Welcome !
CLARIN café
Bilingual and Multilingual corpora

Organisers: Thomas Gaillat, Franck Cinato, Eva Soroli

Support: CLARIN, CORLI K-Centre

CLARIN

Welcome !

CLARIN café

Bilingual and Multilingual corpora

Organisers: Thomas Gaillat, Franck Cinato, Eva Soroli

Support: CLARIN, CORLI K-Centre

CLARIN

CLARIN café
Bilingual/Multilingual corpora

Programme

14.00-14.15

The European Infrastructure CLARIN and its Knowledge Centres, Eva SOROLI, Univ. of Lille, France

14.15-14.30

CORLI (Corpus, Language and Interactions): a CLARIN Knowledge-Centre, Christophe PARISSE, Univ. of Nanterre & Céline POUDAT, Univ. Côte d'Azur

14.30-14.50

The multidialectal corpus of the Crescent dialects: collection, exploitation and analysis, Maximilien GUERIN, University of Paris & CNRS - HTL (UMR 7597), France

14.50-15.00 Questions & Discussion

15.00-15.20

Building CIEP+, the parallel Corpus of Indo-European Prose Plus, Annemarie VERKERK & Luigi TALAMO Universität des Saarlandes, Germany

15.20-15.30 Questions & Discussion

15.30-15.15.50

A dynamic architecture *to structure and analyse comparable learner corpora***:** the case of the French and English Corpus InterLangue (CIL), Thomas GAILLAT, University of Rennes, LIDILE, France

15.50-16.00 Questions & Discussion

16.00-16.15 Wrap-up Session : Franck CINATO

CLARIN ERIC

The European Research Infrastructure and its knowledge centres

Eva SOROLI

CLARIN Ambassador

University of Lille, STL CNRS, CLARIN

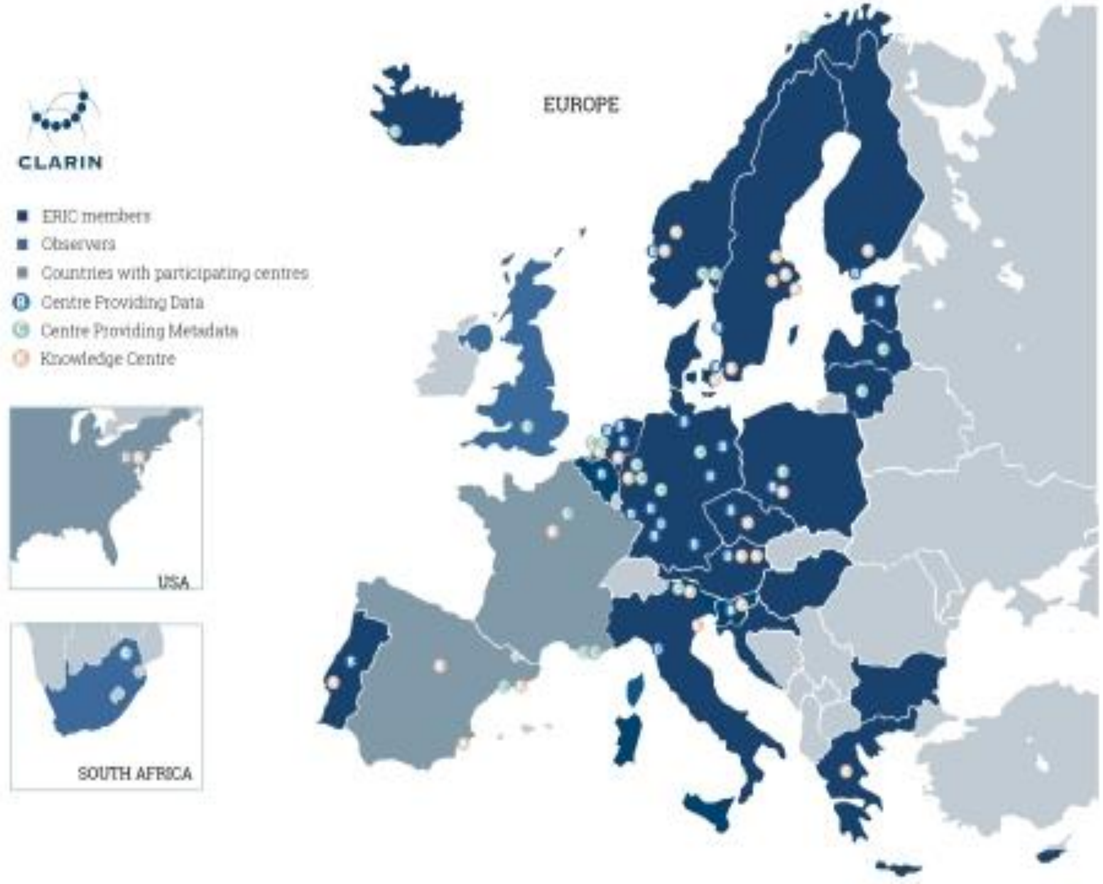
efstathia.soroli@univ-lille.fr

CLARIN ...

- CLARIN : Common Language Resources and Technology Infrastructure
- has the ESFRI **ERIC** status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond to
  - **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing**
- is an integral part of **the European Open Science Cloud (EOSC)**
  - See **clarin.eu/eosc**

CLARIN today

- **70 centres**
- **22 members**: (AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO PL, PT, SE, SI)
- **2 observers:** UK, ZA
- **1 third-part partner :** USA



Ready-To-Use Language Resources

Language resources :

- speech and language data types
- in machine readable form
- tools and services for the processing of language data
- software tools for the preparation, collection, management, conversion, use/re-use of data

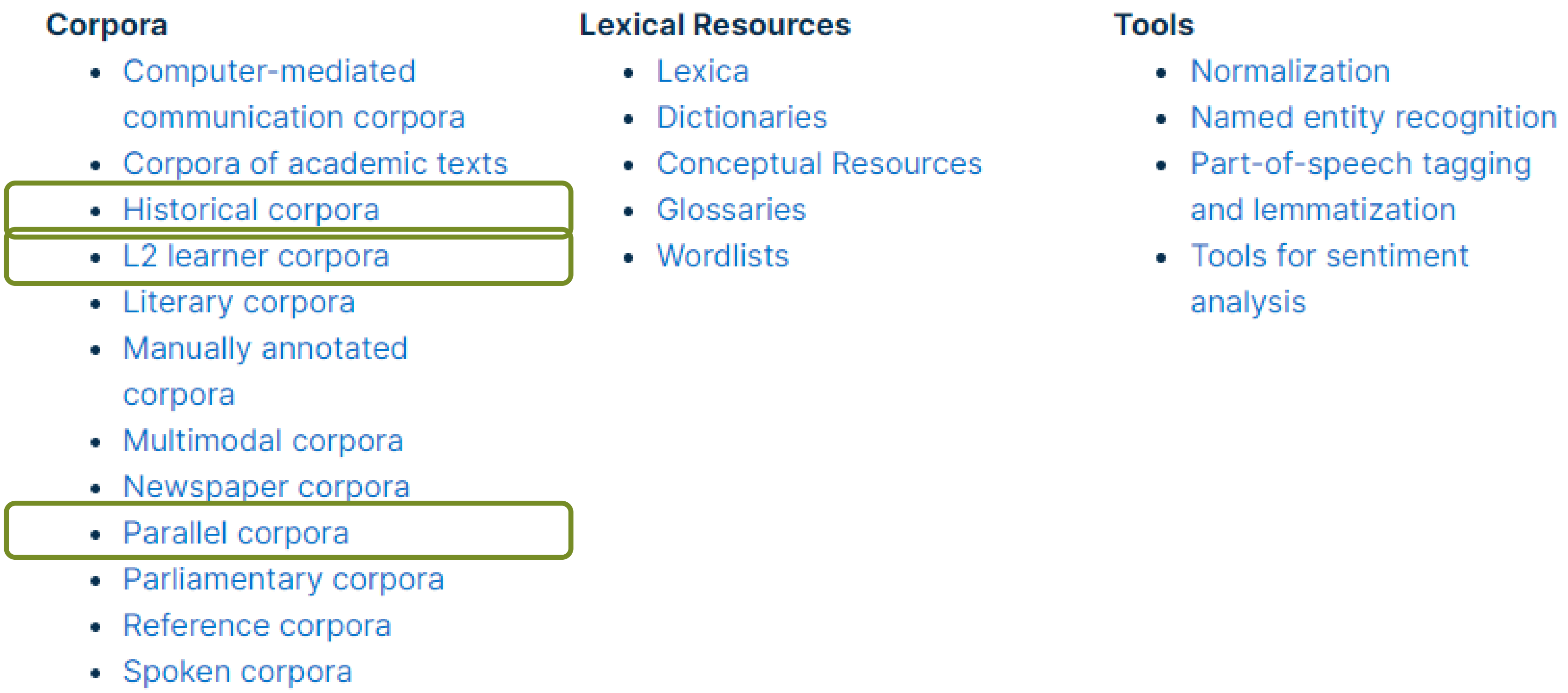
Examples of language resources :

- Written or spoken corpora and lexicons
- Multi-modal resources
- Grammars
- Terminology or domain specific databases and dictionaries
- Ontologies
- Multimedia databases
- Corpus management and exploration systems,
- OCR systems,
- Pipelines,
- Speech processing systems,
- Machine translation systems,
- Environments for annotation and evaluation etc.

CLARIN in Communities of Use

- Digital Humanities
- Linguistics and Philology
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- Teachers
- Industry and Professionals
- General Public
- ….

CLARIN Resource Families - CRFs



Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora
- Sign languages (coming soon)

Lexical Resources

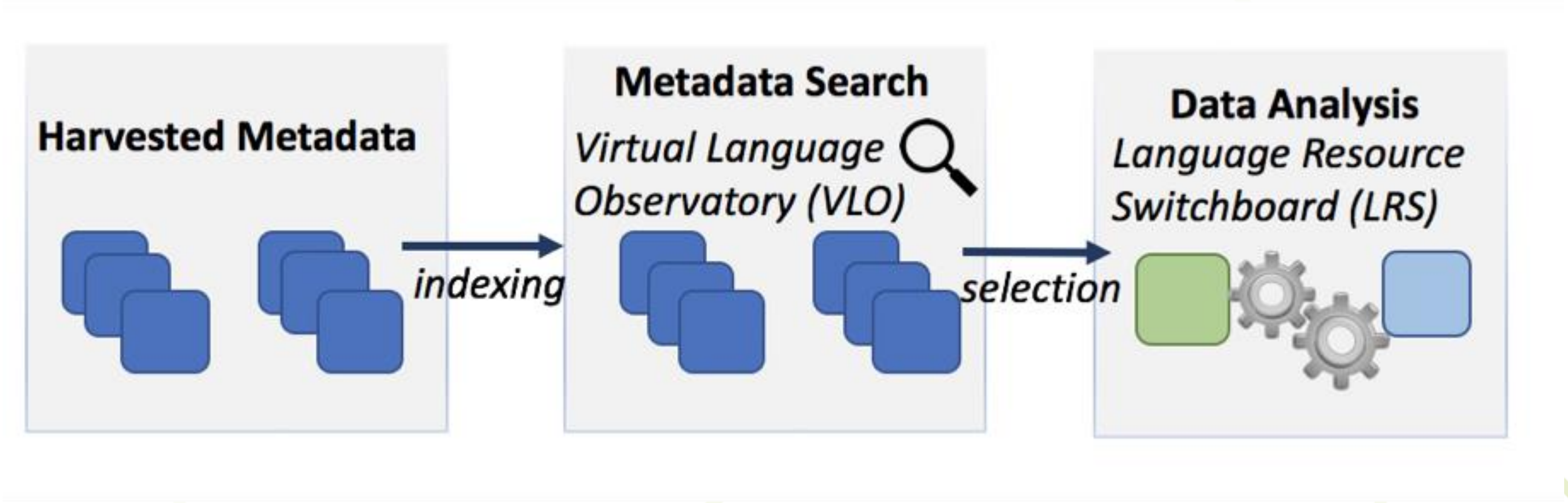
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

See also more information about the CLARIN Resource Families initiative here: https://www.clarin.eu/resource-families

The Technical Infrastructure



Harvested Metadata

indexing

Metadata Search

Virtual Language Observatory (VLO)

selection

Data Analysis

Language Resource Switchboard (LRS)

Findable Accessible Interoperable Reusable

clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

See also here for other useful services : https://www.clarin.eu/content/services

The Knowledge Infrastructure

Knowledge centres

Digital Humanities Course Registry

Tour de CLARIN

Impact Stories

TEACHING WITH CLARIN

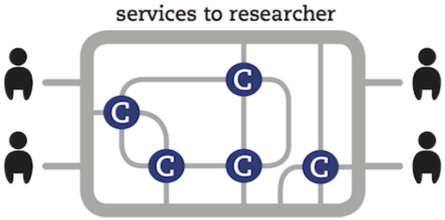
VideoLectures

Funding Opportunities

Support for EU-funded projects

https://www.clarin.eu/content/clarin-for-researchers

A distributed network of knowledge



- **B-centres** (Service Providing Centres) : often a university or an academic institute, offer the scientific community access to resources, services and knowledge on a sustainable basis.
- **C-Centres** (Metadata Providing Centres): Their metadata are integrated with CLARIN, but they need not offer any further services.
- **K-Centres** (Knowledge Centres): Centres sharing their knowledge and expertise on one or more aspects of the domain covered by CLARIN.

K-Centres

- Individual languages (e.g. Danish, Czech, Portuguese), language families (e.g. South Slavic) or groups of languages (e.g. morphologically rich languages, the languages of Sweden)
- Written text and modalities other than written text (e.g. spoken language, sign language)
- Linguistic topics (e.g. language diversity, language learning, diachronic studies)
- Language processing topics (e.g. speech analysis, building treebanks, machine translation)
- Data types other than corpora (e.g. lexical data, word nets, terminology banks)
- Using or processing families of language data that will exist for most languages (e.g. newspapers, parliamentary records, oral history)
- Generic methods and issues (e.g. data management, ethics, IPR, OCR)

Getting involved in CLARIN

- Deposit your data in a CLARIN data centre
- Contact a specialized K-Centre
- Contact your National Coordinator, your national UI representative, CLARIN ambassadors
- Read the Tour de CLARIN to find out about national activities
- Join our NewsFlash and our mailinglists
- Check out our events, funding opportunities and calls
- Check our **#clarincafe**
- Follow us on Twitter @CLARINERIC

CLARIN café
Bilingual/Multilingual corpora

Programme

14.00-14.15

The European Infrastructure CLARIN and its Knowledge Centres, Eva SOROLI, Univ. of Lille, France

14.15-14.30

CORLI (Corpus, Language and Interactions): a CLARIN Knowledge-Centre, Christophe PARISSE, Univ. of Nanterre & Céline POUDAT, Univ. Côte d'Azur

14.30-14.50

The multidialectal corpus of the Crescent dialects: collection, exploitation and analysis, Maximilien GUERIN, University of Paris & CNRS - HTL (UMR 7597), France

14.50-15.00 Questions & Discussion

15.00-15.20

Building CIEP+, the parallel Corpus of Indo-European Prose Plus, Annemarie VERKERK & Luigi TALAMO Universität des Saarlandes, Germany

15.20-15.30 Questions & Discussion

15.30-15.15.50

A dynamic architecture *to structure and analyse comparable learner corpora***:** the case of the French and English Corpus InterLangue (CIL), Thomas GAILLAT, University of Rennes, LIDILE, France

15.50-16.00 Questions & Discussion

16.00-16.15 Wrap-up Session : Franck CINATO

CLARIN café
Bilingual/Multilingual corpora

Programme

14.00-14.15

The European Infrastructure CLARIN and its Knowledge Centres, Eva SOROLI, Univ. of Lille, France

14.15-14.30

CORLI (Corpus, Language and Interactions): a CLARIN Knowledge-Centre, Christophe PARISSE, Univ. of Nanterre & Céline POUDAT, Univ. Côte d'Azur

14.30-14.50

The multidialectal corpus of the Crescent dialects: collection, exploitation and analysis, Maximilien GUERIN, University of Paris & CNRS - HTL (UMR 7597), France

14.50-15.00 Questions & Discussion

15.00-15.20

Building CIEP+, the parallel Corpus of Indo-European Prose Plus, Annemarie VERKERK & Luigi TALAMO Universität des Saarlandes, Germany

15.20-15.30 Questions & Discussion

15.30-15.15.50

A dynamic architecture *to structure and analyse comparable learner corpora***:** the case of the French and English Corpus InterLangue (CIL), Thomas GAILLAT, University of Rennes, LIDILE, France

15.50-16.00 Questions & Discussion

16.00-16.15 Wrap-up Session : Franck CINATO

CLARIN café
Bilingual/Multilingual corpora

Programme

14.00-14.15

The European Infrastructure CLARIN and its Knowledge Centres, Eva SOROLI, Univ. of Lille, France

14.15-14.30

CORLI (Corpus, Language and Interactions): a CLARIN Knowledge-Centre, Christophe PARISSE, Univ. of Nanterre & Céline POUDAT, Univ. Côte d'Azur

14.30-14.50

The multidialectal corpus of the Crescent dialects: collection, exploitation and analysis, Maximilien GUERIN, University of Paris & CNRS - HTL (UMR 7597), France

14.50-15.00 Questions & Discussion

15.00-15.20

Building CIEP+, the parallel Corpus of Indo-European Prose Plus, Annemarie VERKERK & Luigi TALAMO Universität des Saarlandes, Germany

15.20-15.30 Questions & Discussion

15.30-15.15.50

A dynamic architecture *to structure and analyse comparable learner corpora***:** the case of the French and English Corpus InterLangue (CIL), Thomas GAILLAT, University of Rennes, LIDILE, France

15.50-16.00 Questions & Discussion

16.00-16.15 Wrap-up Session : Franck CINATO

CLARIN café
Bilingual/Multilingual corpora

Programme

14.00-14.15

The European Infrastructure CLARIN and its Knowledge Centres, Eva SOROLI, Univ. of Lille, France

14.15-14.30

CORLI (Corpus, Language and Interactions): a CLARIN Knowledge-Centre, Christophe PARISSE, Univ. of Nanterre & Céline POUDAT, Univ. Côte d'Azur

14.30-14.50

The multidialectal corpus of the Crescent dialects: collection, exploitation and analysis, Maximilien GUERIN, University of Paris & CNRS - HTL (UMR 7597), France

14.50-15.00 Questions & Discussion

15.00-15.20

Building CIEP+, the parallel Corpus of Indo-European Prose Plus, Annemarie VERKERK & Luigi TALAMO Universität des Saarlandes, Germany

15.20-15.30 Questions & Discussion

15.30-15.15.50

A dynamic architecture *to structure and analyse comparable learner corpora*: the case of the French and English Corpus InterLangue (CIL), Thomas GAILLAT, University of Rennes, LIDILE, France

15.50-16.00 Questions & Discussion

16.00-16.15 Wrap-up Session : Franck CINATO