Title	Overview of spoken corpora
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	04-10-2018
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2018-1307

Contents

1. Introduction

In the following report, we present an overview of spoken corpora, primarily focusing on those that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). Corpora of spoken language contain transcriptions of spontaneous or planned speech, such as broadcast news or elicited narratives and dialogues. They are often aligned with the accompanying recordings. They are an invaluable resource for various kinds of linguistic research, such as phonology, conversational analysis, and dialectology. Such corpora are carefully sampled and rich in sociodemographic metadata.

The report was conducted in two steps:

(i) manually searching the VLO and the national consortia with keywords like “spoken corpus”, “speech corpus”, etc.
(ii) input provided by CLARIN UI and NC coordinators

The full results are available in a Google Docs Spreadsheet.[1] In total, around 84 spoken corpora were identified. Information on most of the corpora was provided by UI and NC coordinators, whom we would like to thank for their invaluable input. In Section 2, we provide a comprehensive list of the spoken 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 provide a list of spoken corpora are available outside the CLARIN infrastructure.

2. Spoken corpora within the CLARIN infrastructure

Section 2.1 lists those CLARIN corpora that contain both audio recordings and their associated transcriptions. Section 2.2 lists the corpora that contain only transcriptions.

2.1. Corpora that contain transcriptions and recordings

The following table lists a total of 74 spoken corpora that contain both audio recordings and the associated transcriptions.

Corpus	Language	Description
Arabic Speech Corpus **Licence:** CC NC-SA 3.0	Arabic	The corpus is available for download from the Oxford Text Archive.
DIALEKT v1: dialectal corpus with multi-tier transcription **Size:** 100,000 words **Annotation:** orthographically and phonetically (dialect features) transcribed, MSD-tagged, lemmatised **Licence:** Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains traditional dialectological material, mostly unprepared monologue-type speech The corpus is available download (upon request) and through the concordancer KonText.
ORAL2013: balanced corpus of informal spoken Czech (transcriptions & audio) **Size:** 2.8 million words **Annotation:** recordings and transcripts anonymised **Licence:** Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains informal conversations. The corpus is available for download from LINDAT and through the concordancer KonText.

[1] https://docs.google.com/spreadsheets/d/1UusqZ2IqK_dkl9gRSZsgzxlzBpctLoFLS3_1W2iccn8/edit?ouid=102021140038643388505&usp=sheets_home&ths=true. Note that the spreadsheet also lists a number of corpora that contain only audio recordings, or speech corpora tailored to the development of speech technologies. We will include such corpora in an additional survey.

ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions & audio) **Size:** 1 million words **Annotation:** orthographically and phonetically transcribed; MSD-tagged, lemmatised **Licence:** Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains informal conversations. The corpus is available for download from LINDAT and through the concordancer KonText.
OVM – Otázky Václava Moravce **Size:** 35 hours **Annotation:** word-by-word transcriptions, including the transcription of some non-speech events **Licence:** CC BY-NC 3.0	Czech	This corpus contains transcribed recordings from the Czech political discussion broadcast "Otázky Václava Moravce". The corpus is available for download from LINDAT and through the concordancer KonText.
Prague DaTabase of Spoken Czech 1.0 **Size:** 770,000 tokens, 7324 minutes **Licence:** CC BY-NC SA 4.0	Czech	This corpus contains spontaneous dialogue. The corpus is available for download from LINDAT.
Spoken corpus of Karel Makoň **Size:** 1000 hours **Licence:** CC BY-SA 3.0	Czech	The corpus contains talks on Christian mysticism given by Karel Makoň. The corpus is available for download from LINDAT.
Czech Malach Cross-lingual Speech Retrieval Test Collection **Size:** 592 hours **Annotation:** manual annotations of selected topics and interviews' metadata **Licence:** CC BY-NC-ND 4.0	Czech, English, French, German, Spanish	This corpus contains interviews with survivors of the Holocaust. The corpus is available for download from LINDAT.
IFA speech corpus	Dutch	The corpus is available for download; cf. broken link to landing page on LINDAT, however.
IFA Spoken Language Corpus **Size:** 50,000 words (41 minutes/speaker) **Annotation:** Hand-segmented speech	Dutch	The corpus is available for download from an informal webpage.
JASMIN Speech Corpus	Dutch	The corpus contains recordings of human-machine interaction and read speech

Size: 115 hours **Annotation:** PoS-tagged, lemmatised, phonetically transcribed **Licence:** CLARIN RES		performed by children, non-native speakers and senior people. The corpus is available download from LINDAT; cf. broken link to landing page, however.
Air Traffic Control Communication **Size:** 20 hours **Annotation:** speaker information **Licence:** CC BY-NC-ND 3.0	English	This corpus contains recordings of communication between air traffic controllers and pilots The corpus is available for download from LINDAT and through the concordancer KonText.
Boston University Radio Speech Corpus **Size:** 7 hours **Annotation:** PoS-tagged, phonetic alignment, prosodic markers **Licence:** CLARIN RES	English	This corpus contains recordings and texts from radio news. The corpus is available for download from the UPenn repository.
Buckeye Corpus of Conversational Speech **Annotation:** phonetic labels **Licence:** CLARIN RES	English	This corpus contains an interview. The corpus is available for download from ORTOLANG.
ELFA Corpus **Size:** 13 hours **Licence:** restricted	English	This corpus contains recorded lectures and seminars. The corpus is available for download from FIN-CLARIN.
Corpus of Lecture Speech **Size:** 41 hours **Annotation:** orthographically transcribed	Estonian	This corpus contains recordings of academic lectures and oral conference presentations.
Corpus of Radio Interviews **Size:** 36 hours **Annotation:** speech annotation to orthographically transcribed	Estonian	This corpus contains telephone interviews from different radio programmes.
Corpus of Radio News **Size:** 19 hours **Annotation:** speech annotation to orthographically transcribed	Estonian	This corpus contains public broadcast news.
Estonian Dialect Corpus **Size:** 1.3 million words **Licence:** CLARIN ACA **Annotation:** phonetically transcribed, MSD-tagged, partly syntactically parsed	Estonian	This corpus contains interviews. The corpus is available for download from META-SHARE.

Estonian Emotional Speech Corpus **Size:** 1234 sentences **Licence:** CC-BY	Estonian	This corpus contains read sentences that express anger, joy and sadness, or are neutral. The corpus is available for download from META-SHARE.
Estonian North Wind and the Sun Corpus v.1.0.3 **Annotation:** words in standard orthography and phonemes in SAMPA	Estonian	This corpus contains recordings of the tale "Põhjatuul ja päike" (North Wind and the Sun) read by the same speakers who participated in the Phonetic Corpus of Estonian Spontaneous Speech. The corpus is available for download from META-SHARE.
Phonetic Corpus of Estonian Spontaneous Speech v.1.0.4 **Size:** 635,000 words, 90 hours **Annotation:** orthographically and phonetically transcribed, syllables, prosodic feet, intonation phrases, changes in voice quality **Licence:** CLARIN_RES	Estonian	This corpus contains spontaneous speech by speakers with different dialectological and social backgrounds. The corpus is available for download from META-SHARE.
Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg) **Licence:** HZSK-RES (restricted, non-commercial only)	Faroese, Danish	This corpus contains informal interviews.
Aalto University DSP Course Conversation Corpus 2013-2016, Downloadable Version **Size:** 5200 utterances **Licence:** academic	Finnish	This corpus contains spontaneous conversations. The corpus is available for download from FIN-CLARIN.
Finnish Broadcast Corpus **Size:** 18 hours **Licence:** restricted	Finnish	This corpus contains radio and TV broadcasts. The corpus is available for download from FIN-CLARIN and for online querying through the LAT-platform.
Follow-up Study of Dialects of Finnish **Licence:** restricted	Finnish	This corpus contains interviews. This corpus is available for online querying through the LAT-platform.
Route to A wing	Finnish	This corpus contains spontaneous conversations.

Size: 218 tokens **Annotation:** PoS-tagged **Licence:** CC-0		This corpus is available for online querying through the LAT-platform.
Samples of Spoken Finnish **Licence:** CC-BY	Finnish	This corpus contains interviews. This corpus is available for online querying through the LAT-platform and through the concordancer Korp.
The Finnish Dialect Syntax Archive **Size:** 1.2 million words **Annotation:** MSD-tagged **Licence:** CC-BY-NC-ND	Finnish	The corpus contains interviews. The corpus is available through the concordancer Korp.
The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s) **Size:** 210 hours **Licence:** restricted	Finnish	This corpus contains spontaneous speech and interviews. The corpus is available for download from FIN-CLARIN and for online querying through the LAT-platform.
The Corpus of Border Karelia **Size:** 120 hours **Licence:** CC-BY	Finnish, Karelian	This corpus contains interviews. The corpus is available for download from FIN-CLARIN and for online querying through the LAT-platform.
Plenary Sessions of the Parliament of Finland **Size:** 22.5 million words **Licence:** CC-BY-NC-ND	Finnish, Swedish	This corpus contains the proceedings of the Finnish Parliament. The corpus is available through the concordancer Korp.
CLAPI	French	This is a collection containing around 40 corpora which contain social interactions in different contexts: professional, private, institutional, commercial, medical, and educational situations. Most of the corpora can be downloaded and queried through a dedicated concordancer.
Corpus de Français Parlé Parisien des années 2000 **Licence:** CC-BY	French	This corpus contains interviews. The corpus is available for download from a dedicated webpage.
Phonologie du Français Contemporain **Licence:** CC-BY	French	The corpus is available for download from a dedicated webpage.
Australiendeutsch	German	This corpus contains interviews in German

Size: 330,000 words, 65 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed		extraterritorial varieties. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Belgische TV-Debatten **Size:** 10 hours	German	This corpus contains broadcast TV debates The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Berliner Wendekorpus **Size:** 260,000 words, 28 hours **Annotation:** literal and PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains narrative interviews on German reunification. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Biographische und Reiseerzählungen **Size:** 50,000 words, 6 hours **Annotation:** orthographically transcribed	German	This corpus contains narrative and biographic interviews. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Deutsche Hochlautung **Size:** 10,000 words, 2 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains broadcasts in standard German The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Deutsche Mundarten: ehemalige deutsche Ostgebiete **Size:** 838,000 words; 461 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains interviews and elicited speech in German dialects. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Deutsche Standardsprache: König-Korpus **Size:** 50,000 words; 6 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains interviews and elicited speech in standard German The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim). **Note:** Excerpt - complete corpus (around 120 hours) currently in the process of

		curation
Deutsche Umgangssprachen: Pfeffer-Korpus **Size:** 646,000 words, 80 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains interviews in regional varieties of German. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Dialogstrukturen **Size:** 140,000 words, 15 hours **Annotation:** orthographically transcribed, intonation, lemmatised, PoS-tagged, time alignment	German	This corpus contains authentic interaction from various domains. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Elizitierte Konfliktgespräche **Size:** 160,000 words, 12 hours **Annotation:** orthographically transcribed	German	This corpus contains elicited conflict interaction. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Emigrantendeutsch in Israel **Size:** 232,000 words, 285 hours **Annotation:** orthographically transcribed, lemma, PoS-tagged, time alignment	German	This corpus contains interviews in German extraterritorial varieties. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Emigrantendeutsch in Israel: Wiener in Jerusalem **Size:** 225,000 words, 51 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains interviews in German extraterritorial varieties. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Forschungs- und Lehrkorpus gesprochenes Deutsch **Size:** 2.3 million words, 230 hours **Annotation:** literal and PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains authentic interactions from various domains The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Grundstrukturen: Freiburger Korpus **Size:** 600,000 words, 70 hours **Annotation:** orthographically transcribed, intonation, lemmatised, PoS-tagged, time alignment	German	This corpus contains authentic interaction from various domains. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).

Hamburg Modern Times Corpus **Size:** 3 hours **Annotation:** manual annotation of phonetic phenomena, accent/stress marking **Licence:** HZSK-ACA (academic, non-commercial only)	German	This corpus contains task-oriented communcation (e.g., a film retelling) in the context of studying adult L2 acquisition.
Mehrsprachige Kinder im Vorschulalter **Size:** 17,000 words, 13 hours **Annotation:** literal and PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains elicitation tasks with pre-school children. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Russlanddeutsche Dialekte **Size:** 100,000 words, 10 hours **Annotation:** literal and PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German	This corpus contains interviews in German extraterritorial varieties. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Zweite Generation deutschsprachiger Migranten in Israel **Size:** 125 hours	German	This corpus contains interviews in German extraterritorial varieties. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Deutsche Mundarten: Zwirner-Korpus **Size:** 4 million words; 1076 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German, (some Frisian and Dutch)	This corpus contains interviews and elicited speech in German dialects. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Deutsche Mundarten: DDR **Size:** 212,000 words, 385 hours **Annotation:** PoS-tagged, lemmatised, time-aligned, orthographically transcribed	German, (some Sorbian)	This corpus contains interviews and elicited speech in German dialects. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
EXMARaLDA Demo Corpus 1.0 **Size:** 2 hours **Annotation:** suprasegmental information, accentuation/stress marking **Licence:** HZSK-PUB (public, non-commercial only)	German, English, French, Spanish, Turkish, Polish, Vietnamese, Swedish,	This corpus is a demo of the EXMARaLDA system. The corpus is available for download from a CLARIN-D repository.

	Norwegian, Italian, Russian, Afrikaans, Portuguese	
Gesprochene Wissenschaftssprache Kontrastiv **Size:** 760,000 words, 92 hours **Annotation:** literal and PoS-tagged, lemmatised, time-aligned, orthographically transcribed, annotation of discourse phenomena and language mixing	German, English, Polish, Bulgarian	This corpus contains academic interaction. The corpus is available for download and through a concordancer via the Database of Spoken German (AGD @ IDS Mannheim).
Hamburg Adult Bilingual LAnguage (HABLA) **Size:** 79 hours **Licence:** HZSK-RES (restricted, non-commercial only)	German, French, Italian	This corpus contains interviews.
Budapest Sociolinguistic Interview - version 2 **Size:** 270,000 words **Annotation:** MSD-tagged, spoken language phenomena (hesitation, consonant drops) **Licence:** CLARIN RES	Hungarian	This corpus contains sociolinguistic interviews conducted with 50 individuals. The corpus is available for download and through a dedicated concordancer.
Hungarian Speecon Database **Licence:** ELRA	Hungarian	This corpus contains speech tasks involving adults and children. The corpus is available for download from the ELRA catalogue.
CLIPS : corpora e lessici di italiano parlato e scritto **Size:** 100 hours	Italian	This corpus contains speech from 15 different cities in Italy.
Mbochi speech corpus **Size:** 5000 sentences, 4.5 hours **Licence:** ELRA	Mbochi, French	The corpus is available for download from the ELRA catalogue.
Nepali Spoken Corpus **Size:** 31 hours 26 minutes **Annotation:** phonetically transcribed **Licence:** ELRA	Nepali	The corpus is available for download from the ELRA Catalogue.

Nganasan Spoken Language Corpus (NSLC) **Size:** 32 hours **Licence:** HZSK-RES (restricted, non-commercial only)	Nganasan, Russian	
LIA **Size:** 1.5 million tokens **Annotation:** orthographically and phonetically transcribed, MSD-tagged, lemmatised	Norwegian	This corpus contains interviews and conversation in Norwegian dialects. The corpus is available through a CLARINO concordancer (account needed).
NoTa-Oslo **Size:** 1 million tokens **Annotation:** orthographically transcribed, MSD-tagged, lemmatised	Norwegian	This corpus contains interviews and conversations in Oslo sociolects. The corpus is available through a CLARINO concordancer (account needed).
TAUS **Size:** 270 000 tokens **Annotation:** MSD-tagged, lemmatised, orthographically and partially phonetically transcribed	Norwegian	This corpus contains informal interviews in Oslo sociolects. The corpus is available through a CLARINO concordancer (account needed).
The BigBrother Corpus **Size:** 440,300 tokens **Annotation:** orthographically transcribed, msd-tagged, lemmatised	Norwegian	This corpus contains recordings and transcripts from the Norwegian Big Brother in 2001. The corpus is available through a CLARINO concordancer.
Corpus of American Nordic Speech (CANS) **Size:** 251,000 tokens **Annotation:** orthographically and phonetically transcribed, MSD-tagged, lemmatised	Norwegian, Swedish	This corpus contains interviews, conversations. Norwegian and Swedish dialects in America. The corpus is available through a CLARINO concordancer.
Hamburg Corpus of Polish in Germany (HamCoPoliG) **Size:** 38 hours **Licence:** HZSK-RES (restricted, non-commercial only)	Polish	This corpus contains spontaneous speech and reading tasks
Consecutive and Simultaneous Interpreting (CoSi) **Size:** 6 hours **Licence:** HZSK-RES (restricted, non-commercial only)	Portuguese, English	This corpus contains lectures in Portuguese with simultaneous interpretation in English
Skolt Saami Documentation Corpus (2016)	Skolt Saami	This corpus contains interviews.

Size: 19 hours **Annotation:** MSD-tagged		This corpus is available for online querying through the LAT-platform.
Hamburg Corpus of Argentinean Spanish (HaCASpa) **Size:** 19 hours **Annotation:** orthographically transcribed **Licence:** HZSK-RES (restricted, non-commercial only)	Spanish (Argentinian)	This corpus contains spontaneous speech and reading tasks.
Catalan in a bilingual context (PhonCAT) **Size:** 144 hours **Annotation:** orthographically and phonetically transcribed **Licence:** HZSK-RES (restricted, non-commercial only)	Spanish (Catalan)	This corpus contains read, elicited and spontaneous speech.

2.2. Corpora with without audio recordings

The following table contains 11 corpora that contain only transcriptions of recorded speech.

Corpus	Language	Description
ORAL2008: Balanced corpus of informal spoken Czech **Size:** 1 million tokens **Licence:** CC BY-NC-SA 3.0	Czech	This corpus contains informal conversations. The corpus is available for download from LINDAT and through the concordancer KonText.
ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions) **Size:** 1 million tokens **Annotation:** orthographically and phonetically transcribed, MSD-tagged, lemmatised **Licence:** CC BY-NC-SA 4.0	Czech	This corpus contains informal conversations. The corpus is available for download from LINDAT and through the concordancer KonText.
Prague Dependency Treebank of Spoken Language (PDTSL) 0.5 **Size:** 120,000 words **Annotation:** syntactic dependencies **Licence:** ACADEMIC (PDTSL)	Czech	The corpus is available for download from LINDAT.
ParCorFull: A Parallel Corpus Annotated with Full Coreference **Size:** 160,000 tokens **Annotation:** coreference (nominal and clausal)	English, German	This corpus contains planned speech and newswire. The corpus is available for download from LINDAT.

Licence: CC BY-NC-ND 4.0		
The Spoken Wikipedia Corpora **Annotation:** text segmentation, normalization, time-alignment **Licence:** CC-BY SA 4.0	English, German, Dutch	The corpus contains transcripts of read Wikipedia articles The corpus is available for download from a CLARIN-D repository.
Corpus of Spoken Estonian **Size:** 1 million words **Annotation:** unspecified tagging	Estonian	The corpus contains transcripts of recordings from various domains.
ALCEBLA **Size:** 72 hours **Licence:** HZSK-RES (restricted, non-commercial only)	German, Spanish	This corpus contains Speech tasks performed by bilingual children.
Corpus of Doctor-Patient Conversations from Ahus **Size:** 958,830 tokens **Annotation:** orthographically transcribed, MSD-tagged, lemmatised	Norwegian	This corpus contains doctor-patient conversations. The corpus is available through a CLARINO concordancer (account needed).
Spoken corpus Gos 1.0 **Size:** 1 million words, 120 hours **Annotation:** orthographically and phonetically transcribed **Licence:** CC BY-NC-SA 4.0	Slovenian	This corpus contains transcripts from radio and TV shows, school lessons, private conversations, business meetings The corpus is available for download from CLARIN.SI as well as through a dedicated webconcordancer
Spoken corpus Gos VideoLectures 2.0 (transcription) **Size:** 79420 words **Annotation:** PoS-tagged, lemmatised, orthographically and phonetically transcribed **Licence:** CC BY 4.0	Slovenian	This corpus contains public academic speech. The corpus is available for download from CLARIN.SI and through the concordancer KonText.
Gothenburg Dialogue Corpus **Size:** 1,470,000 tokens **Annotation:** MSD-tagged, lemmatised	Swedish	The corpus is available through the concordancer Korp (account needed).

3. Overview of the CLARIN corpora

There are 85 spoken corpora in total in the CLARIN infrastructure. Section 3.1 lists those corpora that are not yet included in the VLO, but can be found through a CLARIN repository. Section 3.2. provides an overview of the availability of the corpora and Section 3.3 an overview of the relevant metadata.

3.1. Identification

The vast majority of the CLARIN corpora can be found in the VLO (82 out of 85). The following 3 corpora cannot be found through the VLO.

[1] Gothenburg Dialogue Corpus (available via the Swedish Language Bank)
[2] Interaction and Variation in Pluricentric Languages (IVIP) (available via the Swedish Language Bank)
[3] LIA (available via a CLARINO node)

3.2. Availability

The following 16 corpora are available for download and through a concordancer:

[1] ORAL2008: Balanced corpus of informal spoken Czech
[2] ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions)
[3] Spoken corpus Gos 1.0
[4] Spoken corpus Gos VideoLectures 2.0 (transcription)
[5] ORAL2013: balanced corpus of informal spoken Czech (transcriptions & audio)
[6] ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions & audio)
[7] DIALEKT v1: dialectal corpus with multi-tier transcription
[8] COLA - Corpus Oral de Lenguaje Adolescente
[9] IVIP demo - Interaction and Variation in Pluricentric Languages
[10] CLAPI
[11] Air Traffic Control Communication
[12] OVM – Otázky Václava Moravce
[13] Budapest Sociolinguistic Interview - version 2
[14] The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)
[15] Finnish Broadcast Corpus
[16] The Corpus of Border Karelia

The following 35 corpora are available through a concordancer:

[1] Gothenburg Dialogue Corpus
[2] Corpus of Doctor-Patient Conversations from Ahus
[3] Interaction and Variation in Pluricentric Languages (IVIP)
[4] The BigBrother Corpus
[5] Corpus of American Nordic Speech (CANS)
[6] LIA
[7] Nordic Dialect Corpus
[8] NoTa-Oslo
[9] TAUS
[10] Route to A wing
[11] Follow-up Study of Dialects of Finnish
[12] Samples of Spoken Finnish
[13] Skolt Saami Documentation Corpus (2016)
[14] The Finnish Dialect Syntax Archive
[15] Plenary Sessions of the Parliament of Finland
[16] Deutsche Mundarten: Zwirner-Korpus
[17] Deutsche Mundarten: ehemalige deutsche Ostgebiete

[18]Deutsche Mundarten: DDR
[19]Deutsche Umgangssprachen: Pfeffer-Korpus
[20]Deutsche Standardsprache: König-Korpus
[21]Deutsche Hochlautung
[22]Australiendeutsch
[23]Russlanddeutsche Dialekte
[24]Emigrantendeutsch in Israel
[25]Emigrantendeutsch in Israel: Wiener in Jerusalem
[26]Forschungs- und Lehrkorpus gesprochenes Deutsch
[27]Zweite Generation deutschsprachiger Migranten in Israel
[28]Gesprochene Wissenschaftssprache Kontrastiv
[29]Grundstrukturen: Freiburger Korpus
[30]Dialogstrukturen
[31]Berliner Wendekorpus
[32]Biographische und Reiseerzählungen
[33]Belgische TV-Debatten
[34]Elizitierte Konfliktgespräche
[35]Mehrsprachige Kinder im Vorschulalter

The following 24 corpora are available for download:

[1] The Spoken Wikipedia Corpora
[2] ParCorFull: A Parallel Corpus Annotated with Full Coreference
[3] Prague Dependency Treebank of Spoken Language (PDTSL) 0.5
[4] EXMARaLDA Demo Corpus 1.0
[5] Nepali Spoken Corpus
[6] Phonetic Corpus of Estonian Spontaneous Speech v.1.0.4
[7] Mbochi speech corpus
[8] Arabic Speech Corpus
[9] Spoken corpus of Karel Makoň
[10]IFA Spoken Language Corpus
[11]IFA speech corpus
[12]JASMIN Speech Corpus
[13]Boston University Radio Speech Corpus
[14]Buckeye Corpus of Conversational Speech
[15]Spoken Portuguese Corpus
[16]Corpus de Français Parlé Parisien des années 2000
[17]Phonologie du Français Contemporain
[18]Prague DaTabase of Spoken Czech 1.0
[19]Estonian North Wind and the Sun Corpus v.1.0.3
[20]Estonian Emotional Speech Corpus
[21]Estonian Dialect Corpus
[22] Hungarian Speecon Database
[23]The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)
[24]Aalto University DSP Course Conversation Corpus 2013-2016, Downloadable Version

The following 16 corpora are unavailable.

[1] Corpus of Spoken Estonian
[2] ALCEBLA
[3] Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg)

[4] Hamburg Adult Bilingual LAnguage (HABLA)
[5] Hamburg Corpus of Polish in Germany (HamCoPoliG)
[6] Consecutive and Simultaneous Interpreting (CoSi)
[7] Hamburg Corpus of Argentinean Spanish (HaCASpa)
[8] Catalan in a bilingual context (PhonCAT)
[9] Corpus of Lecture Speech
[10] Hamburg Modern Times Corpus
[11] Nganasan Spoken Language Corpus (NSLC)
[12] Corpus of Radio News
[13] Corpus of Radio Interviews
[14] Corpus of Lecture Speech
[15] Livonian prosody corpus
[16] CLIPS : corpora e lessici di italiano parlato e scritto

3.3. Metadata

3.3.1. Languages

A vast majority (69 out of 85) of the corpora are monolingual, accounting for the following 15 languages:

[1] Arabic (1 corpus)
[2] Czech (9 corpora)
[3] Dutch (3 corpora)
[4] English (4 corpora)
[5] Estonian (8 corpora)
[6] Finnish (7 corpora)
[7] French (3 corpora)
[8] German (18 corpora)
[9] Hungarian (2 corpora)
[10] Italian (1 corpus)
[11] Nepali (1 corpus)
[12] Norwegian (5 corpora)
[13] Polish (1 corpus)
[14] Skolti Saami (1 corpus)
[15] Slovenian (2 corpora)
[16] Spanish (2 corpora)
[17] Swedish (1 corpus)

16 corpora are multilingual, accounting for the following 16 language combinations:

[1] Czech, English, French, German, Spanish (1 corpus)
[2] English, German (1 corpus)
[3] English, German, Dutch (1 corpus)
[4] Faroese, Danish (1 corpus)
[5] Finnish, Karelian (1 corpus)
[6] Finnish, Swedish (1 corpus)
[7] German, Frisian, Dutch (1 corpus)
[8] German, Sorbian (1 corpus)
[9] German, English, French, Spanish, Turkish, Polish, Vietnamese, Swedish, Norwegian, Italian, Russian, Afrikaans, Portuguese (1 corpus)

[10]German, English, Polish, Bulgarian (1 corpus)
[11]German, French, Italian (1 corpus)
[12]German, Spanish (1 corpus)
[13]Nganasan, Russian (1 corpus)
[14]Norwegian, Swedish (1 corpus)
[15]Mbochi, French (1 corpus)
[16]Portuguese, English (1 corpus)

3.3.2. Size

Information on size is available for the vast majority of corpora – 73 out of total 85. The following 12 corpora lack information on size:

[1] The Spoken Wikipedia Corpora
[2] Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg)
[3] Arabic Speech Corpus
[4] IFA speech corpus
[5] Buckeye Corpus of Conversational Speech
[6] CLAPI
[7] Corpus de Français Parlé Parisien des années 2000
[8] Corpus de Français Parlé à Bruxelles
[9] Phonologie du Français Contemporain
[10] Hungarian Speecon Database
[11] Skolt Saami Documentation Corpus (2016)
[12] Samples of Spoken Finnish

The corpora in Table 1 are inconsistent w.r.t. size documentation. Although they contain both audio recordings and the associated transcriptions, some corpora only report the size of the recordings (i.e., in terms of length) while others only report the size of the transcripts (i.e., in terms of token/sentence number).

The following Table 1 (i.e. corpora that contain both transcriptions and recordings) corpora contain information on the length of the recordings, but lack information on the size of the transcriptions:

[1] Spoken corpus of Karel Makoň
[2] JASMIN Speech Corpus
[3] Hamburg Modern Times Corpus
[4] Nganasan Spoken Language Corpus (NSLC)
[5] Zweite Generation deutschsprachiger Migranten in Israel
[6] Belgische TV-Debatten
[7] Czech Malach Cross-lingual Speech Retrieval Test Collection
[8] OVM – Otázky Václava Moravce
[9] Corpus of Radio News
[10] Corpus of Radio Interviews

The following Table 1 corpora lack information on the size of the recordings, but contain information on the size of the audio recordings:

[1] ORAL2013: balanced corpus of informal spoken Czech (transcriptions & audio)

[2] ORTOFON v1: balanced corpus of informal spoken Czech with multi-tier transcription (transcriptions & audio)
[3] COLA - Corpus Oral de Lenguaje Adolescente
[4] Interaction and Variation in Pluricentric Languages (IVIP)
[5] IVIP demo - Interaction and Variation in Pluricentric Languages
[6] The BigBrother Corpus
[7] Corpus of American Nordic Speech (CANS)
[8] LIA
[9] Nordic Dialect Corpus
[10] NoTa-Oslo
[11] TAUS
[12] Estonian Emotional Speech Corpus

Size overview (for corpora that have information on length available):

- 8 corpora contain <10 hours of recordings;
- 25 corpora contain between 10 and 100 hours of recordings;
- 13 corpora contain >100 hours of recordings

3.3.3. Annotation

Information on annotation is available for more than half of the corpora – 56 out of 85. Information on annotation is missing for the following 29 corpora:

[1] ALCEBLA
[2] Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg)
[3] Hamburg Adult Bilingual LAnguage (HABLA)
[4] Hamburg Corpus of Polish in Germany (HamCoPoliG)
[5] Consecutive and Simultaneous Interpreting (CoSi)
[6] Corpus of Lecture Speech
[7] Mbochi speech corpus
[8] Arabic Speech Corpus
[9] Spoken corpus of Karel Makoň
[10] IFA speech corpus
[11] Nganasan Spoken Language Corpus (NSLC)
[12] COLA - Corpus Oral de Lenguaje Adolescente
[13] Zweite Generation deutschsprachiger Migranten in Israel
[14] Belgische TV-Debatten
[15] CLAPI
[16] Corpus de Français Parlé Parisien des années 2000
[17] Phonologie du Français Contemporain
[18] Prague DaTabase of Spoken Czech 1.0
[19] Estonian Emotional Speech Corpus
[20] Livonian prosody corpus
[21] Hungarian Speecon Database
[22] The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)
[23] Aalto University DSP Course Conversation Corpus 2013-2016, Downloadable Version
[24] ELFA Corpus
[25] Finnish Broadcast Corpus

[26] The Corpus of Border Karelia
[27] Follow-up Study of Dialects of Finnish
[28] Samples of Spoken Finnish
[29] Plenary Sessions of the Parliament of Finland

33 corpora are PoS/MSD-tagged, 28 corpora are lemmatised, 25 corpora contain orthographic transcriptions, and 12 corpora contain phonetic transcriptions.

3.3.4. Licence

Information on license is available for slightly more than half of the corpora – 47 out of 85. Information on license is missing for the following 38 corpora:

[1] Corpus of Spoken Estonian
[2] Gothenburg Dialogue Corpus
[3] Corpus of Doctor-Patient Conversations from Ahus
[4] Corpus of Lecture Speech
[5] IFA Spoken Language Corpus
[6] IFA speech corpus
[7] Interaction and Variation in Pluricentric Languages (IVIP)
[8] The BigBrother Corpus
[9] Corpus of American Nordic Speech (CANS)
[10] LIA
[11] Nordic Dialect Corpus
[12] NoTa-Oslo
[13] TAUS
[14] Deutsche Mundarten: Zwirner-Korpus
[15] Deutsche Mundarten: ehemalige deutsche Ostgebiete
[16] Deutsche Mundarten: DDR
[17] Deutsche Umgangssprachen: Pfeffer-Korpus
[18] Deutsche Standardsprache: König-Korpus
[19] Deutsche Hochlautung
[20] Australiendeutsch
[21] Russlanddeutsche Dialekte
[22] Emigrantendeutsch in Israel
[23] Emigrantendeutsch in Israel: Wiener in Jerusalem
[24] Forschungs- und Lehrkorpus gesprochenes Deutsch
[25] Zweite Generation deutschsprachiger Migranten in Israel
[26] Gesprochene Wissenschaftssprache Kontrastiv
[27] Grundstrukturen: Freiburger Korpus
[28] Dialogstrukturen
[29] Berliner Wendekorpus
[30] Biographische und Reiseerzählungen
[31] Belgische TV-Debatten
[32] Elizitierte Konfliktgespräche
[33] Mehrsprachige Kinder im Vorschulalter
[34] Estonian North Wind and the Sun Corpus v.1.0.3
[35] Corpus of Radio News
[36] Corpus of Radio Interviews

[37] Corpus of Lecture Speech
[38] CLIPS : corpora e lessici di italiano parlato e scritto

17 corpora are available under CC-BY licenses, 10 under HZSK, 5 under CLARIN RES, 3 under ELRA, the rest miscellaneous.

4. Non-CLARIN spoken corpora

The following table lists 18 identified corpora that are not listed in the VLO and are not available through CLARIN repositories.

Corpus	Language	Description
Griffith Corpus of Spoken Australian English **Size:** 32,134 words	English	The corpus is available for download and through the concordancer of the Australian National Corpus.
Spoken BNC2014 **Size:** 10 million words	English	The corpus contains face-to-face conversations between people who speak British English as their first language. The corpus is available through the CQP concordancer.
The Aston Corpus of West Midlands English (ACWME) **Annotation:** orthographically transcribed	English	The corpus contains recordings of performances - comedy, drama, poetry, song and story-telling - and related interviews with performers, members of the audience and local and national celebrities. The corpus is available for download from a dedicated webpage.
Vienna-Oxford International Corpus of English	English	The corpus contains naturally occurring, non-scripted face-to-face interactions in English as a lingua franca (ELF). The corpus is available through a dedicated concordancer.
AN.ANA.S._MT	English, Italian, Spanish	The corpus contains TV-broadcasts and elicited dialogues.
Babel - A Multi Language Database **Annotation**: orthographically transcribed	Hungarian	This corpus contains various elicited speech tasks.
BEA (Hungarian Spontaneous Speech Database) **Size**: 465 recordings **Annotation:** partial transcription **Licence:** restricted	Hungarian	This corpus contains spontaneous speech.

Hungarian Broadcast News Database **Size:** 25,000 words, 3.5 hours **Annotation:** audio-level annotations **Licence:** META_SHARE NC-NoReD	Hungarian	The corpus is available for download (upon request) from META-SHARE.
Hungarian Gigaword Corpus / "spoken language" subcorpus **Size:** 76 million words **Annotation:** PoS-tagged, MSD-tagged	Hungarian	The corpus contains radio broadcasts (reading aloud and spontaneous conversation) The corpus is available through the Hungarian Gigaword Corpus concordancer.
Hungarian Kindergarten Language Corpus **Size:** 192,000 words **Annotation:** PoS-tagged, MSD-tagged **Licence:** restricted	Hungarian	This corpus contains elicited speech tasks (picture descriptions) and guided conversation with children. The corpus is available for download through META-SHARE.
Hungarian Reference Speech Database **Size:** 6 hours **Annotation:** partial phonemic-level annotation **Licence:** META-SHARE No-Redistribution Commercial FF	Hungarian	This corpus contains reading tasks. The corpus is available for download (upon request) from META-SHARE.
Medical Speech Database **Annotation:** phonetic transcription **Licence:** META-SHARE C-NoReD-FF	Hungarian	The corpus is available for download (upon request) from META-SHARE.
Corpus LIP **Size:** 490,000 words	Italian	The corpus is available through a dedicated concordancer.
Corpus AVIP-API **Annotation:** orthographically transcribed	Italian	The corpus contains quasi-spontaneous dialogues (a map task). The corpus is available for download from a dedicated webpage.
Corpus Lips **Size:** 700,000 words, 100 hours **Annotation:** PoS-tagged, lemmatised	Italian	This is a L2-learner corpus. The corpus is available for download from a dedicated webpage.
Selezione dal "Corpus di parlato telegiornalistico. Anni Sessanta vs. 2005 **Annotation:** orthographically	Italian	This corpus contains news broadcast. The corpus is available for download from a dedicated webpage.

transcribed		
SpIt-MDb (Spoken Italian - Multilevel Database) **Annotation:** orthographically transcribed	Italian	This corpus contains spontaneous speech. The corpus is available for download from a dedicated webpage.
Uralic Languages under the Influence (UraLUID) database **Size:** 108,000 tokens, 4 hours **Annotation:** MSD-tagged, time-alignment, phonetic and orthographic transcription	Udmurt, Tundra Nenets, Synya Khanty, Surgut Khanty	This corpus contains narratives (e.g., folk storites). The corpus is available for download from a dedicated website.