Title	Report on the CLARIN corpora in CLARIN Resource Families
Version	1.4
Author(s)	DF, JL
Date	13-06-2018
Status	Draft
Distribution	BoD, NCF, UI
ID	CE-2018-1236

Contents

1. Introduction

In the following report, we provide an in-depth summary of the current state of the *CLARIN Resource families* initiative. The aim of the CLARIN Resource Families initiative is to provide user-friendly overviews of the available corpora in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies. The overviews are based on extensive surveys of resources available in the VLO and the CLARIN repositories. The surveys were generally carried out in an four-step process: (i), we searched for the resources through the VLO; (ii), we checked the repositories and websites of the national consortia, (iii) we checked non-CLARIN repositories, such as META-SHARE and the LRE Map and (iv), contacted NC and UI representatives for their input. In this report, we summarize the information pertaining to corpora that are part of the CLARIN infrastructure – i.e., corpora that are either made available through national repositories or can be found through the VLO. We also provide lists of corpora for each type of resource that cannot be found in the VLO or have problematic metadata. Section 2 focuses on parliamentary corpora, section 3 on CMC corpora, section 4 on parallel corpora, section 5 on newspaper corpora, section 6 on L2-learner corpora and section 7 on historical corpora. Finally, in section 8 we provide a selection of major corpora that are not yet part of the CLARIN infrastructure for possible inclusion.

2. Parliamentary corpora

2.1 Summary

In Table 1, we summarize the information on parliamentary corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size and period, annotation and licence. The summary is based on the version of the overview presented on the Parliamentary corpora subpage of the *Resource families* that was current on 12 June 2018.

Table 1: Summary of information on parliamentary corpora within the CLARIN infrastructure

Identification	• 18 parliamentary corpora part of the CLARIN infrastructure in total • 13 (72%) corpora identified through the VLO • 5 (28%) corpora identified through national repositories, but not through VLO
Availability	• 5 (28%) corpora for download and through a concordancer • 9 (50%) corpora for download • 3 (24%) corpora through a concordancer • 1 (6%) corpora unavailable
Languages	• 17 (94%) corpora are monolingual, 1 (6%) is multilingual • 2 Norwegian corpora • 2 English corpora • 1 corpus per language: Czech, Danish, French, Estonian, Finnish, Greek, German, German (Austrian), Lithuanian, Portuguese, Slovenian, Swedish, Polish
Size and period	• Largest corpus 1.6 billion tokens, smallest corpus 190,000 tokens • Information on size available for all corpora • Period unknown for 1 (6%) corpus
Annotation	• 8 (44%) corpora PoS/MSD-tagged, 8 (44%) corpora lemmatized • Unknown for 4 (22%) corpora
Licence	• 8 (44%) corpora under CC-BY • Unknown for 4 (22%) corpora

2.2 Issues

In Table 2 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 2: Issues

List of parliamentary corpora with metadata issues	List of CLARIN parliamentary corpora not in the VLO
1. Czech Parliamentary Meetings • Unknown period 2. Hansard corpus • Unknown licence 3. Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) • Unclear linguistic annotation 4. Transcripts of Riigikogu (Estonian Parliament) • Unclear linguistic annotation 5. Plenary Sessions of the Parliament of Finland • Unclear linguistic annotation 6. Hellenic Parliament Sittings (2011-2015) • Unclear linguistic annotation 7. ParlAT beta • Unknown licence 8. Polish Parliamentary Corpus • Unknown licence 9. Europarl • Unknown licence	1. Hansard corpus 2. Hellenic Parliament Sittings (2011-2015) 3. Riksdag's Open data 4. Proceedings of Norwegian Parliamentary Debates 5. ParlAT beta

3. CMC corpora

3.1 Summary

In Table 3, we summarize the information on CMC corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size and period, annotation and licence. The summary is based on the version of the overview presented on the CMC corpora subpage of the *Resource families* that was current on 12 June 2018.

Table 3: Summary of information on CMC corpora within the CLARIN infrastructure

Identification	• 12 CMC corpora part of the CLARIN infrastructure in total • 11 (92%) corpora identified through the VLO • 1 (8%) corpus identified through a national repository, but not through VLO
Availability	• 7 (58%) corpora for download and through a concordancer • 4 (33%) corpora for download • 1 (8%) corpus through a concordancer
Languages	• All corpora are monolingual • 5 Slovene corpora • 1 corpus per language: Czech, Dutch, Estonian, Finnish, German, Lithuanian, French
Size and period	• Largest corpus 2.6 billion tokens, smallest corpus 1 million tokens • Information on size available for all corpora

	• Period unknown for 5 (42%) corpora
Annotation	• 9 (75%) corpora PoS/MSD-tagged, 7 (58%) corpora lemmatized • Annotation unknown for 3 (25%) corpora
Licence	• 9 (75%) corpora available under CC-BY • Unknown for 2 (8%) corpora

3.2 Issues

In Table 4 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 4: Issues

List of CMC corpora with metadata issues	List of CLARIN CMC corpora not in the VLO
1. Corpus of contemporary blogs • Unknown period • Unclear linguistic annotation 2. SoNaR New Media • Unknown licence 3. The Mixed Corpus: New Media • Unknown licence 4. LITIS v.1 • Unclear linguistic annotation 5. Wikipedia talk corpus Janes-Wiki 1.0 • Unknown period 6. Forum corpus Janes-Forum 1.0 • Unknown period 7. Blog post and comment corpus Janes-Blog 1.0 • Unknown period 8. News comment corpus Janes-News 1.0 • Unknown period 9. CoMeRe repository • Unclear linguistic annotation	1. The Mixed Corpus: New Media

4. Parallel corpora

4.1 Summary

In Table 5, we summarize the information on parallel corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, level of alignment[1] and licence. The summary is based on the version of the overview presented on the Parallel corpora subpage of the *Resource families* that was current on 12 June 2018.

Table 5: Summary of information on parallel corpora within the CLARIN infrastructure

Identification	• 79 parallel corpora part of the CLARIN infrastructure in total • 45 (54%) corpora can be found in the VLO • 34 (46%) corpora can be found only in a national repository, but not VLO

[1] Alignment refers to the level of textual alignment (e.g., word, sentence, paragraph, document).

Availability	• 4 (5%) corpora both for download and through a concordancer • 8 (10%) corpora through a concordancer • 58 (74%) corpora for download
Languages	• 45 (57%) corpora are bilingual, 34 (43%) corpora are multilingual • Most languages within one corpus – 117
Size	• Largest corpus 2.7 billion tokens, smallest 43,000 tokens • Unknown for 14 (18%) corpora
Alignment	• 37 (47%) corpora sentence-aligned • Unknown for 35 (44%) corpora
Licence	• 30 (38%) corpora available under CC-BY • Unknown for 8 (10%) corpora

4.2 Issues

In Table 6 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 6: Issues

List of parallel corpora with metadata issues	List of parallel corpora not in the VLO
1. The CLUVI parallel corpus • Unclear alignment 2. The English-Slovak Parallel corpus • Unknown size 3. The Polish-Lithuanian Parallel Corpus • Unknown size • Unclear alignment 4. Kacenka • Unknown licence 5. Parallel corpus newsletters IFT FR-GR • Unknown size • Unclear alignment 6. COMPARA • Unknown size 7. ACCURAT balanced test corpus for under resourced languages • Unclear alignment 8. EMEA Corpus • Unknown size 9. DGT-Acquis • Unknown size 10. Text Corpus - EMEL • Unclear alignment 11. FREL • Unclear alignment 12. Inlerlingual Perspectives • Unclear alignment • Typo in name 13. aformes • Unclear alignment 14. GLOSSOLOGIA • Unknown size	1. Parallel corpus newsletters IFT FR-GR 2. ACCURAT balanced test corpus for under resourced languages 3. European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English 4. EMEA Corpus 5. ECDC Translation Memory 6. DGT-Translation Memory 7. DGT-Acquis 8. EAC Translation Memory 9. A parallel corpus collected from the European Constitution 10. A parallel corpus of KDE4 localization files (v.2) 11. European Central Bank parallel corpus 12. OpenSubtitles2011 13. SPC - Stockholm Parallel Corpora 14. Tatoeba 15. DGT-TM-2016 16. QTLP English-Greek Corpus for the MEDICAL domain 17. QTLP German-Greek Corpus for the MEDICAL domain 18. QTLP Portuguese-Greek Corpus for the MEDICAL domain 19. QTLP English-Greek Corpus for the AUTOMOTIVE domain 20. QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain 21. Text Corpus - EMEL 22. FREL

- Unclear alignment

15. Civitas Gentium
    - Unclear alignment
16. Official Journal of the European Union
    - Unknown size
17. SzegedParalell: angol-magyar párhuzamos korpusz
    - Unknown size
    - Unclear alignment
    - Unknown licence
18. Tourism English-Croatian Parallel Corpus 2.0
    - Unclear alignment
19. LOGON parallel tourist corpus of Norwegian-English texts
    - Unclear alignment
    - Unknown licence
20. Serbian-English parallel corpus srenWaC 1.0
    - Unclear alignment
21. Parallel Bible Corpus
    - Unknown size
    - Unclear alignment
    - Unknown licence
22. JRC-Acquis Multilingual Parallel Corpus
    - Unknown licence
23. MLCC Multilingual and Parallel Corpora
    - Unclear alignment
24. MULCOLD - Multilingual Corpus of Legal Documents
    - Unclear alignment
25. PANACEA English-French and English-Greek parallel corpus
    - Unknown size
    - Unclear alignment
26. HindEnCorp 0.5
    - Unclear alignment
27. English-Luganda Parallel Corpus
    - Unknown licence
28. English-Urdu Religious Parallel Corpus
    - Unknown size
29. Polish-Bulgarian-Russian Parallel Corpus
    - Unknown size
    - Unclear alignment
    - Unknown licence
29. European Parliament Interpretation Corpus (EPIC)
    - Unclear alignment

Additionally, the level of alignment is unclear for the following corpora:

30. CsEnVi Pairwise Parallel Corpora
23. Inlerlingual Perspectives
24. aformes
25. GLOSSOLOGIA
26. Civitas Gentium
27. Official Journal of the European Union
28. INTERA Corpus - the Greek-English part
29. Greek-Bulgarian Bul-TM parallel corpus
30. Opus, Helsinki Korp Version
31. SzegedParalell: angol-magyar párhuzamos korpusz
32. LOGON parallel tourist corpus of Norwegian-English texts
33. PELCRA Polish-English parallel corpora
34. FTA/Eng and FTA/Spa

31. English-Czech Corpus from Wikipedia 32. DGT-Translation Memory 33. FTA/Eng and FTA/Spa 34. The Norwegian-Spanish Parallel Corpus 35. CRATER 2 Corpus	

5. Newspaper corpora

5.1 Summary

In Table 7, we summarize the information on newspaper corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the version of the overview presented on the Newspaper corpora subpage of the *Resource families* that was current on 12 June 2018.

Table 7: Summary of information on newspaper corpora within the CLARIN infrastructure

Identification	• 27 newspaper corpora part of the CLARIN infrastructure in total • 15 (56%) corpora can be found in the VLO • 12 (44%) corpora can be found only in a national repository, but not VLO
Availability	• 11 (41%) corpora both through a concordancer and for download • 5 (16%) corpora available through a concordancer • 10 (37%) corpora for download

Languages	• 23 (85%) corpora are monolingual
	  • 11 Swedish corpora
	  • 4 German corpora
	  • 2 Czech corpora
	  • 2 French corpora
	• 1 Arabic, 1 Finnish, 1 Norwegian, 1 Polish corpus
	• 4 (13%) corpora are multilingual

Size	• Largest 8.8 billion tokens, smallest corpus 0.5 million tokens • Unknown for 2 (7%) corpora
Annotation	• 14 (52%) corpora PoS/MSD-tagged; 4 (15%) corpora lemmatized • Unknown for 10 (37%) corpora
Licence	• 14 (52%) corpora available under CC-BY • Unknown for 4 (15%) corpora

5.2 Issues

In Table 8 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 8: Issues

List of newspaper corpora with metadata issues	List of CLARIN newspaper corpora not in the VLO

1. An-Nahar Newspaper Text Corpus
  • Unknown size
  • Unclear linguistic annotation
2. The Karelian Finnish Newspaper Corpus
	1. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
	2. DN 1987
	3. GP 1994 and 2001-2011

• Unclear linguistic annotation 3. BREF-80 • Unclear linguistic annotation 4. Corpus journalistique issu de l'Est Républicain • Unknown size • Unclear linguistic annotation 5. The Norwegian Newspaper Corpus • Unknown licence 6. MLCC Multilingual and Parallel Corpora • Unclear linguistic annotation 7. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version • Unclear linguistic annotation 8. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) • Unknown size • Unclear linguistic annotation 9. Corpora of Newspaper Texts • Unclear linguistic annotation 10. Mannheim Corpus of Historical Newspapers and Magazines • Unclear linguistic annotation • Unknown licence	4. Kubhist 5. The Webbnyheter corpus 6. Dagny 7. Hertha 8. Idun 9. Kvinnorans Tidning 10. Rösträtt för Kvinnor 11. Morgonbris 12. Smittskydd

6. L2-learner corpora

6.1 Summary

In Table 9, we summarize the information on L2-learner corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the version of the overview presented on the L2 learner corpora subpage of the *Resource families* that was current on 12 June 2018.

Table 9: Summary of information on L2-learner corpora within the CLARIN infrastructure

Identification	• 34 L2-learner corpora part of the CLARIN infrastructure in total • 28 (82%) corpora identified through the VLO • 6 (18%) corpora identified through a national repository, but not through VLO
Availability	• 3 (9%) corpora for download and through a concordancer • 4 (12%) corpora through a concordancer • 16 (47%) corpora for download
Languages	• 24 (71%) monolingual corpora, 10 (29%) multilingual • Monolingual corpora for the following 9 languages: English (10), Finnish (4), Swedish (3), German (2), Arabic (1), Czech (1), French (1), Hungarian (1), Norwegian (1) • Multilingual corpora for the following 5 combinations: English and French (2); Finnish and English (2); English and German (1); Finnish, English and Swedish (1);

	French, Italian, Spanish (1). An additional 3 corpora with more than 5 languages each.
Size	• Largest corpus 3 million words • Unknown for 10 (29%) corpora
Annotation	• Unknown for 24 (70%) corpora
Licence	• 12 (35%) corpora available under CLARIN RES, 10 (29%) under CC-BY • Unknown for 7 (21%) corpora

6.2 Issues

In Table 10 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 10: Issues

List of L2 corpora with metadata issues	List of CLARIN L2-corpora corpora not in the VLO
1. British Academic Written English Corpus • Unclear linguistic annotation 2. ICLE International Corpus of Learner English • Unclear linguistic annotation • Unknown licence 3. The Uppsala Student English corpus • Unclear linguistic annotation 4. The Advanced Finnish Learners' Corpus • Unclear linguistic annotation 5. Testipiste Corpus • Unclear linguistic annotation 6. Commented Learner Corpus Academic Writing • Unclear linguistic annotation 7. FinSveStud 79-80 • Unclear linguistic annotation 8. The Anglish Corpus • Unknown size • Unknown licence 9. GLBCC (Giessen - Long Beach Chaplin Corpus) • Unclear linguistic annotation 10. A Learners' Corpus of Reading Texts • Unknown size • Unclear linguistic annotation 11. LANGMAN • Unknown size • Unclear linguistic annotation 12. Arabic Learner Corpus • Unclear linguistic annotation 13. English as a Foreign Language Corpus • Unclear linguistic annotation • Unknown licence	25. The Hanken Corpus of Academic Writing 26. Testipiste Corpus 27. ASK – Norsk andrespråkskorpus 28. SW1203-essays 29. DIALUKI: Diagnosing reading and writing in a second or foreign language 30. TAITO: Written and Oral Data of the TAITO-project

14. The Long Second Corpus • Unknown size • Unclear linguistic annotation 15. LETEC (Learning and Teaching Corpus) • Unknown size • Unclear linguistic annotation 16. CEFLING Project Corpus • Unknown size • Unclear linguistic annotation • Unknown licence 17. DIALUKI: Diagnosing reading and writing in a second or foreign language • Unclear linguistic annotation 18. Topling - Paths in Second Language Acquisition • Unclear linguistic annotation 19. AixOx • Unclear linguistic annotation 20. LeaP: The Learning the Prosody of a Foreign Language • Unknown licence 21. Repiso/Contrefactualité • Unknown size • Unclear linguistic annotation 22. Openprodat • Unknown size • Unclear linguistic annotation 23. TAITO: Written and Oral Data of the TAITO-project • Unknown size • Unclear linguistic annotation • Unknown licence 24. YKI National Certificates corpus • Unknown size • Unclear linguistic annotation	

7. Historical corpora

7.1 Summary

In Table 11, we summarize the information on historical corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size and period, annotation and licence. The historical corpora subpage is forthcoming as of 12 June 2018.

Table 11: Summary of information on historical corpora within the CLARIN infrastructure

Identification	• 68 historical corpora part of the CLARIN infrastructure in total • 56 (82%) corpora identified through the VLO • 12 (18%) corpus identified through a national repository, but not through VLO
Availability	• 9 (13%) corpora for download and through a concordancer • 22 (32%) corpus through a concordancer • 51(32%) corpora for download
Languages	• 51 (75%) monolingual corpora, 17 (25%) multilingual • Monolingual corpora for the following 26 languages: English (15), German (12), Finnish (6), French (2), Dutch (2) Greek (2), Slovenian (2), Scots (2), Akkadian (1), Chinese (1), Estonian (1), Hungarian (1), Old Norse (1), Polish (1), Swedish (1), Welsh (1) • Multilingual corpora for the following 11 combinations: Finnish, Swedish (6); English, Latin (2); Czech, German, Latin, Polish (1); Czech, Latin (1); English, Hebrew (1); English, German, Latin, Old Norse, Swedish (1); English, Sumerian (1); Finnish, Karelian, Ludian, Latin, Swedish, Olonets, Izhorian, Votic (1); Finnish, Russian, German, Latin (1); Latin, German (1); Swedish, German, French, etc. (1)
Size and period	• Largest corpus 8.7 billion tokens, smallest corpus 11, 300 tokens • Unknown for 15 (22%) corpora • Oldest corpus contains Sumerian texts
Annotation	• 18 (26%) corpora are PoS/MSD-tagged, 13 (19%) corpora are lemmatized and 6 (9%) corpora are syntactically parsed • Unknown for 21 (31%) corpora
Licence	• 35 (51%) corpora are available under CC-BY, 11 (16%) corpora under the Oxford Text Archive licence, 6 (9%) under CLARIN ACA, 3 (4%) under EUPL v.1.1 SA • Unknown for 8 (12%) corpora

7.2 Issues

In Table 12 we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 12: Issues

List of historical corpora with metadata issues	List of CLARIN historical corpora not in the VLO
1. Hungarian Historical Corpus • Unclear linguistic annotation • Unknown licence 2. Sheffield Corpus of Chinese • Unknown size	1. Menota 2. Greek Medieval Texts 3. Austrian Baroque Corpus 4. OROSSIMO Corpus - History 5. DDR-Presseportal (GDR press portal)

- Unclear linguistic annotation

3. Historical Corpus of the Welsh Language 1500-1850
   - Unclear linguistic annotation
   - Unknown licence
4. GerManC. A Historical Corpus of German Newspapers 1650-1800
   - Unclear linguistic annotation
5. "PolDiLemma" Middle Polish Diachrone Lemmatized Corpus
   - Unknown size
6. Deutsches Textarchiv (DTA)
   - Unknown size
   - Unclear linguistic annotation
7. Corpus of biblical text in Scots / John Kirk,
   - Unknown size
   - Unclear linguistic annotation
8. Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn
   - Unknown size
   - Unclear linguistic annotation
9. Corpus of Late Modern English prose / David Denison
   - Unknown size
   - Unclear linguistic annotation
10. The Helsinki corpus of Older Scots : [1450-1700]
    - Unknown size
11. The Lampeter Corpus of Early Modern English Tracts
    - Unknown size
    - Unclear linguistic annotation
12. Dictionary of Old English Corpus in Electronic Form (DOEC)
    - Unknown size
    - Unclear linguistic annotation
13. Greek Medieval Texts
    - Unclear linguistic annotation
14. Mannheimer Korpus Historischer Zeitungen und Zeitschriften
    - Unclear linguistic annotation
    - Unknown licence
15. Austrian Baroque Corpus
    - Unknown licence
16. Hansard Corpus
    - Unknown licence
17. Die Grenzboten (journal)
    - Unknown licence

6. The Morpho-Syntactic Database of Mikael Agricola's Works
7. Classics Library of the National Library of Finland - Kielipankki version
8. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
9. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1875-1920)
10. The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
11. The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
12. Språkbanken's historical corpora

18. Brieven als buit (Letters as loot)
    - Unknown licence
19. Partonopeus de Blois: transcriptions of all manuscripts and fragments
    - Unknown size
    - Unclear linguistic annotation
20. The Electronic Text Corpus of Sumerian Literature. Revised edition.
    - Unknown size
21. The Lancaster Newsbooks Corpus
    - Unknown size
    - Unclear linguistic annotation
22. A Corpus of English Dialogues 1560-1760 (CED)
    - Unclear linguistic annotation
23. Aleksis Kivi Corpus (SKS)
    - Unclear linguistic annotation
24. Finnish Folk Poetry
    - Unclear linguistic annotation
25. Classics of Finnish Literature, Kielipankki Version
    - Unclear linguistic annotation
26. Corpus of Old Literary Finnish
    - Unclear linguistic annotation
27. Corpus of Early Modern Finnish, Kielipankki Version
    - Unclear linguistic annotation
28. The Letters of Paul Sinebrychoff, Kielipankki Version
    - Unclear linguistic annotation
29. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
    - Unclear linguistic annotation
30. Classics Library of the National Library of Finland - Kielipankki version
    - Unknown size
    - Unclear linguistic annotation
31. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
    - Unknown size
    - Unclear linguistic annotation
32. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
    - Unknown size
    - Unclear linguistic annotation

8. List of corpora not yet in the CLARIN infrastructure

Table 9 provides a selection of the most relevant corpora we have identified in the METAShare repository, the LRE Map and via Google which are still missing from the CLARIN infrastructure for the four all the types of language resources. Contact information is provided for the relevant authors so that they can be invited to deposit their corpus.

Table 9: List of corpora that are not yet included in the CLARIN infrastructure

Corpus name	Language	Contact information
Parliamentary corpora		
DutchParl	Dutch	Maarten Marx M.J.Marx@uva.nl
Korpusbasierte Analyse österreichischer Parlamentsreden	German (Austrian)	Colin Sippl colin.sippl@stud.uni-regensburg.de
SAEIMA	Latvian	Ilze Auziņa ilzea@latnet.lv
CMC corpora		
Corpus of Highly Emotive Internet Discussions	Polish	Antoni Sobkowicz antoni.sobkowicz@opi.org.pl
The Corpus of Welsh Language Tweets	Welsh	D.B. Jones d.b.jones@bangor.ac.uk
Parallel corpora		
United Nations Parallel Corpus	Arabic, English, Spanish, French, Russian, Chinese	Michał Ziemski mziemski@unog.cz
MultiUN: Multilingual UN Parallel Text 2000—2009	English, French, Spanish, Arabic, Russian, Chinese, German	Andreas Eisele andreas.eisele@dfki.de
µtopia	English, Mandarin, Russian, Korean, Japanese	Wang Ling lingwang@cs.cmu.edu
LILA parallel corpus	Lithuanian and Latvian	Andrius Utka a.utka@hmf.vdu.lt
TED-Parallel-Corpus	Multilingual (109 languages)	Ajinkya kulkarni ajinkyakulkarni14@gmail.com
EUbookshop	Multilingual (48 languages)	Raivis Skadiņš raivis.skadins@tilde.lv
Newspaper corpora		
Europeana Newspapers NER Corpora	Dutch, French, German	Clemens Neudecker clemens.neudecker@europeana-newspapers.eu
Zurich English Newspaper Corpus	English (historical)	Udo Fries ufries@es.uzh.ch
"LA REPUBBLICA" Corpus	Italian	Marco Baroni marco.baroni@unitn.it
Timestamped JSI web corpus	Multilingual	Mitja Trampus mitja.trampus@ijs.si
L2 corpora		

EIC: The Estonian Interlanguage Corpus of Tallinn University	Estonian	Pille Eslon peslon@tlu.ee
SPLLOC: Spanish Learner Language Oral Corpus	Spanish	Professor Rosamond Mitchell R.F.Mitchell@soton.ac.uk
Historical corpora		
Historical Corpora at Lancaster University	English	Andrew Hardie a.hardie@lancaster.ac.uk
M.I.DIA. (Morfologia dell'Italiano in DIAcronia)	Italian	Aurelio de Rosa a.derosa@audero.it
Old Hungarian Corpus	Hungarian	mgtsz@nytud.mta.hu
Parsed Corpus of Historical Portuguese	Portuguese	corpustb@gmail.com

9. General comments

Author: Koenraad de Smidt

Sometimes a resource is available at multiple sites but sometimes with different metadata, and sometimes pointing to different versions, sometimes with obsolete links. This is a known problem with highly distributed infrastructures, which is why we should try to devise CLARIN-wide solutions, e.g. by using the concept of “authoritative metadata”, by mirroring, by more clearly requesting that sites that copy should not link to urls, but to PIDS, etc.