CLARIN's Key Resource Families

Darja Fišer[1,2], Jakob Lenardič[1], Tomaž Erjavec[2]
Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{darja.fiser, jakob.lenardic}@ff.uni-lj.si, tomaz.erjavec@ijs.si

Abstract

CLARIN is a European Research Infrastructure that has been established to support the accessibility of language resources and technologies to researchers from the Digital Humanities and Social Sciences. This paper presents CLARIN's Key Resource Families, a new initiative within the infrastructure, the goal of which is to collect and present in a uniform way the most prominent data types in the network of CLARIN consortia that display a high degree of maturity, are available for most EU languages, are a rich source of social and cultural data, and as such are highly relevant for research from a wide range of disciplines and methodological approaches in the Digital Humanities and Social Sciences as well as for cross-disciplinary and trans-national comparative research. The four resource families that we present each in turn are newspaper, parliamentary, CMC (computer-mediated communication), and parallel corpora. We focus on their presentation within the infrastructure, their metadata in terms of size, temporal coverage, annotation, accessibility and license, and discuss current problems.

Keywords: language resources, research infrastructure, open science, digital humanities and social sciences, CLARIN

1. Introduction

CLARIN is a European Research Infrastructure that has been established to support the accessibility of language resources and technologies to researchers from the Humanities and Social Sciences (Krauwer and Hinrichs, 2014). CLARIN's vision, mission and design are aimed at findability, accessibility, interoperability and re-usability of its resources, tools and services to support researchers in the Humanities and Social Sciences (SSH) (de Jong et al., 2018; De Smedt et al., 2018). At the time of writing, CLARIN has 20 member and 2 observer countries which provide numerous language resources and tools through certified data centres. Access to these resources is enhanced by the Virtual Language Observatory (VLO) portal which enables searching for resources and provides a uniform display of highly-granular Component MetaData Infrastructure (CMDI) metadata (Van Uytvanck et al., 2012).

Similar to other service-oriented e-research infrastructures (Chunpir et al., 2015), the developmental phase of the CLARIN infrastructure and services was followed by analyses of user experience and needs. Odijk (2014) looked into the problems with resource descriptions, granularity of metadata and resources, and string as well as faceted search in the VLO through the eyes of a linguist. Lušicky and Wissik (2016) conducted a similar survey of resource discovery that was tailored to the needs in translation studies. Sanders (2017) carried out two focus groups in order to obtain user experience and desiderata from two distinct types of users: humanities and social sciences researchers, and language technology and information technology experts. All these surveys reached similar conclusions: both SSH researchers as well as IT experts need easier access to the desired resources and more data type-specific guidance, including comprehensive metadata on the provenance and annotation of the resources, standard formatting, uniform concordancing and text analytics options that enable comparisons across corpora, as well as showcases and best practice examples from different disciplines that shows the value and utilization of the CLARIN infrastructure to fellow researchers in a real-life setting.

Inspired by these findings, the aim of this paper is to present current CLARIN's Key Resource Families, a new initiative within the infrastructure, the goal of which is to collect and present in a uniform way prominent data types in the network of CLARIN consortia that display a high degree of maturity, are available for most EU languages, are a rich source of social and cultural data, and as such highly relevant for research from a wide range of disciplines and methodological approaches in SSH as well as for cross-disciplinary and trans-national comparative research.

2. Survey protocol

For all data types except parallel corpora (see Section 3.4) we limited our search to the 22 countries that are either members or observers of CLARIN ERIC: Austria, Bulgaria, the Czech Republic, Denmark, Belgium, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, the Netherlands, Norway, Poland, Portugal, France, Slovenia, and Sweden, and the UK.

Corpora were identified through the following procedure: (1) searching the VLO, (2) national CLARIN repositories and websites, (3) the META-SHARE[1] repository and LRE Map,[2] (4) Google Scholar[3], and (5) reaching out to the representatives of the national CLARIN consortia.

The primary aim of the surveys was to ascertain three general characteristics of the corpora. First, we checked whether the corpora are available through the VLO. Since the VLO is a CMDI-facet browser, which is an easy-to-use interface that ensures uniform access to resources from all the national CLARIN repositories (van Uytvanck et al., 2012), it is worthwhile to know whether there exist corpora that still lack VLO entries. In our surveys, we listed such corpora for future inclusion in the VLO. Second, we provided information on the availability of the corpora – that is, whether a corpus can be downloaded,

[1] http://metashare.csc.fi
[2] http://lremap.elra.info
[3] https://scholar.google.com

accessed through a concordancer or both, which is essential for re-use of the corpora. Finally, we described the presentation of their metadata in terms of (token) size, covered time period, levels of linguistic annotation and licence, highlighting missing information.

3. Survey results

3.1 Newspaper corpora

Due to large volumes, good document structure and rich and reliable metadata, one of the most traditional sources for corpus compilation projects are newspapers. They are an invaluable source of data for synchronic as well as diachronic studies of neologisms and other lexicographic phenomena (Smørdal Losnegaard and Inger Lyse, 2012), gender studies (Baker, 2012) and politology (Baker and McEnery, 2005).

We have identified 40 newspaper corpora, 30 of which are part of the CLARIN infrastructure. Due to space restrictions, we focus only on presenting their salient characteristics in terms of identification (i.e., whether the CLARIN corpora are also listed in the VLO), availability and metadata in this paper but provide a comprehensive overview in the report published on the CLARIN website.[4]

In relation to the metadata, information on size is available for the majority of the corpora (24 out of 30). *The Newspaper and Periodical Corpus of the National Library of Finland* (FIN-CLARIN, 2013)[5] is the largest, containing 8.6 billion tokens and covers the longest time period from 1771 to 2011. The smallest is the German *Tiger Corpus* (CLARIN-D, 2003), which contains 900,000 tokens. Information regarding annotation is available for 21 of the 30 corpora in the CLARIN infrastructure, almost half of which are tagged and annotated for syntactic dependencies, as is the case of the *Press 65-98* corpus (Språkbanken, 2017a). In terms of availability. 5 of the corpora can be downloaded (e.g. the *Tiger Corpus*), 7 can be accessed through concordancers (e.g. *The Newspaper and Periodical Corpus of the National Library of Finland* corpus) and 10 are both for download and accessible through concordancers (e.g. the *Press 65-98* corpus). It is worth noting that all of the 10 corpora available both for download and online querying are accessible through the same concordancer *Korp*[6], which is provided by the Swedish CLARIN repository. Information on license is available for 21 corpora, 17 of which are available under public CC-BY (Creative Commons) licences, 4 under the academic ELRA END USER licence, 1 under CLARIN PUB (i.e., publicly available) and 1 under the restrictive CLARIN RES licence. Only 16 of the 30 corpora that are part of the CLARIN infrastructure have VLO entries.

[4]https://office.clarin.eu/v/CE-2017-1128-Newspaper-corpora.pdf.

[5] All referenced corpora are accompanied by their URL in the Sec. 8.

[6]https://www.kielipankki.fi/tuki/korp/. Finnish corpora are available through the *Kielipankki* (Language Bank of Finland) variant of Korp : https://korp.csc.fi/#?stats_reduce=word.

3.2 Parliamentary corpora

The second family of resources are corpora of parliamentary proceedings, which are a quintessential resource for a wide range of research questions from a number of SSH disciplines, such as Critical Discourse Analysis (Voutilainen, 2017), History (Pančur and Šorn, 2016) as well as Sociolinguistics (Rheault et al., 2015). Their most distinguishing characteristic is that they are transcriptions of spoken language produced in controlled and regulated circumstances. For this reason, they are rich in invaluable (sociodemographic) metadata as well as easily available under the Freedom of Information Acts set in place to enable informed participation by the public and to improve effective functioning of democratic systems, making the datasets even more valuable.

In total, we identified 22 corpora of parliamentary records. They exist for all CLARIN countries except Italy. We found one corpus per CLARIN country except in the case of the Czech Republic and Norway, where we found two corpora for each. Out of the 22 existing parliamentary corpora (the full list of which is available on the CLARIN webpage[7]), the following 16 are available within the CLARIN infrastructure:

(1) *Czech Parliament Meetings* (Pražák and Šmídl, 2012);
(2) *DK-CLARIN Almensprogligt korpus* (CLARIN-DK, 2012);
(3) *Transcripts of Riigikogu* (Center for Estonian Language Resources, 2017);
(4) *Eduskunta Corpus* (Parliament of Finland, 2017);
(5) *Hellenic Parliament Sittings* (clarin:el, 2017);
(6) *Proceedings of Norwegian Parliamentary Debates* (Språkbanken, 2016);
(7) *Riksdag's Open Data* (Språkbanken, 2017b);
(8) *PTPARL Corpus* (ELRA, 2017a);
(9) *SlovParl* (Pančur et al. 2016);
(10) *Hungarian National Corpus* (Váradi, 2005);
(11) *Hansard Corpus* (Hansard-corpus.org, 2017);
(12) *Lithuanian Parliament Corpus for Authorship Attribution* (Kapočiūtė-Dzikienė et al., 2017);
(13) *Talk of Norway* (Lapponi and Søyland, 2016);
(14) *Parliamentary Debates on Europe at the House of Commons (1998-2015)* (Truan, 2016a);
(15) *Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)* (Truan, 2016b);
(16) *Parliamentary Debates on Europe at the Bundestag (1998-2015)* (Truan, 2016c).

The *Hansard Corpus* and the *Riksdag's Open Data* are the largest, comprising well over 1 billion tokens. Other corpora are significantly smaller (most between 10 and 100 million tokens) with the *PTPARL Corpus* (1 million tokens) and the *Czech Parliament Meetings* corpus (0.5 million tokens) being the smallest. All corpora cover various contemporary periods from the 1970s onwards except for the *Hansard Corpus*, which contains parliamentary sessions from the period between 1803 and 2005. In relation to linguistic annotation, all of the above corpora are tokenized, lemmatized and tagged for parts of speech except for the *Riksdag's Open Data* corpus, which is tokenized, lemmatized, MSD-tagged, and displays syntactic dependency relations, and the *Proceedings of*

[7]https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report.pdf.

Norwegian Parliamentary Debates corpus, which is only tokenized. In terms of availability, 10 corpora – that is, (1), (2), (5), (8), (9), and (12)-(16) – are available for download, corpora (4), (6), (10), (11) are available through online search environments and corpora (3) and (7) are available both for download and through online search environments. Corpora (4) and (7) can be accessed through through the Swedish *Korp*[8] concordancer, while corpus (6) can be queried through *Corpuscle*[9], the concordancer of the Norwegian consortium CLARINO. The rest (corpora 3, 6, 10, 11) are available through different dedicated non-CLARIN online environments. Corpora (1), (4), (5), (7), (9) and (8)-(16) are available under the CC-BY license; the license is unknown for corpora (10)-(12). All the corpora can be found through the VLO except for (5), which can be found in the repository of the Greek consortium; (7), which can be found in the repository of the Swedish consortium; and (11), which can be found on the website of the British observer.

3.3 Computer-mediated communication corpora

The third data type included in the initiative are corpora of computer-mediated communication (CMC) that are compiled from user-generated content (such as blogs, forums, and chats) as well as from interactions on social media (such as Twitter, Facebook, and WhatsApp). CMC corpora are a rich data type which can be collected in real time and for which a multitude of metadata can be harvested automatically. Such corpora therefore have a big potential for reuse and re-purposing in many fields of SSH. CMC corpora can serve as a basis for researching contemporary language variation and change (Danescu-Niculescu-Mizil et al., 2013) as well as changes in social and cultural dynamics (Östman and Turtiainen, 2016). Unlike most traditional text types, these corpora contain high levels of "noise" due to non-standard language phenomena that are frequent in informal on-line communication settings. Compilation and further dissemination of such corpora is hindered by terms of use and privacy protection limitations.

In total, we identified 21 CMC corpora covering 14 languages, 2 of which were multilingual. The full survey is again available through the CLARIN webpage.[10] 12 of the 20 identified corpora are part of the CLARIN infrastructure:

(1) the Estonian *Mixed Corpus: New Media* (Segakorpus 2011);
(2) the Finnish *Suomi24* corpus (Aller Media ltd., 2014);
(3) the Lithuanian *LITIS v.1. corpus* (Amilevičius and Petkevičius, 2016);
(4) the Dutch *SoNaR New Media Corpus* (Radboud University et al., 2013);
(5) the German *Dortmund Chat Corpus* (Technische Universität Dortmund, 2013);
(6) the Czech *Corpus of Contemporary Blogs* (Grác, 2011);
(7) the Slovene *Twitter corpus Janes-Tweet 1.0* (Ljubešić et al., 2017a);
(8) the Slovene *Wikipedia talk corpus Janes-Wiki 1.0* (Ljubešić et al., 2017b);
(9) the Slovene *Forum corpus Janes-Forum 1.0* (Erjavec et al., 2017a);
(10) the Slovene *Blog post and comment corpus Janes-Blog 1.0* (Erjavec et al., 2017b);
(11) the Slovene *News comment corpus Janes-News 1.0* (Erjavec et al., 2017c) and;
(12) the French *CoMeRe Repository* (Chanier et al., 2014).

The *Suomi24* corpus is the largest, containing 2.6 billion tokens while the German *Dortmund Chat Corpus* the smallest with 1 million tokens. Information on the time span is available only for corpora (1), (2), (3), (4) and (7) – among these, corpus (2) has the longest span, covering the period between 2001 and 2016, whereas corpus (7) contains data from the shortest period between 2013 and 2017. Corpus (1) is tokenized, corpus (2) is tokenized and morphosyntactically tagged, (4) and (5) are tokenized and part of speech tagged and corpora (7)-(11) are tokenized, word-level normalised (standardised), morphosyntactically tagged and lemmatized, while the annotation levels are unknown for corpora (3) and (6). Corpora (3) and (5) are available for download, corpora (4) and (6) for online searching and corpora (1), (2), and (7)-(12) both for download and online searching. Here corpus (2) is available through *Korp* and (7)-(12) through KonText and noSketch Engine, all CLARIN-provided concordancers. While the license is unknown for corpora (1) and (4), the rest are available under various CC licences, save for corpus (3), which is available under ACA_CLARIN-LT_End-User-License-Agreement_EN-LT – that is, under an academically-restricted licence. All of the corpora can be found on the VLO except for (1), which is available only on the website of the Estonian consortium.

3.4 Parallel corpora

The most recent family of resources from our survey are parallel corpora. Unlike the rest of the surveyed resource families, this overview had a broader scope not limited to official languages of CLARIN member states because other languages are also relevant for many researchers in the CLARIN network. In total, we were able to identify 106 parallel corpora, 81 of which are already part of the CLARIN infrastructure. Due to space restrictions, we provide here only a summary of the identified 81 corpora in the CLARIN infrastructure, and give the full account in the report available on the CLARIN webpage.[11]

The largest corpus is *Opus – the Helsinki Korp Version* (Tiedemann, 2004), which contains 2.7 billion tokens, whereas the smallest is *Text Corpus – EMEL* (clarin:el, 2016), which contains 43,000 tokens. 32 of the 81 corpora are multilingual, with the *Parallel Bible Corpus* (Christodouloupoulos and Steedman, 2014) and the *Tatoeba* corpus (Tiedemann, 2012a) containing data from more than 100 languages. Information on annotation,

[8]https://www.kielipankki.fi/tuki/korp/ and https://korp.csc.fi/#?stats_reduce=word.
[9] http://clarino.uib.no/korpuskel/page.
[10]https://office.clarin.eu/v/CE-2017-1064-Resources-for-computer-mediated-communication.pdf.
[11]https://office.clarin.eu/v/CE-2017-1095-Parallel-corpora-report.pdf.

which primarily pertains to the level of alignment in the case of parallel corpora, is available for 43 out of 81 corpora, 39 of these are sentence-aligned whereas the *Czech-English Manual Word Alignment* corpus (Mareček, 2008) and *GeFRePaC - German French Reciprocal Parallel Corpus* (ELRA, 2018) are also word aligned, *ParRus* (Bartis, 2017) is only paragraph aligned, *Czech and English abstracts of ÚFAL papers* (Rosa, 2016) is only document-aligned. In terms of availability, 45 corpora are available for download (e.g. the *Czech-English Manual Word Alignment* corpus); 23 corpora, such as *European Parliament Interpretation Corpus* (ELRA, 2017b), are only listed in the national repositories or the VLO but are unavailable (often due to various license restrictions, sometimes due to broken links); 10 corpora are available through concordancers (e.g. *Parallel Bible Corpus*) and 3 corpora - (the *OPUS* corpus (Tiedemann, 2012b), the *KOTUS Finnish-Swedish Parallel Corpus* (Institute for the Languages of Finland, 2014) and *The Norwegian-Spanish Parallel Corpus* (Hareide, 2013) - are available both for download and through concordancers. Licensing information is available for 68 corpora – 32 of these are available under CC-BY. 46 corpora can be found through the VLO while the rest are listed on the national repositories only (e.g., *Tatoeba* can only be found on the repository of the Greek consortium).

4. Discussion

We observed very uneven levels of inclusion into the CLARIN infrastructure across the types of resources that we surveyed. While many corpora have been added to national repositories, they still cannot be identified through VLO directly due to lacking, idiosyncratic or vernacular names, keywords or description fields, such as the Portuguese *PTPARL corpus* and the Danish *DK-CLARIN Almensprogligt korpus*. For some of the corpora, only older versions can be found through VLO, even though more recent ones are available on national repositories (e.g. *Hungarian National Corpus*). Granularity of the deposited resources ranges from complete archives to single-file records (e.g. 148 records of *Flemish Parliament Debates)*, which makes navigation and use of the resources much more difficult. The most frustrating of all the accessibility issues, however, are cases where successfully identified records lead to empty or broken download links, such as the parallel corpus *The Croatian-Slovenian Parallel Corpus* (Tadić, 2014).

The second group of major issues is the incomplete documentation (metadata) for many of the corpora that can range from corpus size (e.g. *Parallel Bible Corpus*), period (e.g. the German *Dortmund Chat* corpus), linguistic annotation (e.g. Finnish *Eduskunta corpus*), or license information (e.g. *SoNaR New Media Corpus*). For parallel corpora, the biggest issue in terms of metadata is the directionality of translations which is not available for most of the corpora (e.g. on the *MUSA Multilingual Multimodal corpus* (Piperidis et al., 2004)).

Of all four resource types discussed in this paper, the metadata on the parliament corpora are generally the best, since the only information that was lacking was related to annotation in the case of 2 out of the 11 corpora within the infrastructure. By contrast, the metadata on the parallel corpora are much poorer. Information regarding corpus size fares best as it is available for 67 (83%) of the corpora. Information regarding annotation, which primarily has to do with the level of textual alignment, is available for 43 (52%) corpora.

5. Conclusion

In this paper we provided a comprehensive overview of four key resource families across the CLARIN network: newspaper corpora, parliamentary corpora, corpora of computer-mediated communication, and parallel corpora. In addition to generating the first entry point for these rich and versatile resource families of its kind, and important secondary aim is to establish how their integration to the infrastructure could be further improved and enriched. In our future work we plan to extend CLARIN's Key Resource Families with other major datatypes, such as historical and learner corpora, as well as increase their visibility by developing a rich online research environment with tutorials and exercise kits for them as well as offer updates on their recent enhancements.

6. Acknowledgements

The work reported in this paper has been supported by the member countries and observers in the CLARIN ERIC, and it has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 676529 for project CLARIN-PLUS. We would like to thank all the national User Involvement Coordinators and researchers who have provided invaluable feedback on our surveys.

7. Bibliographical References

Baker, P. (2012). Corpora and Gender Studies. In Hyland, K., Huat, C.H., & Handford, M. (eds.): *Corpus Applications in Applied Linguistics*, pp. 100-116.

Baker, P. & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. In *Journal of Language and Politics* 4 (2), pp. 197-226.

Chunpir, H. I., Ludwig, T., & Williams, D. N. (2015). Evolution of E-Research: From Infrastructure Development to Service Orientation. Proceedings of the Fourth International Conference on Design, User Experience, and Usability: Interactive Experience Design (DUX'15), pages 25–35. Cham, Springer. http://dx.doi.org/10.1007/978-3-319-20889-3.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J. & Potts, C. (2013). No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 307-318.

De Jong, F., Maegaard, B., De Smedt, K., Fišer, D., & Van Uytvanck, D. (2018). CLARIN: Towards FAIR and Responsible Data Science. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18).

De Smedt, K., de Jong, F., Maegaard, B., Fišer, D., & Van Uytvanck, D. (2018) Towards an Open Science Infrastructure for the Digital Humanities: The Case of CLARIN. In Proceedings of the third international conference Digital Humanities in the Nordic Countries (DHN2018).

Krauwer, S. & Hinrichs, E. (2014). The CLARIN research infrastructure: resources and tools for e-Humanities scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1525-1531), Portorož, Slovenia. European Language Resources Association (ELRA).

Lušicky, V., & Wissik, T. (2016). Evaluation of CLARIN services, User requirements, Usability, VLO, Translation Studies. Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, France, pages 63-75. Linköping University Electronic Press, Linköpings universitet.

Odijk, J. (2014). Discovering Resources in CLARIN: Problems and Suggestions for Solutions. Utrecht University Repository, Netherlands.

Östman, S. & Turtiainen, R. (2016). From Research Ethics to Researching Ethics in an Online Specific Context. Media and Communication, 4 (4), pp. 66-74.

Pančur, P. & Šorn, M. (2016). Smart Big Data: use of Slovenian parliamentary papers in digital history. *Prispevki za novejšo zgodovino*, 56 (3), pages 130-146.

Rheault, L., Beelen, K., Cochrane, C. & Hirst, G. (2015). Measuring Emotion in Parliamentary Debates Using Methods of Natural Language Processing. http://ww.cs.toronto.edu/pub/gh/Rheault-etal-CPSA-2015.pdf .

Sanders, W. (2017). Focus Group on User Involvement. Sofia, Bulgaria. CLARIN. https://office.clarin.eu/v/CE-2017-1091-Focus-group-UI-2017-03-27.pdf

Smørdal Losnegaard, G. & Inger Lyse, G. A data-driven approach to Anglicism identification in Norwegian. In Andersen, G. (ed.): *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*.

Sobkowicz, A. (2016). Political Discourse in Polish Internet – Corpus of Highly Emotive Internet Discussions. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Sobkowicz_Political-Discourse-in-Polish-Internet.pdf.

Tiedemann, J. (2012a). Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages, 2214-2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 1029-1034, Istanbul, Turkey. European Language Resources Association (ELRA).

Voutilainen, E. (2017). Parliamentary Records as Data for Linguistic Discourse Studies. http://videolectures.net/clarinplusworkshop2017_voutilainen_studies/.

8. Language Resource References

Aller Media ltd. (2014). The Suomi 24 Sentences Corpus. Distributed via Kielipankki, http://urn.fi/urn:nbn:fi:lb-2017021505.

Amilevičius, D. & Petkevičius, M. (2016). LITIS v.1, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/11.

Bartis, I. (2017). ParFin 2016, ParRus 2016, Finnish-Russian / Russian-Finnish Parallel Corpus of Literary Texts. http://metashare.csc.fi/repository/browse/parfin-2016-parrus-2016-finnish-russian-russian-finnish-parallel-corpus-of-literary-texts/98799e72ec5811e6ba62005056be118e5894ba3bf8f54477acd3e0ac9bbd3ff1/.

Center of Estonian Language Resource. (2017). Transcripts of Riigikogu, https://keeleressursid.ee/en/resources/corpora.

Chanier,T., Poudat,C., Sagot, B., Antoniadis, G., Wigham,C. R. , Hriba, L.,Longhi, J. & Seddah, D. (2014). *The CoMeRe Repository.* https://hdl.handle.net/11403/comere.

Christodouloupoulos, C. & Steedman, M. (2014). Parallel Bible Corpus, https://github.com/christos-c/bible-corpus.

CLARIN-D. (2003). Tiger Corpus, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html.

CLARIN-DK. (2012) DK-CLARIN Almensprogligt korpus, https://clarin.dk/clarindk/item.jsp?id=dkclarin:986010.

clarin:el. (2016). Text Corpus – EMEL. http://hdl.grnet.gr/11500/AUTH-0000-0000-2C5A-B.

clarin:el. (2017). Hellenic Parliament Sittings. http://hdl.grnet.gr/11500/AEGEAN-0000-0000-2545-9.

Daelemans, W. (2005.) De Standaard Corpus, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11372/LRT-391.

ELRA. (2017a). PTPARL Corpus, http://catalog.elra.info/product_info.php?products_id=1179.

ELRA. (2017b). European Parliament Interpretation Corpus, http://catalog.elra.info/product_info.php?products_id=1145.

ELRA. (2018). GeFRePaC - German French Reciprocal Parallel Corpus. http://catalog.elra.info/product_info.php?products_id=633.

Erjavec, T., Ljubešić, N. and Fišer, D. (2017a). Forum corpus Janes-Forum 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1139.

Erjavec, T., Ljubešić, N. and Fišer, D. (2017b). Blog post and comment corpus Janes-Blog 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1138.

Erjavec, T., Ljubešić, N. and Fišer, D. (2017c). News comment corpus Janes-News 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1140.

FIN-CLARIN. (2013). The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version, https://www.kielipankki.fi/corpora/.

Grác, M.. (2011). Corpus of contemporary blogs, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-000E-011B-8.

Hansard-corpus.org. (2017). *Hansard Corpus,* https://www.hansard-corpus.org/.

Hareide, L. (2013). The Norwegian-Spanish Parallel Corpus, Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/73.

Institute for the Languages of Finland. (2014). KOTUS Finnish-Swedish Parallel, http://islrn.org/resources/430-116-345-758-1.

Kapočiūtė-Dzikienė, J., Šarkutė, L. & Utka, A. (2017). Lithuanian Parliament Corpus for Authorship Attribution, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/17.

Lapponi & Sørland. (2016). Talk of Norway, https://github.com/ltgoslo/talk-of-norway.

Ljubešić, N., Erjavec, T. and Fišer, D. (2017a). Twitter corpus Janes-Tweet 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1142.

Ljubešić, N., Erjavec, T. and Fišer, D. (2017b). Wikipedia talk corpus Janes-Wiki 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1137.

Mareček, D. (2008). Czech-English Manual Word Alignment, https://ufal.mff.cuni.cz/czech-english-manual-word-alignment.

Parliament of Finland. (2017). Plenary Sessions of the Parliament of Finland, Helsinki Korp Version. Distributed via the Language Bank of Finland. http://urn.fi/urn:nbn:fi:lb-2017020202.

Piperidis, S., Demiros, I., Prokopidis, P., Vanroose, P., Hoethker, A., Daelemans, W., Sklavounou, E., Konstantinou, M. & Karavidas, Y. (2004). MUSA Multilingual Multimodal Corpus, http://metashare.elda.org/repository/browse/musa-multilingual-multimodal-corpus/9f5d29a263c211e29fc5842b2b6a04d7a2d7266c56224f90ae4cb8f4757bf8ed/.

Pražák, A. & Šmídl, L. (2012). Czech Parliament Meetings, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Radboud University, CLST; Tilburg University, ILK; University of Twente, HMI. (2013). *SoNaR New Media Corpus*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11372/LRT-1502.

Rosa, R. (2016). Czech and English abstracts of ÚFAL papers, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1731.

Segakorpus. (2011). The Mixed Corpus : New Media, http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/.

Språkbanken. (2016). Proceedings of Norwegian parliamentary debates, https://www.nb.no/sprakbanken/show?serial=oai:clarino.uib.no:stortinget&lang=.

Språkbanken. (2017a). Press 65-98, https://spraakbanken.gu.se/eng/resources.

Språkbanken. (2017b). The Riksdag's Open Data, https://spraakbanken.gu.se/eng/resources.

Tadić, M. (2014). Croatian-Slovenian Parallel Corpus, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-239.

Technische Universität Dortmund. (2013). The Dortmund Chat Corpus, http://www.chatkorpus.tu-dortmund.de/.

Tiedemann, J. (2004). Opus, Helsinki Korp Version [text corpus]. Distributed via Kielipankki, http://urn.fi/urn:nbn:fi:lb-2015102201.

Tiedemann, J. (2012a). Tatoeba. Distributed via clarin:el, http://hdl.grnet.gr/11500/ATHENA-0000-0000-2589-C.

Tiedemann, J. (2012b). Opus, http://opus.lingfil.uu.se/.

Truan, N. (2016a). Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/uk-parl.

Truan, N. (2016b). Parliamentary Debates on Europe at the assemblée nationale [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/fr-parl/v1/.

Truan, N. (2016c). Parliamentary Debates on Europe at the Bundestag [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/de-parl/v1/.

Váradi, T. (2005). *Hungarian National Corpus*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11372/LRT-345.