Social media corpora, datasets and tools: An overview

Darja Fišer
Director for User Involvement CLARIN ERIC
Darja.Fiser@ff.uni-lj.si

Jakob Lenardič
Assistant to Director for User Involvement CLARIN ERIC
jakob.lenardic@ff.uni-lj.si

CLARIN-PLUS workshop "Creation and Use of Social Media Resources"

Kaunas, Lithuania

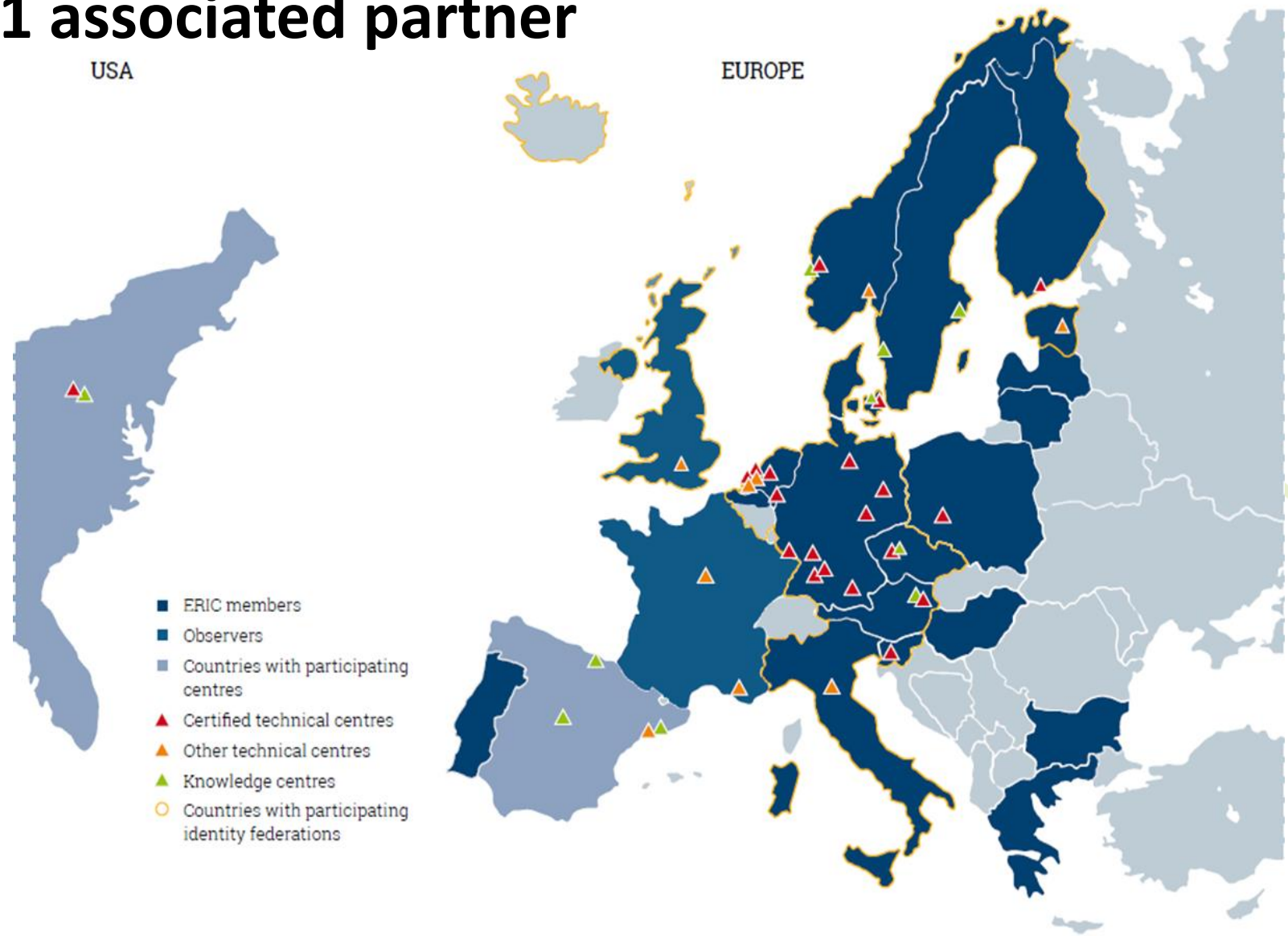
18-19 May 2017

CLARIN

CLARIN in five bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form),
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment.

CLARIN ERIC: 19 members, 2 observers, 1 associated partner



Prehistory of this workshop

- H2020 project CLARIN-PLUS: outreach to new users, focus on four specific data types
  - oral history collections
  - newspaper archives
  - parliamentary records
  - social media data
- Joint proposals and research collaboration
  - national initiatives
  - R&D proposals (FP7, H2020, COST)
  - international projects

Long-term vision

- easy access to social media material
- services suited for this type data can easily be found and employed
- encourage researchers to develop and address discipline-specific hypotheses and scholarly questions

Challenges and multidisciplinary potential

Social media data sets are considered a rich data type that

- is suited for both *close reading* and *distance reading*
- is often presenting itself as messy or noisy data
- is calling for links with data in other modalities than text
- under specific circumstances that need to be well understood before strong conclusions can be drawn

Social media data sets have a big potential for reuse and re-purposing within many fields of study in the humanities and social sciences (and beyond):

- *Humanities*: language variation and change, discourse analysis, …
- *Social sciences*: social and cultural dynamics, political sciences, economics, …

Aims for this workshop

- explore existing and envisoned approaches for analyzing social media records (text, multimedia) with the use of CLARIN-compatible standards and processing tools
- generation of an overview of relevant resources to stimulate synergy and cross-country collaboration
- creation of an action plan

Corpora

- Surveyed for all CLARIN members
- 15 corpora identified
  - 12 different lang (de, uk, ee, fi, nl, lt, cy, fr, no, pl, it, si)
  - most common data types: forums, blogs, tweets
  - 4 available through concordancer, 4 for download, 2 both download and concordancer, 5 unavailable
  - License info available for 7 corpora
    - 4 under CC BY: Suomi24 Sentence Corpus, Dortmund Chat Corpus, CoMeRe, DWDS-Blogs
    - Re-licensing forbidden: Monitor Corpus of Austrian Tweets
    - ACA_CLARIN-LT_End-User-Licence-Agreement_EN-LT: LITIS
    - Restricted in accordance with Twitter Terms of Use: Corpus of Welsh Language Tweets
  - Most for German (4 corpora)
- 6 corpora available through the CLARIN infrastructure
  - ee: Mixed Corpus: New Media — CLARIN ESTONIA
  - fi: Suomi24 — VLO
  - lt: LITIS — VLO
  - nl: SoNaR New Media Corpus — VLO
  - no: NTAP — CLARINO
  - de: Dortmund Chat Corpus — VLO

Overview of corpora (1/2)

Lang	Name of corpus	Data types	Size	Period	Anno	Avail	Found
German	Dortmund Chat Corpus	Chats	1,06m	/	T,P,L	/	VLO
German	DEREKO subcorpus	News & German Wiki	670m	/	T	D,C	P.C.
German, English	Monitor corpus of tweets from Austrian users	Tweets	30-40m	2007-2017	T, L	/	Google
German	DWDS subcorpus – Blogs	Blogs	102m	/	/	C	P.C.
Estonian	Mixed Corpus: New Media	Forums, chats, comments	25m	2000-2008	T	D, C	CLARIN Estonia
Finnish	Suomi 24	Forums	2,600m	2001-2016	T,P	C	VLO
Lithuanian	LITIS v.1	News	190k cmnts	2010-2014	/	D	VLO
Dutch	SoNaR New Media Corpus	Tweets, chats, SMS	35m	2005-2012	T,P,L	C	VLO
Dutch	Flemish online teenage talk	Facebook, Whatsapp	2.9m	2015-2016	T	/	Survey
Welsh	Corpus of Welsh Language Tweets	Tweets	7m tweets	/	/	D	P.C.
Norwegian, English, French	NTAP climate change blog corpora	Blogs related to climate change	5,000m	2000-2014	T	C	CLARINO
Polish	Corpus Highly Emotive	Tweets	160m	/	T	D	Google

Overview of corpora (2/2)

Lang	Name	Data types	Size	Period	Anno	Avail	Found
French	CoMeRe Repository	Emails, forums, chats, tweets, Wiki, etc.	75-80m	Various	/	D	P.C.
Italian	Web2Corpus_it	Forums, Blogs, Newsgroups, social networks, chats	/	/	T,P,L	/	P.C.
Slovenian	JANES	Slovene CMC	200m	2013-2016	T,P,L	/	CLARIN.SI

- Italian *Web2Corpus_it* and Slovene *JANES* are still in preparation
- *Monitor corpus of tweets from Austrian users* and *Flemish Online Teenage Talk* are unavailable.

Problems

- Missing metadata
  - Unknown temporal span for DWDS, DEREKO, Dortmund Chat, Corpus of Welsh Language Tweets
  - Unknown annotation process for DWDS, LITIS, Corpus of Welsh Language Tweets
- Licence info
  - Unclear for most of the surveyed corpora

Datasets

- 17 datasets identified
  - by language:
    - 9 different languages (cz, dk, el, de, it , es ,se, si, uk)
    - 1 multilingual
    - most for Slovene (6), English (3) and Italian (3)
  - by data type:
    - Tweets (10)
    - Facebook comments (2)
    - mixed (3)
    - blogs (1)
    - Reddit (1)
  - by task: sentiment analysis (5), NER (1), entity linking (1), rest miscellaneous
- 8 of these integrated in the CLARIN infrastructure

Tools

- **Within the CLARIN infrastructure:**
  - GATE tools (CLARIN-UK)
  - JANES tools (Clarin.si)
- **Elsewhere:**
  - Hunaccent (Hungarian)
    - Accentizer of Hungarian text
  - Twython (language-independent)
    - Python wrapper for the Twitter API
  - dmi-tcat
    - A set of tools to retrieve and collect tweets from Twitter for statistical analysis
  - Tweet NLP
    - A tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated corpora and web-based annotation tools.

Lessons learned

- User ambitions tend to be conservative, so ….
  a bit of technology push can be good, but ...
- ... the functionality that tools have to offer should support users in the workflows they know, rather than steer the exploration of data or the application of tools in ways that are not understood, so …
  user needs should be kept in focus.
- Scholarly insights and conlusions without modes for validating and/or replicating the results have difficulty to gain trust, so …
  black boxes have little added value
- For collaboration across disciplinary boundaries, communication pitfalls will never stop to exist, so …
  keep talking after this workshop!

CLARIN: Infrastructural support for the study and use of language as social and cultural data

darja.fiser@ff.uni-lj.si

CLARIN-PLUS workshop "Creation and Use of Social Media Resources"

Kaunas, Lithuania

18-19 May 2017

CLARIN