Title	Overview of Newspaper Corpora
Version	2
Author(s)	DF, JL
Date	20-11-2017
Status	For distribution
Distribution	NCF, UI
ID	CE-2017-1128

Contents

1. Background and approach

In the following survey, our aim is to provide an overview of newspaper corpora for languages relevant for all the countries that are members or observers of CLARIN ERIC. Our motivation was to identify to what extent these resources exist and are easily available, and check which information about the resources is available, thereby highlighting the aspects in which accessibility of these corpora as well as the presentation of the relevant information can be optimised from a User Involvement perspective.

The corpora were identified in five steps:

(i) through the VLO,
(ii) on the repositories or websites of the national consortia,
(iii) through META-SHARE,
(iv) through the LRE Map, and
(v) through the national UI coordinators and participants of the Working with Digital Collections of Newspapers workshop.[1]

The overview is split into three parts: section (2) presents the newspaper corpora that are in the CLARIN infrastructure, section (3) provides an overview of the corpora that are not part of the infrastructure but were identified through the LRE Map, through META-SHARE and through the national UI coordinators; section (4) presents an overview of the collections of digitized newspapers which were provided to us by the participants of the workshop.

2. Newspaper corpora in the CLARIN infrastructure

The following table lists the newspaper corpora that were identified in the CLARIN infrastructure in alphabetical order of the language of the corpus. Each corpus is described in terms of its size, the time period covered and information on the annotation, availability and licence. In the left-hand column, a link is given to the most relevant webpage for a specific corpus, prioritizing the CLARIN repository where the corpus is available, then (in a descending order of priority), VLO and dedicated webpages. In total, we were able to identify 30 such corpora.

The hyperlinks were last accessed 20 November 2017.

[1] https://www.clarin.eu/event/2016/clarin-plus-workshop-working-digital-collections-newspapers.

Table 1: Overview of the newspaper corpora in CLARIN infrastructure

Corpus name	Description
An-Nahar Newspaper Text Corpus Arabic Unclear size Unclear annotation Unavailable	This corpus consists of data from the Arabic newspaper An-Nahar for the period between 1995 and 2000. The size of the corpus is unclear. It is unclear how the corpus is annotated. It can be found through the VLO though it is unavailable because of the restrictive ELRA END USER licence.
SYN2006PUB: corpus of Czech newspapers Czech 300 million tokens Tokenised, lemmatised, PoS-tagged For download	This corpus consists of data from 11 Czech newspapers for the period between 1989 and 2004. It consists of 300 million tokens. The corpus is tokenised, lemmatised and PoS-tagged. The corpus can be found through the VLO and is available for download through the Czech repository LINDAT under the CC-BY licence.
SYN2013PUB: corpus of written Czech newspapers Czech 935 million tokens Lemmatised, morphologically tagged For download	This corpus consists of data from Czech newspapers for the period between 2005 and 2009. It is unclear which newspapers are represented. The corpus consists of 935 million tokens. The corpus is lemmatised and morphologically tagged. It can be found through the VLO and is available for download in the Czech repository LINDAT under the Czech National Corpus (Shuffled Corpus Data) licence.
The Karjalainen Corpus Finnish Unclear size Unclear annotation Unavailable	This corpus consists of data from the Finnish newspaper Karjalainen from the 1990s. The size of the corpus is unclear. It is unclear how the corpus is annotated. The corpus can be found through the VLO but is unavailable. The licence is likewise unclear.
The Karelian Finnish Newspaper Corpus Finnish Unclear size Unclear annotation Unavailable	This corpus consists of data from the Finnish newspaper Karjalan Sanomat for the period between 2012 and 2014. It is unclear how the corpus is annotated. The corpus can be found in the FIN-CLARIN repository though it is unavailable as it is still in development. The licence is under negotiation.
BREF-80 French Unclear size Unclear annotation Unavailable	This corpus consists of data from the French newspaper Le Monde from an unknown period. The size of the corpus is unclear. It is unclear how the corpus is annotated. It can be found

	through the VLO though it is unavailable because of the restrictive ELRA END USER licence.
Corpus journalistique issu de l'Est Républicain French Unclear size Unclear annotation For download	This corpus contains data from the French newspaper l'Est Républicain for the period between 1999 and 2003.The size of the corpus is unclear. It is unclear how the corpus is annotated. It can be found through the VLO and is available for download in the ORTOLANG repository under CC-BY.
Tübingen Treebank of Written German / Newspaper Corpus German 1.8 million tokens MSD tagged, lemmatised, syntactic constituency, named-entities Concordancer	This corpus consists of data from the German newspaper Die Tageszeitung from an unknown period. The corpus consists of 1.8 million tokens. The corpus is tagged for MSD, lemmatised, syntactic constituency and dependencies as well as named entities, anaphora and coreference relations. The corpus can be found through the VLO and is available through a dedicated concordancer though not freely – an institutional account is required. The licence is restricted.
TIGER Corpus German German 900,000 tokens PoS-tagged, syntactic structure, lemmatised For download	This corpus consists of data from the German newspaper Frankfurter Rundschau from an unknown period. The corpus consists of 900,000 tokens. The corpus is PoS-tagged, annotated with syntactic structure and lemmatised. It can be found through the VLO and is available for download through a dedicated webpage. The corpus is publicly available, though the licence isn't further specified.
MTP Annotated German corpus - tagged version German 500,000 tokens MSD tagged Unavailable	This corpus consists of data from the two German newspapers Die Frankfurter Allgemeine Zeitung and Die Zeit from 1992. The corpus consists of 500,000 tokens. The corpus is tokenised and MSD tagged. The corpus can be found through the VLO though it is unavailable due to the restrictive ELRA END USER licence.
Mannheim Corpus of Historical Newspapers and Magazines German 4.1 million tokens Unclear annotation For download	This corpus consists of data from 21 German newspapers from the 18th and 19th century. The corpus consists of roughly 4.1 million tokens. It is unclear how the corpus is annotated. It is available for download. The licence is not clear.
The Norwegian Newspaper Corpus Norwegian 700 million tokens multitagged	This corpus consists of data from 24 Norwegian newspapers from 1998 onwards. The corpus consists of 700 million tokens. The corpus is "multitagged" though a clearer description is

Concordancer	of the annotation process is missing. The corpus can be found through the VLO and is available through the concordancer *Corpuscule*. The license is unclear.
ChronoPress Corpus of Polish Press Texts Polish 20 million tokens Tokenised, PoS-tagged, Named entities Concordancer	This corpus consists of data from various Polish newspapers from 1945 and 1962. The corpus consists of 20 million tokens. The corpus is tokenised and tagged for PoS and Named Entities. The corpus can be found through the VLO repository and is available through a dedicated concordancer. The licence is CLARIN PUB.
Romanian corpus of newspaper articles Romanian 50 million tokens Not annotated Unavailable	This corpus consists of data from Romanian newspapers. The corpus consists of 50 million tokens is not annotated. All the other metadata are missing. The corpus can be found through the VLO but is not available (the download link to the project page is broken). The licence is also unclear.
DN 1987 Swedish 5 million tokens Tokenised, PoS-tagged, semantic dependency, compounds For download and concordancer	This corpus consists of data from the Swedish newspaper Dagens Nyheter from 1987. The corpus consists of 5 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations and compounds. It can be found through the Språkbanken repository and is available both for download and through the concordancer Korp under the CC-BY licence.
GP 1994 and 2001-2011 Swedish 271 million tokens Tokenised, PoS-tagged, semantic dependency, compounds For download and concordancer	This corpus consists of data from the Swedish newspaper Göteborgsposten from 1994 and the period between 2001 and 2011. The corpus consists of 271 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations and compounds. It can be found through the Språkbanken repository and is available both for download and through the concordancer Korp under the CC-BY licence.
Kubhist Swedish 1 billion tokens Tokenised, PoS-tagged, semantic dependency Concordancer	This is a historical newspaper corpus for Swedish for the period between the 1740s and 1920s. It consists of 1 billion tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available through the concordancer Korp under the CC-BY licence.
8 sidor Swedish 678,000 tokens	This corpus consists of data from the Swedish newspaper 8 sidor for the period between 2003 and 2012. It consists of 678,000 tokens.

Tokenised, PoS-tagged, semantic dependency, compounds For download and concordancer	The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations and compounds. It can be found through the Språkbanken repository and is available both for download and through the concordancer Korp under the CC-BY licence.
The Webbnyheter corpus Swedish 272 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations Concordancer	This corpus consists of data from various Swedish online newspapers for the period between 2001 and 2013. The corpus consists of 272 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Dagny Swedish 8.1 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Dagny for the period between 1886 and 1913. The corpus consists of 8.1 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Hertha Swedish 3.8 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Hertha for the period between 1914 and 2015. The corpus consists of 3.8 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Idun Swedish 2 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Idun for the period between 1887 and 1917. The corpus consists of 2 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Kvinnorans Tidning Swedish 5.5 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Kvinnornas Tidning for the period between 1921 and 1925. The corpus consists of 5.5 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Rösträtt för Kvinnor	This corpus consists of data from the newspaper Rösträtt

Swedish 2.2 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	för Kvinnor for the period between 1912 and 1919. The corpus consists of 5.5 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Morgonbris Swedish 3.5 million tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Morgonbris for the period between 1904 and 1924. The corpus consists of 3.5 million tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Smittskydd Swedish 691,000 tokens Tokenized, PoS-tagged, tagged for semantic dependency relations For download and concordancer	This corpus consists of data from the newspaper Smittskyd for the period between 2002 and 2010. The corpus consists of 691,000 tokens. The corpus is tokenized, PoS-tagged, tagged for semantic dependency relations. It can be found through the Språkbanken repository and is available under the CC-BY licence.
Multilingual corpora:	
MLCC Multilingual and Parallel Corpora Dutch, English, French, German, Italian and Spanish 100 million tokens Unclear annotation Unavailable	This corpus is multilingual and consists of data from newspapers in Dutch, English, French, German, Italian and Spanish for the period between 1986 and 1994. The corpus contains approx. 100 million tokens. It is unclear how the corpus is annotated. The corpus can be found through the VLO though it is unavailable due to the restrictive ELRA END USER licence.
The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version Finnish and Swedish 8.8 billion tokens Unclear annotation Concordancer	This corpus consists of data from a large variety of Finnish and Swedish newspapers (over 100 for each language) for the period between 1770 and 2011.The corpus consists of 8.8 billion tokens. It is unclear how the corpus is annotated. The corpus can be found in the FIN-CLARIN repository and is available through the concordancer Korp under the CC-BY licence.
The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) Finnish and Swedish Unclear size	This corpus consists of data from a large variety of Finnish and Swedish newspapers (more than 100 for each language) for the period between 1771 and 1874. The size of the corpus is unclear. It is unclear how the corpus is annotated. The corpus can

Unclear annotation Concordancer	be found through the VLO and is available through the concordancer Korp under the CC-BY licence.
Corpora of Newspaper Texts Swedish, English and Finnish 435 million tokens Unclear annotation Unavailable	This corpus consists of data from Swedish, English and Finnish newspapers though the documentation does not explicitly name the newspapers. The period is also unclear. The corpus consists of 435 million tokens. It is unclear how the corpus is annotated. The corpus can be found through FIN-CLARIN though its availability and licence are still under negotiation.

2.1. Identification

In total we were able to identify 30 newspaper corpora that are part of the CLARIN infrastructure. The following 16 were identified via the VLO:

(i) *The Norwegian Newspaper Corpus*
(ii) *SYN2006PUB: corpus of Czech newspapers*
(iii) *SYN2013PUB: corpus of written Czech newspapers*
(iv) *Romanian corpus of newspaper articles*
(v) *Tübingen Treebank of Written German / Newspaper Corpus*
(vi) *TIGER Corpus*
(vii) *The Karjalainen Corpus*
(viii) *Corpus journalistique issu de l'Est Républicain*
(ix) MLCC Multilingual and Parallel Corpora
(x) MTP Annotated German corpus - tagged version
(xi) BREF-80
(xii) An-Nahar Newspaper Text Corpus
(xiii) The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
(xiv) The Karelian Finnish Newspaper Corpus
(xv) ChronoPress Corpus of Polish Press Texts
(xvi) 8 sidor

The following 14 corpora were identified through the national repositories:

(i) DN 1987
(ii) GP 1994 and 2001-2011
(iii) Kubhist
(iv) The Webbnyheter corpus
(v) Dagny
(vi) Hertha
(vii) Idun
(viii) Kvinnorans Tidning

(ix) Morgonbris
(x) Smittskydd
(xi) Rösträtt för Kvinnor
(xii) Corpora of Newspaper Texts
(xiii) The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
(xiv) Mannheim Corpus of Historical Newspapers and Magazines

Corpora (i)-(xi) were found via the Swedish Språkbanken while corpora (xii)-(xiii) were found via the Finish the Language Bank of Finland. Corpus (xiv) was found on the website of the CLARIN centre IDS-Mannheim.

2.2. Availability

2.2.1. For download and through a concordancer priority

The following 10 corpora are available both for download (all through the Swedish Språkbanken) and through a concordancer (*Korp*, Swedish distribution).

(i) DN 1987
(ii) GP 1994 and 2001-2011
(iii) 8 sidor
(iv) Dagny
(v) Hertha
(vi) Idun
(vii) Kvinnorans Tidning
(viii) Morgonbris
(ix) Rösträtt för Kvinnor
(x) Smittskydd

2.2.2. For download

The following 5 corpora are available for download:

(i) SYN2006PUB: corpus of Czech newspapers
(ii) SYN2013PUB: corpus of written Czech newspapers
(iii) TIGER Corpus
(iv) Corpus journalistique issu de l'Est Républicain
(v) Mannheim Corpus of Historical Newspapers and Magazines

Corpora (i) and (ii) can be downloaded through the Czech repository LINDAT. Corpora (iii) and (iv) are available for download through dedicated webpages. Corpus (v) is available via IDS-Mannheim.

2.2.3. Through a concordancer

The following 7 corpora are available through a concordancer:

(i) The Norwegian Newspaper Corpus
(ii) Tübingen Treebank of Written German / Newspaper Corpus
(iii) The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
(iv) The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
(v) The Webbnyheter corpus
(vi) ChronoPress Corpus of Polish Press Texts
(vii) Kubhist

Corpus (i) is accessible through the concordancer *Corpuscule* (provided by CLARINO), while corpora (iii), (iv), (v), (vi) and (viii) are available through *Korp* (Finnish distribution for (iii), (iv), (v) and Finnish for (vi) and (viii). The rest (corpora (ii) and (vii)) are searchable via dedicated concordancers. Access to corpus (ii) requires an institutional account for login.

2.2.4. Unavailable

Despite being listed in the CLARIN infrastructure, the following 8 corpora are unavailable:

(i) Romanian corpus of newspaper articles
(ii) The Karjalainen Corpus
(iii) Corpora of Newspaper Texts
(iv) An-Nahar Newspaper Text Corpus
(v) BREF-80
(vi) MLCC Multilingual and Parallel Corpora
(vii) MTP Annotated German corpus - tagged version
(viii) The Karelian Finnish Newspaper Corpus

The link to the landing pages of corpus (i) is broken. The availability and licence of corpus (v) are still under negotiation. Corpora (iv), (vi), (vii) and (viii) are unavailable because of the restrictive ELRA END USER licence. Finally, it is unclear why corpora (ii) and (iii) are unavailable, though it must be pointed out that they are all listed as part of the *LRT + Open Submissions Data & Tools* collection.

2.3. Metadata

2.3.1. Languages

The corpora in table (1) are mostly monolingual (26 out of 30 corpora). These monolingual corpora represent the following languages: Swedish (12 corpora), German (4 corpora), Czech (2 corpora), Finnish (2 corpora), French (2 corpora), Romanian (1 corpus), Arabic (1 corpus), Polish (1 corpus), Norwegian (1 corpus). The following corpora are multilingual:

(i) Corpora of Newspaper Texts (English, Swedish, Finnish)
(ii) MLCC Multilingual and Parallel Corpora (Dutch, English, French, German, Italian and Spanish)

(iii) The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version (Finnish and Swedish)
(iv) The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) (Finnish and Swedish)

2.3.2. Size and period

The largest corpus is The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version, which consists of approximately 8.8 billion tokens, while the smallest corpus is MTP Annotated German corpus - tagged version, which contains 500,000 tokens. Information on the size of the corpus is missing for the following 6 out of the total 30 corpora:

(i) The Karjalainen Corpus
(ii) An-Nahar Newspaper Text Corpus
(iii) The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
(iv) The Karelian Finnish Newspaper Corpus
(v) Trove Newspaper Corpus
(vi) Corpus journalistique issu de l'Est Républicain

In relation to the time span of the data, the corpus The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version covers the longest period (1770—2011), while MTP Annotated German corpus - tagged version , for instance, covers only one year (1992). The following corpora are historical:

(i) The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
(ii) ChronoPress Corpus of Polish Press Texts (1945—1962)
(iii) Dagny (1886—1913)
(iv) Hertha (1914—2015)
(v) Idun (1887—1917)
(vi) Kvinnorans Tidning (1921—1925)
(vii) Morgonbris (1904—1924)
(viii) Mannheim Corpus of Historical Newspapers and Magazines

Information on the time span is missing for the following 6 out of 30 corpora:

(i) Romanian corpus of newspaper articles
(ii) Tübingen Treebank of Written German / Newspaper Corpus
(iii) TIGER Corpus
(iv) Corpora of Newspaper Texts
(v) An-Nahar Newspaper Text Corpus
(vi) BREF-80

2.3.3. Annotation and licence

Most of the corpora in table (1) are tokenised, PoS-tagged and lemmatised. Additionally, all the 12 Swedish corpora (cf. section 2.1 for the list) are also annotated for semantic dependencies, while the Tübingen Treebank of Written German / Newspaper Corpus and ChronoPress Corpus of Polish Press Texts are also annotated for Named Entities.

It is unclear how the following 9 out of the total 30 corpora are annotated:

(i) The Karjalainen Corpus
(ii) Corpora of Newspaper Texts
(iii) An-Nahar Newspaper Text Corpus
(iv) BREF-80
(v) MLCC Multilingual and Parallel Corpora
(vi) The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki VersionThe Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
(vii) The Karelian Finnish Newspaper Corpus
(viii) Corpus journalistique issu de l'Est Républicain
(ix) Mannheim Corpus of Historical Newspapers and Magazines

17 corpora are available under CC-BY, 4 under the ELRA END USER licence, 1 is available under the Czech National Corpus (Shuffled Corpus Data) licence, 1 is available under CLARIN PUB and 1 has restricted access. Information on the licence is missing for the following 6 corpora:

(i) The Norwegian Newspaper Corpus
(ii) Romanian corpus of newspaper articles
(iii) The Karjalainen Corpus
(iv) Corpora of Newspaper Texts
(v) The Karelian Finnish Newspaper Corpus
(vi) Mannheim Corpus of Historical Newspapers and Magazines

3 Newspaper corpora that are not part of the CLARIN infrastructure

The following table lists the newspaper corpora that were identified through the LRE Map, on META-SHARE, through personal correspondence, but are not in the CLARIN infrastructure. They are listed in an alphabetical order according to the language of the corpus. Each corpus is described in terms of its size, the time period covered and information on the annotation, availability and licence. In total, we were able to identify 10 such corpora. While the list most likely is not comprehensive, we believe we have identified the most visible examples of existing newspaper corpora that are still missing from the CLARIN infrastructure. We will be updating the list with new entries on a need basis.

Table 2: Newspaper corpora that are not part of the CLARIN infrastructure

Corpus name	Description
Zurich English Newspaper Corpus English 1.6 million tokens Unclear annotation For download	This corpus consists of data from various English newspapers (mainly newspapers from London) from the 17th and 18th century. The corpus consists of 1.6 million tokens. It is unclear how the corpus is annotated. The corpus can be found through Google and is seemingly available for download, though the author needs to be contacted beforehand. The licence is unclear.
Trove Newspaper Corpus English Unclear size Named entities Concordancer	This corpus consists of data from English newspapers, though it is unclear which. The period and the size of the corpus are unclear. The corpus is annotated for named entities. It can be found through the LRE Map and is available through a dedicated concordancer. The licence is not clear.
deu_newscrawl_2011 German 426 million tokens Unclear annotation Concordancer	This corpus consists of data from various German newspapers from 2011. The specific newspapers are not named. The corpus consists of 426 million tokens. It is unclear how the corpus is annotated. It can be found through Google and is available through a dedicated concordancer. The licence is not clear.
CRIPCO Italian 43,000 documents Coreference resolution For download	This corpus consists of data from the Italian newspaper L'Adige for the period between 1999 and 2006. The corpus consists of 43,000 documents The corpus is annotated for coreference resolution of named entities. The corpus can be found through META-SHARE and is available for download under a proprietary licence.
WItaC - NewsReader Wikinews Italian Corpus Italian Unclear size Multiple annotation For download	This corpus contains Italian translations of 120 English Wikinews articles from an unknown period. The size of the corpus is unclear. The corpus is annotated at multiple levels, including entities, events, event factuality, temporal information, semantic roles, and intra-document and cross-document event and entity coreference. The link to the corpus was obtained through personal correspondence. It is available for download under CC-BY.
"LA REPUBBLICA" CORPUS Italian 380 million tokens	The corpus consists of data from the Italian newspaper La Repubblica from an unknown period. The corpus contains 380 million tokens.

Tokenised, PoS-tagged, lemmatised Concordancer	The corpus is tokenized, PoS tagged and lemmatised. The link to the corpus was obtained through personal correspondence. It is available through a concordancer under the CC-BY licence.
Corpus of Contemporary Serbian Newspapers and Magazines Serbian 916 million tokens Tokenised, PoS-tagged and lemmatised Unavailable	This corpus consists of data from over a 100 Serbian newspapers for the period between 2004 and 2012. The corpus consists of 916 million tokens. The corpus is tokenised, PoS-tagged and lemmatised. The corpus can be found through META-SHARE but is seemingly unavailable due to the fact that the link to the landing page is broken. The corpus is available under the CC-BY licence.
Multilingual corpora:	
Europeana Newspapers NER Corpora Dutch, French and German Unclear size Named entities For download	This corpus consists of Dutch, French and German data from the Europeana newspapers from an unknown period. The size of the corpus is unclear. The corpus is annotated for named entities. It can be found through the LRE Map and is available for download. The licence is CC0.
Timestamped JSI web corpus Multilingual 35 billion tokens Tokenised, PoS-tagged Concordancer	This corpus contains data from newsfeed for 18 languages for the period between 2014 and 2017. The corpus contains approx. 35 billion tokens. The corpus is tokenised and PoS-tagged. The link to the corpus was obtained through personal correspondence. It is available through noSketchEingine. The licence is not clear.

3.1. Identification

The following 2 corpora were identified through META-SHARE.

(i) Corpus of Contemporary Serbian Newspapers and Magazines
(ii) CRIPCO

The following 2 corpora were identified through the LRE Map.

(i) Trove Newspaper Corpus
(ii) Europeana Newspapers NER Corpora

The following 2 corpora were identified through personal correspondence.

(i) WItaC - NewsReader Wikinews Italian Corpus
(ii) "LA REPUBBLICA" CORPUS

The following 3 corpora were identified through Google.

(i) Timestamped JSI web corpus
(ii) Zurich English Newspaper Corpus
(iii) deu_newscrawl_2011

3.2. Availability

3.2.1. For download and through a concordancer

None of these corpora are available both for download and through a concordancer

3.2.2. For download

The following 4 corpora are available for download:

(i) CRIPCO
(ii) Zurich English Newspaper Corpus
(iii) Europeana Newspapers NER Corpora
(iv) WItaC - NewsReader Wikinews Italian Corpus

Though corpus (ii) is available for download, its authors need to be contacted beforehand.

3.2.3. Through a concordancer

The following 4 corpora are available through a concordancer:

(i) Zurich English Newspaper Corpus
(ii) Trove Newspaper Corpus
(iii) "LA REPUBBLICA" CORPUS
(iv) Timestamped JSI web corpus

3.3. Metadata

2.3.4. Size and period

The largest corpus is Timestamped JSI web corpus, which consists of approximately 35 billion tokens. Information on the size of the corpus is missing for the following 3 out of the total 9 corpora:

(i) Trove Newspaper Corpus
(ii) Europeana Newspapers NER Corpora
(iii) WItaC - NewsReader Wikinews Italian Corpus

Information on the time span is missing for the following 4 out of the 9 corpora:

(i) Trove Newspaper Corpus
(ii) Europeana Newspapers NER Corpora

(iii) WItaC - NewsReader Wikinews Italian Corpus
(iv) "LA REPUBBLICA" CORPUS

2.3.5. Annotation and licence

It is unclear how the following 2 out of the total 9 corpora are annotated:

(i) Zurich English Newspaper Corpus
(ii) deu_newscrawl_2011

Furthermore, information on licence is missing for the following 4 corpora:

(i) Zurich English Newspaper Corpus
(ii) deu_newscrawl_2011
(iii) Trove Newspaper Corpus
(iv) Timestamped JSI web corpus

4 Collections of digitized newspapers

In what follows we provide a list of archives that are not corpora (i.e. structured linguistically annotated sets of data) but larger collections of digitized newspapers. Links to all the collections below were obtained through personal correspondence with the attendees of the Working with Digital Collections of Newspapers and none can be found through the VLO except for the Nederlab.

- Delpher open newspaper archive
  - A collection of 351.000 various Dutch newspapers for the period between 1618-1876. The archive is available through an online search environment and a subset can be downloaded here.
- Croatian Historic Newspapers
  - A collection of 100 historic newspapers in German, Croatian, Hungarian, Italian, Latin, Slovenian and Serbian for the period between 1789 and 1920. The archive is available through an online search environment.
- Archivio storico
  - A collection of the Italian newspaper La Stampa for the period between 1867 and 2005. The archive is available through an online search environment.
- Archivio storico Corriere della Sera
  - A collection of the Italian newspaper Corriere della Sera for the period between 1867 and 2016. The archive is available through an online search environment; however, a paid subscription is required to browse the newspapers.
- Archivio La Repubblica
  - A collection of the Italian newspaper La Repubblica for the period between 1924 and 2008. The archive is available through an online search environment.
- I giornali del Piemonte

    - A collection of the Italian newspaper Giornali de Piemonte for the period between 1846 and 2016. The archive is available through an online search environment.
- Europeana historic newspapers
    - A collection of 18 million newspaper pages from the cultural heritage webpages Europeana and the European Library.
- 19th Century British Library Newspapers Database
    - A collection of 70 UK and Irish national and local newspaper titles from the 19th century. The archive is available through an online search environment; however, it requires institutional access.
- Dagblad Vooruit
    - A collection of the historical Dutch (Flemish) newspaper Vooruit for the period between 1884 and 1918. The archive is available for download.
- The Belgian WAR Press
    - A collection of historical Dutch (Flemish) newspapers for the periods between 1914 and 1918 and 1940 and 1945. The collection is available through an online search environment.
- AMSAB collection
    - A collection of 12 newspapers in Dutch (Flemish) and French from 1800 onward. The collection is available through an online search environment.
- BelgicaPress
    - A collection of 84 newspapers in Dutch (Flemish) and French for the period between 1831 and 1970. The collection is available through an online search environment.
- Nederlab
    - A collection of Dutch newspapers for the period before 1900. It is available through an online search environment.
- JPRESS
    - A collection of 143 newspapers in 10 languages, most of them being in Yiddish and Hebrew. It is available through an online search environment.
- Compact Memory
    - A collection of 209 newspapers, mostly in German (199) but also in Hebrew (9), Yiddish (4), and English (3). Hebrew. It is available through an online search environment.
- AustriaN Newspapers Online
    - A collection of Austrian newspapers for the period between 1568 and 1946. It is available through an online search environment.