens in size.

Aside from manually assigned Named Entity labels, the corpora are also manually annotated for the following layers:

[1] PoS/MSD tagging (7 corpora)
[2] Lemmatisation (6 corpora)
[3] Syntactic parsing (2 corpora)

9 of the corpora with manual Named Entity recognition are available under CC-BY licences, 1 under the MIT Licence and 1 under CLARIN RES.

Table 5: Corpora with manual Named Entity recognition

Corpus	Language	Description
Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0 **Size:** 89,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Croatian	This corpus contains Tweets. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

Training corpus hr500k 1.0 **Size:** 500,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, morphosyntactic tagging, lemmatisation and Named Entity recognition. Half of corpus also syntactically parsed **Licence:** CC BY-SA 4.0	Croatian	The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Czech Named Entity Corpus 1.1 **Size:** 5868 sentences, 35220 NEs **Annotation (purpose):** Named Entity recognition **Licence:** CC BY-NC-SA 3.0	Czech	The corpus is available for download from LINDAT.
xLiMe Twitter Corpus XTC 1.0.1 **Size:** 364,000 tokens **Annotation (purpose):** PoS tagging, Named Entity recognition, sentiment analysis **Licence:** MIT License	German, Italian, Spanish	This corpus contains Tweets. The corpus is available for download from the CLARIN.SI repository.
KPWr (Polish Corpus of Wrocław University of Technology) 1.2 **Size:** 447,000 tokens **Annotation (purpose):** chunks and selected predicate-argument relations, Named Entity recognition, relations between named entities, anaphora relations, word senses, events, temporal expressions, spatial relations between entities, keywords and semantic roles within nominal and adjective phrases **Licence:** CC BY-SA 3.0	Polish	This corpus contains texts in a variety of domains (blogs, science, stenographic recordings, etc.). The corpus is available for download from the CLARIN-PL repository.
Polish Spatial Texts 1.0 **Size:** 46,000 tokens **Annotation (purpose):** Named Entity recognition (spatial expressions) **Licence:** CC BY-SA 4.0	Polish	This corpus contains travel blogs. The corpus is available for download from the CLARIN-PL repository
CINTIL-Corpus Internacional do Português **Size:** 1 million tokens **Annotation (purpose):** morphosyntactic tagging, Named Entity recognition **Licence:** CLARIN RES	Portuguese	The corpus contains transcriptions of spoken communication as well as written texts from several genres (news, literature, magazines, etc.). The corpus is available for download from the ELRA Catalogue.

Training corpus SETimes.SR 1.0 **Size:** 87,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, morphosyntactic tagging, lemmatisation, syntactic parsing, and Named Entity recognition **Licence:** CC BY-SA 4.0	Serbian	This corpus contains posts from the Southeast European Times news portal, which is now no longer being updated. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0 **Size:** 92,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY 4.0	Serbian	This corpus contains Tweets. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
CMC training corpus Janes-Tag 2.0 **Size:** 75,000 tokens **Annotation (purpose):** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains computer-mediated communication (CMC). The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Training corpus ssj500k 2.1 **Size:** 586,000 tokens **Annotation (purpose):** fully – tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. Half of the corpus – syntactic parsing, Named Entity recognition, and verbal multiword expression tagging. Quarter of corpus: semantic roles **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains standard Slovenian. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

3.5. Sentiment analysis

There are 7 corpora with manual sentiment analysis in the CLARIN infrastructure, presented in detail in Table 6.

6 corpora are available for download only, whereas Facebook Data for Sentiment Analysis is available for download and through a concordancer.

The following 6 languages are represented:

[1] German, Italian, Spanish (1 corpus)
[2] Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish (1 corpus)
[3] Croatian (1 corpus)
[4] Czech (2 corpora)
[5] Norwegian (1 corpus)
[6] Slovenian (1 corpus)

The corpora are between 364,000 (xLiMe Twitter Corpus XTC 1.0.1) and 1.6 million (Twitter sentiment for 15 European languages) tokens in size.

Aside from manually assigned sentiment labels, which characterize all the corpora in this section, the xLiMe Twitter Corpus XTC 1.0 is also manually annotated with PoS tags and Named Entity labels.

6 of the corpora are available under CC-BY licences, 1 under the MIT Licence and 1 under the PARSEME Shared Task Data Agreement licence.

Table 6: Corpora with manual sentiment analysis

Corpus	Language	Description
xLiMe Twitter Corpus XTC 1.0.1 **Size:** 364,000 tokens **Annotation (purpose):** PoS tagging, Named Entity recognition, sentiment analysis **Licence:** MIT License	German, Italian, Spanish	This corpus contains Tweets. The corpus is available for download from the CLARIN.SI repository.
Twitter sentiment for 15 European languages **Size:** 1.6 million tweets **Annotation (purpose):** sentiment analysis **Licence:** CC BY-SA 4.0	Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish	This corpus contains Tweet IDs with sentiment annotations. The corpus is available for download from the CLARIN.SI repository.
Dataset and baseline model of moderated content FRENK-STYRIA-24sata 1.0 **Size:** 407.5 million words **Annotation (purpose):** sentiment analysis (socially unacceptable discourse) **Licence:** CC BY-SA 4.0	Croatian	This corpus contains news comments from the website 24sata.hr. The corpus is available for download from CLARIN.SI.
Aspect-Term Annotated Customer Reviews in Czech	Czech	This corpus contains online user-product reviews.

Size: 2200 reviews **Annotation (purpose):** sentiment analysis **Licence:** CC BY-NC-SA 3.0		The corpus is available for download from LINDAT.
Facebook Data for Sentiment Analysis **Size**: 10,000 Facebook posts **Annotation (purpose)**: sentiment analysis **Licence:** CC BY-SA 3.0	Czech	This corpus contains Facebook posts. The corpus is available for download from LINDAT and through the concordancer KonText.
NoReC: The Norwegian Review Corpus **Size:** 14.8 million tokens **Annotation (purpose):** sentiment analysis **Licence:** CC BY-NC 3.0	Norwegian	This corpus contains reviews in different domains (e.g., literature, videogames, etc.). The corpus is available for download from the CLARINO repository.
Manually sentiment annotated Slovenian news corpus SentiNews 1.0 **Size:** 10,427 articles **Annotation (purpose):** sentiment analysis **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains news articles. The corpus is available for download from the CLARIN.SI repository.

3.6. Other annotation layers

There are 24 corpora and 1 corpus collection (Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1)) in the CLARIN infrastructure, presented in detail in Table 9, with the following other types of manual annotation

[1] Comma placement (1 corpus)
[2] Coreference/anaphora (4 corpora)
[3] Discourse phenomena and connectives (1 corpus)
[4] Elliptical constructions (1 corpus)
[5] Linked Data (1 corpus)
[6] Logical entailment (1 corpus)
[7] Identification of reported speech (1 corpus)
[8] Mark-up of terminology (3 corpora)
[9] Mark-up of Verbal Multi Word expressions (2 corpora)
[10] Semantic role labelling (2 corpora)
[11] Sentence relations – CST Theory (1 corpus)
[12] Summarisation (1 corpus)
[13] Temporal semantic annotations (1 corpus)
[14] Word normalisation (3 corpora)
[15] Word sense disambiguation (1 corpus)

The following corpora and corpus collection are available through a concordancer and for download:

[1] Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0
[2] Czech Legal Text Treebank 2.0
[3] Prague Discourse Treebank 2.0
[4] Prague Czech-English Dependency Treebank 2.0 Coref
[5] Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0
[6] CMC training corpus Janes-Tag 2.0
[7] Training corpus ssj500k 2.1
[8] Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1)
[9] The ACL RD-TEX 2.0

The following corpora are available for download only:

[1] Artificial Treebank with Ellipsis
[2] SoNaR-1
[3] Greek Coreference Corpus
[4] Greek Textual Entailment Corpus
[5] KPWr (Polish Corpus of Wrocław University of Technology) 1.2
[6] Grundtvig's Works Corpus
[7] Polish Summaries Corpus
[8] WUT Relations Between Sentences Corpus
[9] Corpus of comma placement Vejica 1.3
[10] Terminology identification dataset KAS-term 1.0
[11] Bilingual terminology extraction dataset KAS-biterm 1.0
[12] Speech, Thought and Writing Presentation Corpus
[13] TimeML annotated corpus of Estonian newspaper articles
[14] [12] Estonian Treebank annotated with coreference relations
[15] Semantically disambiguated corpus of Estonian

The following 14 languages are represented:

[1] Basque, Bulgarian, Croatian, English, French, German, Hebrew, Hindi, Hungarian, Italian, Lithuanian, Modern Greek, Persian, Polish, Portuguese, Romanian, Slovenian, Spanish, Turkish (1 corpus)
[2] Croatian (1 corpus)
[3] Czech (2 corpora)
[4] Czech, English (1 corpus)

[5] Czech, English, Finnish, Russian, Slovak (1 corpus=
[6] Danish (1 corpus)
[7] Dutch (1 corpus)
[8] English (2 corpora)
[9] Estonian (3 corpora)
[10] Greek (2 corpora)
[11] Polish (3 corpora)
[12] Serbian (1 corpus)
[13] Slovenian (4 corpora)
[14] Slovenian, English (1 corpus)

The corpus collection Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1) represents 16 languages.

The following corpus lacks information on size:

[1] Grundtvig's Works Corpus

The LR's in this section are between 78,500 (the corpus Bilingual terminology extraction dataset KAS-biterm 1.0) and 5.8 million (the corpus collection Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1)) tokens in size.

The following corpus lacks licence information: SoNaR-1.

19 corpora are available under CC-BY licences, the rest under miscellaneous. The corpus collection Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1) is available under PARSEME Shared Task Data Agreement licence.

Table 7: Corpora with other types of manual annotation

Corpus	Language	Description
Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0 **Size:** 89,000 tokens **Annotation (other type):** word normalisation **Licence:** CC BY 4.0	Croatian	This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Czech Legal Text Treebank 2.0 **Size:** 1121 sentences **Annotation (other type):** semantic role labelling **Licence:** CC BY-NC-SA 4.0	Czech	This corpus contains legal texts. The corpus is available through the concordance KonText, the PML-TQ tool and for download from the LINDAT repository.
Prague Discourse Treebank 2.0 **Size:** 49,500 sentences **Annotation (other type):** mark-up of discourse phenomena enriched by the annotation of secondary connectives **Licence:** CC-BY	Czech	This corpus is a subset of the Prague Dependency Treebank 3.5. The corpus is available through the PML-TQ tool.

The ACL RD-TEX 2.0 **Size:** 33216 tokens **Annotation (other type)**: terminology extraction/classification **Licence:** CC BY-NC-SA 4.0	English	This corpus contains 6818 terms extracted from abstracts of computational linguistics papers. The corpus is available for download from LINDAT and through KonText.
Speech, Thought and Writing Presentation Corpus **Size**: 260,000 words **Annotation (other type)**: identification of reported speech **Licence:** CC BY-NC-SA 3.0	English	This corpus contains literary, newspaper and biography texts. The corpus is available for download from the Oxford Text Archive.
TimeML annotated corpus of Estonian newspaper articles **Size:** 22,000 words **Annotation (other):** temporal semantic annotations **Licence:** CC-BY-SA	Estonian	This corpus contains newspaper articles. The corpus is available for download from META-SHARE (CELR distribution).
Estonian Treebank annotated with coreference relations **Size:** 107,000 words **Annotation (other)**: anaphora relations **Licence:** GPL	Estonian	This corpus contains newspaper texts plus one scientific medical text. The corpus is available for download from META-SHARE (CELR distribution).
Semantically disambiguated corpus of Estonian **Size:** 375,733 tokens **Annotation (other):** word sense disambiguation **Licence:** CLARIN ACA	Estonian	The corpus is available for download from META-SHARE.
Prague Czech-English Dependency Treebank 2.0 Coref **Size:** 49,000 sentences **Annotation (other type):** mark-up of coreference **Licence:** CC-BY-NC-SA + LDC99T42 (restricted use)	Czech, English	This corpus is an extended version of Prague Czech-English Dependency Treebank 2.0, with added mark-up of coreference. The syntactic parsing follows the PDT 2.0 style. The corpus is available for download from the LINDAT repository. The version without coreference annotation is available through the concordancer KonText and the PML-TQ tool.
Artificial Treebank with Ellipsis **Size:** 106,000 tokens, 10,604 sentences **Annotation (other type):** mark-up of elliptical constructions **Licence:** Licence Universal dependencies v2.1	Czech, English, Finnish, Russian, Slovak	The syntactic parsing follows the Universal Dependencies schema. The corpus is available for download from the LINDAT repository.
Grundtvig's Works Corpus **Annotation (other type):** linked data (places, persons, bible citations, etc.)	Danish	This corpus contains the literary works of the Danish bishop N.F.S Grundtvig.

Licence: CC BY-NC 4.0		The corpus is available for download from the CLARIN-DK repository.
SoNaR-1 **Size:** 1 million words **Annotation (other type)**: semantic role labelling	Dutch	This is a manually annotated subset of the much larger (approx. 500 million) word) SoNaR corpus. The corpus is available for download from the Dutch Language Institute.
Greek Coreference Corpus **Size:** 62,988 tokens **Annotation (other type)**: coreference **Licence:** CC-BY-NC-SA	Greek	In addition to coreference, the corpus is annotated for identity and bridging relations. The corpus is available for download from the clarin:el repository.
Greek Textual Entailment Corpus **Size:** 600 sentence-pairs **Annotation (other type)**: logical entailment **Licence:** CC-BY	Greek	This corpus contains texts from the domains of politics, law and travel. This corpus is available for download from the clarin:el repository.
KPWr (Polish Corpus of Wrocław University of Technology) 1.2 **Size:** 447,000 tokens **Annotation (other type):** selected predicate-argument relations, relations between named entities, anaphora relations, word senses, events, temporal expressions, spatial relations between entities, keywords and semantic roles within nominal and adjective phrases **Licence:** CC BY-SA 3.0	Polish	This corpus contains texts in a variety of domains (blogs, science, stenographic recordings, etc.). The corpus is available for download from the CLARIN-PL repository.
Polish Summaries Corpus **Size:** 10845 summaries **Annotation (other type):** summarization **Licence:** CC BY 3	Polish	This corpus is available for download from the ZIL IPI PAN repository.
WUT Relations Between Sentences Corpus **Size:** 5654 sentences **Annotation (other type):** relations between sentences - Cross-document Structure Theory (CST) **Licence:** CC BY-SA 3.0	Polish	This corpus contains news items. The corpus is available for download from the CLARIN.PL repository.
Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0 **Size:** 92,000 tokens **Annotation (other type):** word normalisation **Licence:** CC BY 4.0	Serbian	This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
CMC training corpus Janes-Tag 2.0 **Size:** 75,000 tokens **Annotation (other type):** word normalisation **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains computer-mediated communication (CMC). The corpus is morphosyntactically tagged following the MULTEXT-East Version 5 tagset. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

Corpus of comma placement Vejica 1.3 **Size:** 104,000 sentences **Annotation (other type):** comma placement **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains texts from various Slovenian corpora (KUST, Šolar, Lektor, JANES-Vejica) and from the Slovenian Wikpedia. The corpus is available for dow.nload from CLARIN.SI.
Terminology identification dataset KAS-term 1.0 **Size:** 22,950 term candidates **Annotation (other type):** monolingual term extraction **Licence:** CC BY-SA 4.0	Slovenian	This corpus contains term candidates from PhD theses in chemistry, computer science and political science. The corpus is available for download from the CLARIN.SI repository.
Training corpus ssj500k 2.1 **Size:** 586,000 tokens **Annotation (other type):** verbal multiword expression tagging, semantic role labelling **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains standard Slovenian. The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.
Bilingual terminology extraction dataset KAS-biterm 1.0 **Size:** 1,950 sentences, 78,500 tokens, 3,700 terms **Annotation (other type):** bi-lingual term extraction **Licence:** CC BY-SA 4.0	Slovenian, English	This corpus contains PHD theses. The corpus is available for download from the CLARIN.SI repository.
Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1) **Size:** 5.8 million tokens **Annotation (other type):** identification of verbal multi-word expressions (idioms, light-verb constructions, verb-particle constructions, inherently reflexive verbs, multi-verb constructions) **Licence:** PARSEME Shared Task Data (v. 1.1) Agreement	16 languages	This corpus collection is available for download from LINDAT. The PARSEME corpora can be queried individually through *KonText.* We provide the individual links to each corpus: 1. Parseme VMWE 1.0 – Czech 2. Parseme VMWE 1.0 – German 3. Parseme VMWE 1.0 – Greek 4. Parseme VMWE 1.0 – Spanish 5. Parseme VMWE 1.0 – Persian 6. Parseme VMWE 1.0 – French 7. Parseme VMWE 1.0 – Hungarian 8. Parseme VMWE 1.0 – Italian 9. Parseme VMWE 1.0 – Maltese 10. Parseme VMWE 1.0 – Polish 11. Parseme VMWE 1.0 – Portuguese 12. Parseme VMWE 1.0 – Romanian 13. Parseme VMWE 1.0 – Slovenian 14. Parseme VMWE 1.0 – Swedish 15. Parseme VMWE 1.0 – Turkish