Title	Overview of L2 corpora and resources
Version	1.0
Author(s)	Jakob Lenardič, Therese Lindström Tiedemann, Darja Fišer
Date	04-04-2018
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2018-1202

Table of contents

1 Introduction

In the following report, we present an overview of second language (L2) 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). The report was conducted in several steps:

(i) manually searching the VLO and the national consortia with keywords like "learner corpus", "academic writing", and "longitudinal corpus";
(ii) cross-referencing L2 corpora that are listed on the UCL[1] website with the VLO and the national repositories; and
(iii) input provided by CLARIN UI and NC coordinators as well as the participants of the Workshop on interoperability of L2 resources and tools, which took place in December 2017.

The full results are available in a Google Docs Spreadsheet.[2] In total, more than 180 resources were identified but the majority seems to be unavailable through regular channels (i.e., download, concordancers), which is why in this report we focus on those that are available. In Section 2, we provide a comprehensive list of the L2 corpora that are part of the CLARIN infrastructure, describing their identification (i.e., listed in the VLO or not), their availability (download or through an online environment), and their metadata (size, annotation, target language, license). In section 3, we provide

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[2] https://docs.google.com/spreadsheets/d/1YO9aGhmFndaCZKZYLaKIR1MCC7pmt9vztkc0prbqSsw/edit?usp=sharing

a list of L2 corpora from languages spoken in CLARIN and that are available outside the CLARIN infrastructure. In section 4, we provide a list of non-corpus resources related to L2 teaching.

2 L2 Learner Corpora in the CLARIN infrastructure

The second language (L2) learner corpora[3] consist mainly of written essays, often argumentative essays. But there are also some spoken corpora, some mixed and even one CMC corpus including material from a course taught over Adobe connect.

There degree of inclusion of L2 learner corpora into the CLARIN infrastructure is very good.

In this section we present the identified L2 learner corpora in the CLARIN infrastructure. We first list the monolingual corpora (1. written corpora; 2. spoken corpora; 3. Video and multimodal corpora.) and then the multilingual corpora in the same order.

Table 1: Overview of L2 learner corpora found in VLO and/or national CLARIN repositories

Monolingual corpora	
Written	
CzeSL – Czech as a Second Language Czech (L1 various) 0.9 million words Download and concordancer	This corpus contains essays collected in 2013 written by non-native learners of Czech from 54 different language backgrounds. The corpus is tagged and lemmatized, and original forms and automatic corrections are assigned error labels. Additionally, texts have metadata attributes about the author and the text. The corpus is listed in the VLO and can be downloaded from the Czech repository LINDAT under the CC BY-SA 3.0 licence. For a related publication, see Rosen (2016).
The 2002 and 2006 Entrance Exam Essays of The University of Helsinki, English Philology English (L1 Finnish, Swedish) 164 essays Unavailable	This is a corpus of English written by native Finnish speakers. The corpus consists of 164 essays (293 word is the average length of an individual essay). The corpus is still under development so metadata on annotation and licence are not yet available. Otherwise, the corpus is set to be available through the Finnish Language Bank. The corpus is listed in the VLO.
British Academic Written English Corpus English (L1 English mainly, various) 2761 texts Download	This is primarily a L1 corpus although it also contains L2 data. It consists of 2761 pieces of student writing in English. It is unclear how the corpus is annotated. It can be downloaded from University of Oxford Text Archive under CC-BY. The corpus is listed in the VLO.
The Hanken Corpus of Academic Writing English	This corpus consists of written academic texts in English by students with Finnish and Swedish as the L1 background. The corpus consists of 500,000 words.

[3] L2 is here used in its broad sense including foreign language, L2, L3, L4 etc. Any language learnt which is not an L1. Many corpora also tell you something of the informant's language learning background in the metadata.

(L1 unclear) 500,000 tokens Unavailable	The corpus is not yet available and the level of annotation is unclear. When published, it will be available under a CC-BY licence. The corpus is listed in the VLO.
ETS Corpus of Non-Native Written English English (L1 various) 12,100 essays (1100 / language) Download	The corpus consists of written texts in English produced by speakers of 11 non-English native languages as part of an international text of academic English proficiency. The corpus consists of 1,100 essays for each of the 11 native languages, so 12,100 essays in total. Prompts as well as proficiency level is part of the metadata. It is unclear how the corpus is annotated. The corpus can be downloaded from the LDC catalogue under an unspecific restricted licence. The corpus is listed in the VLO.
ICLE International Corpus of Learner English English (L1 various) 3 million words Unavailable online	This is a 3 million word corpus of writing by learners of English from 14 different mother tongue backgrounds. Annotation of the corpus is unclear, as is the licence. The corpus is listed both in the VLO and LINDAT; however, the link to the external landing page is broken, so the corpus is unavailable online. However, it can be purchased on CD-ROM and a new version (ICLE v.3) is in development.
USE: The Uppsala Student English corpus English (L1 Swedish) 1.2 million tokens Download	This text corpus consists of student essays written as part of courses during the first three semesters of English studies at Uppsala University, although most of them were written during the first semester. The corpus consists of text files, each with a student ID and text ID including the course level, and information about the different prompts are available. The corpus is found in the VLO. Information about the informants is available in a separate document.
International Corpus of Learner Finnish (ICLFI) Corpus Finnish (L1 various) 1 million words Concordancer	This is a 1 million word corpus of Finnish texts (i.e. fiction and non-fiction) written by non-native speakers. The corpus is morphologically annotated. The corpus provides information on a large number of variables concerning the linguistic background of the learner, the learning task, the learning context, etc. It is available through the concordancer *Korp* (*Language Bank of Finland* distribution) under the CLARIN RES licence – i.e. restricted for individuals. The corpus is listed in the VLO. For a related publication, see Jantunen (2011).
Testipiste Corpus Finnish (L1 various) 840,000 tokens Unavailable	This is a corpus of essays in Finnish written by adult migrants of various L1 backgrounds. The corpus consists of 840,000 tokens. The corpus is still in development, so annotation is unclear. The corpus will be made available through the concordancer *Korp* under a CLARIN_RES licence. The corpus is listed in the VLO.
LAS2: The Advanced Finnish Learners' Corpus Finnish	This is a 288,500 token corpus of Finnish texts (academic writing) written by non-native MA students and collected in 2009.

Finnish (L1 unclear) 0.3 million tokens Concordancer, Download	It is unclear how the corpus is annotated. It is available through the concordancer *Korp* (*Language Bank of Finland* distribution) under the CLARIN RES licence – i.e. restricted for individuals. The corpus is listed in the VLO. The corpus consists of two subcorpora -- The Exam Essays Subcorpus and the Course Papers Subcorpus, both of which are also available through *Korp*.
KoLaS: Commented Learner Corpus Academic Writing German (L1 various) 853 texts Download	This corpus consists of written texts in German produced by students at the University of Hamburg. Students have various L1 backgrounds. The corpus consists of 853 texts. The annotation of the corpus is unclear. The corpus is available for download through the repository of the University of Hamburg under a non-specific academic licence. The corpus can be found in the VLO. More information is available here.
ASK – Norsk andrespråkskorpus Norwegian (L1 various, incl. Norwegian L1) 618,000 tokens Concordancer	This is a written corpus of Norwegian produced by non-native speakers (students) from 10 different native language backgrounds. The texts consist of written essays and tests, amounting to 618,000 tokens in the corpus. The corpus also contains L1 control essays. The corpus is PoS-tagged and displays error annotation; the concordancer also allows its users to add their own annotation. The corpus can be queried through a concordancer provided by CLARINO under the CLARIN RES licence. The corpus is not listed in the VLO.
FinSveStud 79-80 Swedish (L1 Finnish) 175,000 tokens Concordancer	This corpus consists of written texts in Swedish produced by students with Finnish as the L1 background. It consists of 175,000 tokens. Apart from the fact that the corpus is "tagged in a number of ways", annotation is unclear. The corpus can be accessed online through *Korp* under the CLARIN_RES licence. The corpus is listed in the VLO.
SW1203-essays Swedish (L1 various) 52025 tokens Download and concordancer	This is a text corpus consisting of essays written in Swedish L2. The corpus consists of 52025 tokens. The corpus displays PoS-tagging and MSD tagging and annotation of lemgrams and compound word forms. It is available for download through Språkbanken and can be queried online through Korp under CC-BY. Together with the Tisus corpus this makes up the pilot SweLL corpus (Volodina et al. 2016).
Tisus corpus Swedish Unclear size (L1 various) Concordancer	Together with the Tisus corpus this makes up the pilot SweLL corpus (Volodina et al. 2016). The corpus consists of 59639 tokens. The corpus is annotated with MSD and for compounds and lemgrams. It can be downloaded from Sprakbanken and queried through *Korp* under CC-BY. It cannot be found in the VLO.
Spoken	
Arabic Learner Corpus Arabic (L1 various) 0.3 million tokens Download	This is a corpus of recordings and associated transcripts in Arabic produced by students of 67 different nationalities. The corpus consists of 280,000 tokens.

	It is unclear how the corpus is annotated. The corpus can be downloaded from the LDC catalogue under an unspecified restricted licence. The corpus is listed in the VLO.
LUCEA: Longitudinal Corpus of University College English Accents English (L1 various) 1100 recordings, approx. 20 minutes each Concordancer	This is an audio corpus of recordings in English produced by both L1 and L2 learners. The corpus consists of approximately 1100 recordings averaging 20 minutes each. It is unclear how the corpus is annotated. The corpus can be accessed online through a CLARIN node under a restricted licence. The corpus is listed in the VLO. For a related publication, see Orr and Quené (2017).
The Anglish Corpus English (L1 French) 60 speakers	This is a spoken corpus of English materials produced by learners with French as the L1 language. The corpus is annotated for interpausal units. It is available for download through the ORTOLANG repository.
GLBCC (Giessen - Long Beach Chaplin Corpus) English (L1 various, mainly German) Unclear size Download	This is a corpus containing recordings and transcriptions of English spoken both by native speakers and non-native speakers (mostly German). The total size is unclear, although transcripts average 2472 word each. It is unclear how the corpus is annotated. It can be downloaded from the University of Oxford Text archive under the CC-BY licence. It is listed in the VLO.
A Learners' Corpus of Reading Texts English (L1 French) Size unknown (54 speakers) Download	This audio corpus consists of unprepared readings of English texts by first-year students at an English department who speak French as a native language. The size of the corpus is unclear but consists of 54 speakers. It is unclear how the corpus is annotated. It can be downloaded from ORTOLANG under the CLARIN RES licence. The corpus is listed in the VLO.
ISLE Speech Corpus English (L1s: German & Italian) c. 18 hours Download	This is an audio corpus of English spoken by German and Italian learners. The corpus consists of 17 hours, 54 minutes and 44 seconds worth of speech data. The annotation constitutes automatic phone-level annotation and manual annotation of phone and stress errors. The corpus is available for download in the ELRA catalogue. It is listed in the VLO
French Learner Language Oral Corpora (FLLOC) French (L1 English, Dutch) 1375 transcripts Download	This is a corpus of audio recordings and associated transcriptions of French produced by native English and Dutch learners. The corpus consists of 1375 transcripts. The corpus is MSD-tagged. The transcripts can be downloaded from University of Oxford Text Archive under CC-BY. The corpus is listed in the VLO. For related publications, see this hyperlink.

Hamburg Modern Times Corpus German (L1 various) 24,000 words Unavailable	This audio corpus consists of recordings of German spoken by non-native speakers and the accompanying transcriptions. The written transcriptions consist of 24464 words. The corpus is characterized by manual annotation of phonetic phenomena and prosody. It is unclear if the corpus is available. The corpus is listed in the VLO under a restricted licence.
LANGMAN Hungarian (L1 Chinese) Size unknown Download	The VLO lists 11 subcorpora of LANGMAN, which is a spoken corpus of Hungarian produced by Chinese learners. The total size is unknown. The 11 subcorpora can be downloaded from the VLO under CC-BY.
Spoken	
AixOx English and French (L1s: French, English) 40 minutes Download	This is an audio corpus of French and English consisting of 40 1-minute presentations of non-native learners of French and English (with these two languages as L1 backgrounds). It is unclear how the corpus is annotated. The corpus can be downloaded from the Ortolang repository under a restricted licence.
Openprodat Dutch, English, French, German, Italian, Arabic, Spanish, Hungarian, Japanese, Thai, Norwegian, Chinese (L1 various) Unclear size: 24 speakers (2013) Download	This is an audio corpus of various languages, available as separate subcorpora. It consists of paragraph readings by participants in both their L1 and in as many L2 as they felt they could manage (Hirst et al 2013). Link to Ortolang repository, where the corpus can be downloaded under the public Publique Générale GNU licence.
Written	
LETEC (Learning and Teaching Corpus) English and French (L1 various) 7 subcorpora Unavailable.	Seven subcorpora of LETEC are listed in the VLO. The corpora contain L2 texts in English and French. The total size is unknown. It is unclear how the corpus is annotated. The corpus is unavailable since the download links are broken in the VLO. The licence is listed as CC-BY.
LeaP: The Learning the Prosody of a Foreign Language - a phonological corpus of Learner English and Learner German English; German (L1s: various) 12 hours Download	This is an audio corpus of English and German spoken by non-native speakers from 31 different native language backgrounds. The corpus consists of 12 hours' worth of recordings. The recordings were annotated manually and automatically on 8 different tiers including pitch, tones, segments, syllables, words, phrasing, parts-of speech and lemmata. Each file can be downloaded from the VLO under an unknown licence.
CEFLING Project Corpus Finnish and English (L1 Various)	This corpus consists of written texts in Finnish and English collected from primary secondary school students (years 7-9). The size of the corpus is unclear.

Unknown size Unavailable	It is unclear how the corpus is annotated. The corpus is unavailable, since the link to the landing page is broken, but latest project webpage says that one should contact the distributor. The corpus is listed in the VLO under an unknown licence.
DIALUKI: Diagnosing reading and writing in a second or foreign language Finnish, English (L1 Finnish, Russian, English) 8,600 texts Unavailable	This is a corpus of texts both in Finnish (produced by Russian learners) and English (produced by Finnish learners). The corpus consists of 8,600 texts. Metadata on annotation are not available. The corpus will be made available through *Korp* under a CLARIN_RES licence. The corpus is listed in the VLO.
Topling - Paths in Second Language Acquisition Finnish, English, Swedish (L1 various) 0.16 million tokens Concordancer	This corpus consists of written texts in English, Swedish and Finnish produced by students in the Finnish educational system and is an extension of the CEFLING corpus which it also includes. The corpus consists of approximately 165,000 tokens. The annotation of the corpus is unclear. Each subcorpus (i.e. English, Swedish and Finnish) can be accessed through the concordancer *Korp*. The corpus is available under the CLARIN End User Licence Agreement with various restrictions (e.g. non-commercial purposes; prohibited distribution to third-parties).
Video	
English as a Foreign Language Corpus English (L1 Finnish; Swedish?) 24 hours Unavailable	The corpus consists of video materials – i.e., videotaped lessons both in English performed by non-native speakers at Finnish secondary schools. The videos are 24 hours long it total. It is unclear how the corpus is annotated. The corpus is currently unavailable. Licence etc. are under negotiation.
Multimodal	
The Long Second Corpus Finnish (L1 mainly Estonian, Russian) Unknown size Unavailable	This is a multimodal corpus consisting of materials in Finnish produced by immigrants from various backgrounds (Estonian, Macedonian, Kurdish, Portuguese, Russian, English). The corpus is still in preparation so all the metadata are unavailable. It is set to be made available on the LAT platform under the CLARIN_RES licence. The corpus is listed in the VLO.
Multilingual corpora	
Video	
Content-and-Language-Integrated Learning Corpus Finnish and English (L1 mainly Finnish) 51 hours Unavailable	The corpus consists of video materials – i.e., videotaped lessons both in English and Finnish performed by non-native speakers at Finnish lower secondary schools. The videos are 51 hours long it total. It is unclear how the corpus is annotated. It is unclear how the corpus is available, since the VLO record links only to the starting webpage of FIN-CLARIN.
Multimodal	
YKI National Certificates corpus	This corpus consists of written and spoken materials in Italian, Swedish, Spanish, English, German, French and Russian. All the other metadata are

Italian, Swedish, Spanish, English, Finnish, German, French, Russian (L1 various) Unclear size Unavailable	unclear, as is the availability of the corpus. It is listed in the VLO under the CLARIN_RES licence.
TAITO: Written and Oral Data of the TAITO-project English, French, German, Italian, Swedish (L1 various, mainly FInnish probably) Size unknown Unavailable	This corpus consists of written, audio and video materials in English, French, German, Italian and Swedish. The corpus is still in development so all the metadata are unknown.

2.1 Summary

2.1.1 Identification

Table 1 lists a total of 36 L2 learner corpora that are integrated with the CLARIN infrastructure. All can be found in the VLO except for the following 3 corpora:

(i) ASK – Norsk andrespråkskorpus
(ii) SW1203-essays
(iii) Tisus corpus

Corpus (i) was identified on the webpage of CLARINO (but not its repository)[4] and corpora (ii) and (iii) through Språkbanken.

2.1.2 Availability

SW1203-essays, Tisus Corpus and The Advanced Finnish Learners' Corpus Finnish are available both for download and through the concordancer *Korp* while CzeSL – Czech as a Second Language is available for download through LINDAT and can be queried through *Kontext*.

The following 4 corpora are available only through a concordancer:

(i) International Corpus of Learner Finnish (ICLFI) Corpus
(ii) ASK – Norsk andrespråkskorpus
(iii) Topling - Paths in Second Language Acquisition
(iv) LUCEA: Longitudinal Corpus of University College English Accents

In the case of corpora (i) and (iii), the concordancer is *Korp* (Språkbanken and Language Bank of Finland distribution), while corpus (ii) can be accessed (though under a restricted licence) by a dedicated concordancer provided by CLARINO.

[4] https://repo.clarino.uib.no

The following 16 corpora are available only for download:

(i) Arabic Learner Corpus, through the LDC catalogue;
(ii) SamtaleBank Dansk_som_Andetsprog Corpus, through a dedicated page provided by the Talk Bank;
(iii) ANGLISH, through ORTOLANG;
(iv) GLBCC (Giessen - Long Beach Chaplin Corpus, through the University of Oxford Text archive;
(v) ISLE Speech Corpus, through the ELRA catalogue;
(vi) LeaP: The Learning the Prosody of a Foreign Language, through the VLO;
(vii) British Academic Written English Corpus, through the University of Oxford Text Archive;
(viii) A Learners' Corpus of Reading Texts, through ORTOLANG;
(ix) ETS Corpus of Non-Native Written English, through the LDC catalogue;
(x) French Learner Language Oral Corpora (FLLOC), through the University of Oxford Text Archive;
(xi) Commented Learner Corpus Academic Writing, through the repository of the University of Hamburg;
(xii) AixOx, through ORTOLANG;
(xiii) Pedagogical Greek L2 textbooks corpus, through the repository of *clarin:el*;
(xiv) LANGMAN, through the VLO;
(xv) The Uppsala Student English corpus, through the University of Oxford Text Archive;
(xvi) Openprodat, through ORTOLANG.

The following 14 corpora are unavailable:

(i) ICLE International Corpus of Learner English
(ii) English as a Foreign Language Corpus
(iii) Hamburg Modern Times Corpus
(iv) LETEC (Learning and Teaching Corpus)
(v) CEFLING Project Corpus
(vi) Content-and-Language-Integrated Learning Corpus
(vii) The 2002 and 2006 Entrance Exam Essays of The University of Helsinki, English Philology
(viii) The Hanken Corpus of Academic Writing
(ix) Testipiste Corpus
(x) The Long Second Corpus
(xi) DIALUKI: Diagnosing reading and writing in a second or foreign language
(xii) Content-and-Language-Integrated Learning Corpus
(xiii) YKI National Certificates corpus
(xiv) TAITO: Written and Oral Data of the TAITO-project

In the case of corpora (i), (iv) and (v), the download links listed in the VLO are broken. Corpora (ii) and (vi) seem to be unavailable because the hyperlinks given to the VLO redirect to the main page of FIN-CLARIN (and not the resource landing page). Corpora (vii)-(xiv) are still in preparation.

2.1.3 Metadata

2.1.3.1 Languages covered and type of data

24 of the 36 corpora are monolingual. The following target languages are represented in the monolingual corpora:

- English (10 corpora);
- Finnish (4 corpora);
- Swedish (3 corpora);
- German (2 corpora);
- Arabic (1 corpus);
- Czech (1 corpus);
- French (1 corpus);
- Hungarian (1 corpus);
- Norwegian (1 corpus).

The multilingual and bilingual corpora cover the following languages:

- English and Finnish (3 corpora);
- English and French (2 corpora);
- English and German (1 corpus);
- Dutch, English, French, German, Italian, Arabic, Spanish, Hungarian, Japanese, Thai, Norwegian, Chinese (1 corpus);
- Italian, Swedish, Spanish, English, German, French, Russian (1 corpus);
- English, French, German, Italian, Swedish (1 corpus);
- English, Finnish and Swedish (1 corpus).

2.1.3.2 Size

Information on size (i.e. tokens/number of documents or temporal length in case of video/audio corpora) is missing for the following corpora 11 out of 36 corpora:

(i) SamtaleBank Dansk_som_Andetsprog Corpus
(ii) ANGLISH
(iii) GLBCC (Giessen - Long Beach Chaplin Corpus)
(iv) A Learners' Corpus of Reading Texts
(v) LANGMAN
(vi) LETEC (Learning and Teaching Corpus)
(vii) CEFLING Project Corpus
(viii) The Long Second Corpus
(ix) Openprodat
(x) YKI National Certificates corpus
(xi) TAITO: Written and Oral Data of the TAITO-project

The largest corpus in terms of token size is ICLE International Corpus of Learner English, which consists of 3 million tokens. Amongst the audio/video corpora, Content-and-Language-Integrated Learning Corpus is the longest, consisting of 51 hours' worth of videos.

2.1.3.3 Annotation and licence

Information on linguistic annotation is not available for the following 24 out of 36 corpora:

(i) Arabic Learner Corpus
(ii) SamtaleBank Dansk_som_Andetsprog Corpus
(iii) GLBCC (Giessen - Long Beach Chaplin Corpus)
(iv) British Academic Written English Corpus
(v) A Learners' Corpus of Reading Texts
(vi) ETS Corpus of Non-Native Written English

(vii) ICLE International Corpus of Learner English
(viii) The Advanced Finnish Learners' Corpus
(ix) English as a Foreign Language Corpus
(x) Commented Learner Corpus Academic Writing
(xi) LANGMAN
(xii) Pedagogical Greek L2 textbooks corpus
(xiii) LETEC (Learning and Teaching Corpus)
(xiv) CEFLING Project Corpus
(xv) Topling - Paths in Second Language Acquisition
(xvi) Content-and-Language-Integrated Learning Corpus
(xvii) LUCEA: Longitudinal Corpus of University College English Accents
(xviii) The Long Second Corpus
(xix) AixOx
(xx) Openprodat
(xxi) DIALUKI: Diagnosing reading and writing in a second or foreign language
(xxii) Content-and-Language-Integrated Learning Corpus
(xxiii) YKI National Certificates corpus
(xxiv) TAITO: Written and Oral Data of the TAITO-project

Otherwise, the annotation varies greatly from corpus to corpus. One of the text corpora is morphologically annotated (International Corpus of Learner Finnish (ICLFI) Corpus), while 2 text corpora display error annotation aside from tokenisation (ASK – Norsk andrespråkskorpus and CzeSL – Czech as a Second Language). Annotation of the audio corpora covers the markup of prosody errors (ISLE Speech Corpus), general prosody annotation (Hamburg Modern Times Corpus), general "interpausal units annotation" (ANGLISH) and 8-level annotation of pitch, tones, segments, syllables, words, phrasing, parts-of-speech and lemmata (LeaP: The Learning the Prosody of a Foreign Language).

Information on licence is missing for the following 7 corpora:

(i) ISLE Speech Corpus
(ii) LeaP: The Learning the Prosody of a Foreign Language
(iii) ICLE International Corpus of Learner English
(iv) English as a Foreign Language Corpus
(v) CEFLING Project Corpus
(vi) Content-and-Language-Integrated Learning Corpus
(vii) AixOx
(viii) The 2002 and 2006 Entrance Exam Essays of The University of Helsinki, English Philology

Otherwise, 12 corpora are available under CC-BY, 14 corpora are listed under restricted licences and the remainder under miscellaneous licences.

3 L2 Learner Corpora from CLARIN countries outside the CLARIN infrastructure

Apart from the L2 learner corpora included in the CLARIN infrastructure there are also many L2 learner corpora in the CLARIN countries which have not been included in the infrastructure for various reasons. Below we list those that we are aware of at the moment.

Table 2: Overview of L2 Learner corpora outside the VLO and/or national CLARIN repositories, but in countries within CLARIN

Written	
ALEC: Advanced learner corpus of English English (L1 mainly Swedish, but some other L1 and some L1 English) 1.3 million tokens Unavailable	This text corpus contains essays written by students of English (Uppsala University) in their third-fifth semester of studying English. Each file includes metadata about language background and year of study etc.
The Anglia Polytechnic University (APU) Learner Spanish Corpus Spanish (L1 various) c. 120 000 Unavailable	This contains written compositions by learners of Spanish with different L1s. Annotation is unclear and it appears to be currently unavailable.
The ASU (andraspråkets strukturutveckling) corpus Swedish Unavailable	This corpus of Swedish L2 was compiled by Björn Hammarberg and uses the ITG system.
THE BAT MAT corpus: EFL academic writing English (L1 Swedish, Finnish) 90 BA/MA dissertations Unavailable	This corpus was created to study academic writing. It contains BA and MA dissertations of Finland-Swedish students at the University of Turku, Finland. There is background information about all writers however it unclear what the format of the corpus is and if it has been annotated.
CLC: The Cambridge Learner Corpus English (L1 various) c. 50 million Partly available	This well-known corpus contains essays from learners of many different languages. The essays have been transcribed and error-annotated. Part of the CLC is available through SketchEngine. An earlier version of the corpus was described in a publication by D. Nicholls (2003).

EIC: The Estonian Interlanguage Corpus of Tallinn University Estonian (L1 various) Concordancer	This is a corpus of texts written by learners of Estonian as a second / foreign language. It can be searched online through a specific search interface. Annotation is unclear.
The Finnish as a Second Language (S2) Matriculation Examination Essay Collection 2001-2002 Corpus Finnish (L1 Swedish?) Unavailable	This is not yet included in the VLO. FIN-CLARIN states that negotations about availability and licence are currently under way.
The Japanese learner corpus of Spanish Spanish (L1 Japanese) c. 83 000 Unavailable	It is unclear exactly what this corpus contains and whether it has been annotated. The corpus appears to be unavailable. Contact person: Yoshihito Kamakura
LANCAWE: The Lancaster Corpus of Academic Written English English (L1 various) Unavailable	This corpus consists of IELTS academic writing tests according to the Université catholique de Louvain list of L2 learner corpora. Weisser's list however claims that it consist of more varied writing from courses in academic L2 English. The corpus also contains a small control corpus of L1 data. Annotation is unclear however there are subcorpora according to task, writer, L1 etc. Contains longitudinal data.
Learner's corpus at the University of Tartu Estonian (L1 Various) c. 300 000 words Availability unclear	The information available about this corpus is only in Estonian and rather sparse. Contact person: kristiina.praakli@ut.ee
The LINCS Corpus German (L1 English, German)	This corpus is under development according to the UCL list. Contact person: Elizabeth Thoday, Heriot-Watt University Edinburgh, UK.
Longi Corpus Swedish (L1 Finnish) 150 000 words	This consists of texts written by upper secondary students in Finland with Finnish as their L1. The corpus has been POS tagged. Licence: under negotiation (FIN-CLARIN)

unavailable	
The Longman Learners' Corpus English (L1 various) c. 10 million Availability unclear	This corpus contains essays and exam scripts according to the UCL list. The UCL list it as a commercial, however availability is unclear.
The Lund CEFLE Corpus (Corpus Écrit de Français Langue Étrangère) French (L1: Swedish plus L1 French control group) c. 100 000 words Available here	This text corpus contains essays written by Swedish learners of French (15-19 years old) as well as an L1 control group of the same age. The essays are available as text files without annotation. The corpus has been used within the Direkt profil project which provides some form of annotation but not quite clear which.
The Tartu Learner Corpus of Spanish as a L3+ Spanish (L1 Estonian) 885 000 words	This corpus contains texts written by Estonian-speaking learners of Spanish. Contact person: Mari Kruse, University of Tartu, Estonia
Spoken	
CLAG & GAP: Comasan Labhairt ann an Gàidhlig & Gaelic Adult Proficiency Gaelic (L1 various) Unavailable	This spoken corpus is based on a variety of oral tasks performed by the learners. It is unclear how it has been annotated and to whom it is available.
ProoF - Pronunciation of Finnish by Immigrants in Finland, version 1.0 Finnish (L1 various) 20 hours Available through LAT	This corpus contains read aloud and spontaneous speak by L2 speakers as well as an L1 control group. The corpus ha been transcribed and annotated on several levels in Praat (FIN-CLARIN).
SPLLOC: Spanish Learner Language Oral Corpus Spanish (L1 English)	This corpus consists of recordings of learners of Spanish at different levelss as well as an L1 control group. The material contains narratives, interviews and picture description tasks.

Download Search	The corpus contains orthographic transcriptions and some error annotation. POS tagging and morphosyntactic tagging appears to have been done or be under way.

4 Textbook corpora in the CLARIN infrastructure

Apart from corpora of L2 learner texts there are also a few corpora of course book texts used in teaching L2 languages. These can be used to study the input that learners receive but also the expected receptive skills that they are expected to have at different proficiency levels.

So far we only know of two such corpora within the CLARIN countries, none of which are listed in the VLO, but the COCTAILL-corpus is available through Språkbanken's concordancer *Korp*.

Course book corpora	
COCTAILL Swedish c. 710 000 tokens Concordancer	This corpus is a collection of the reading texts in course books for Swedish as a second / foreign language at different levels. The corpus has been POS-tagged, syntactically annotated as well as annotated with CEFR-levels. More information in: Volodina et al 2014. Pilan et al 2016
Pedagogical Greek L2 textbooks corpus Greek 276,000 tokens Download	This corpus is a compilation of texts from Greek L2 textbooks. The corpus consists of 276,000 tokens. It is unclear how the corpus is annotated. The corpus is available for download from the clarin:el repository under the CC-BY licence. It is not listed in the VLO.