ride. A review journal for digital editions and resources

published by the IDE

EDITORIAL: Digital Text Collections, the third

By Ulrike Henny-Krahmer (University of Würzburg), ulrike.henny (at) uni-wuerzburg.de and Frederike Neuber (Berlin-Brandenburgische Akademie der Wissenschaften), neuber (at) bbaw.de.

1 We are happy to present the third volume on the topic of Digital Text Collections (DTC) which is also the ninth volume in the entire RIDE series. With this issue, we publish the last reviews that we received as a reaction to the first call for reviews on DTCs, together with several additional contributions. The current issue also represents a milestone for RIDE as a whole: With the five reviews of the current issue, a total of 50 reviews are now available in RIDE, including 20 on DTC and 30 on Digital Scholarly Editions.

2 As in the previous volumes, the rationale of the reviews is based on the guidelines for evaluating text collections. In addition, each article is accompanied by a factsheet that summarises the most important information on the resource evaluated. A selection of this data, gathered with a questionnaire and summarized for all the factsheets on DTCs so far, is visualized here as a set of charts.

Contents

3 The current volume contains five reviews, two in English and three in German. Three of the reviews in this volume are dedicated to linguistic corpora, including Czech, Spanish, and multilingual resources. One review addresses a collection of literary texts in English. The fifth review is devoted to a collection of texts from a completely different context: Ancient Papyrology. The reviewed resources are all DTCs that already look back on a longer history of existence and development. The most recent of the resources reviewed in this issue is *Papyri.info* which was still already launched in 2010 and has been reviewed by Lucia Vannini.[1] *InterCorp*, reviewed by Agnes Kim[2],

ShakespearePlaysPlus, reviewed by Katharina Mahler[3], and the *European Parliament Proceedings Parallel Corpus (Europarl)*, reviewed by Claes Neuefeind[4], have already surpassed the first decade of their existence and were published in 2005, 2006, and 2001, respectively. *CORLEC*, a corpus of oral conversations in contemporary Spanish which has been reviewed by Katrin Betz[5], has even existed for almost two decades (since 1991).

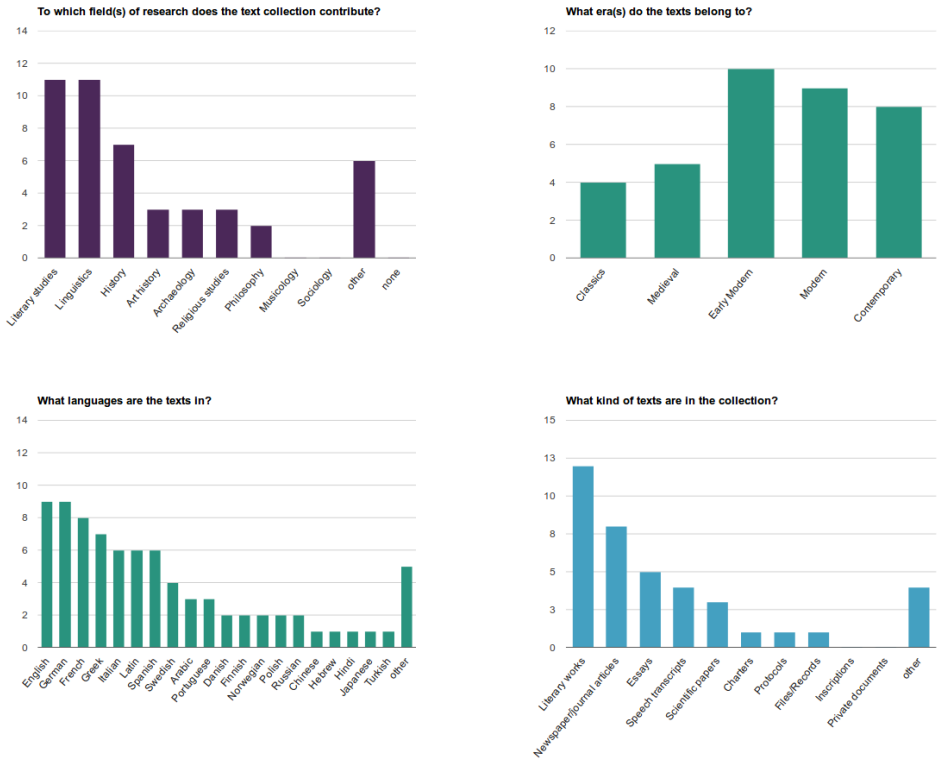

Fig. 1: Research fields (top left), eras (top right), languages (bottom left), and types of text (bottom right) for DTCs.

Fig. 2: Word clouds of the five reviews in RIDE issue 9.

4 Looking at the diversity of content of all the DTCs reviewed in RIDE so far, most reviews addressed resources from the fields of Literature and Linguistics, followed by resources contributing to research in History. There are more reviews on texts belonging to the early modern, modern, and contemporary era than to the medieval or classical era. Regarding languages, English and German texts are most common, but especially the multilingual resources led to a broad language spectrum of the DTCs reviewed in RIDE so far. As for the types of text, DTCs containing literary works have been evaluated most often, followed by collections which include newspaper and journal articles (see Fig. 1). When compared to the general trends, the composition of the current issue reflects them quite well, with a special focus on linguistic corpora. Despite the different contents and languages of the resources examined, it is possible to have a comparative glance at the DTCs of the current issue in terms of their history, methods and practices.

Temporality

5 A recurring challenge in the evaluation of digital resources, which we have also discussed in the editorial of the previous issue on DTCs, is to take into account the temporality of a digital resource on several levels when it is reviewed, used, and evaluated. If older resources are no longer actively developed and updated, they reflect a particular historical methodological state of the art which may not bear comparison with similar resources which were created later on. Nonetheless, they are still valuable,

especially if they grew out of major projects with a comprehensive goal and if they constitute fundamental research which has not been repeated since then.

6 Regarding *CORLEC*, conceived as a reference corpus for contemporary spoken Spanish, Betz states that it is still an important resource for researchers interested in the field, despite its somewhat outdated methods:[6] “It must be anticipated here that certain features and characteristics of *CORLEC*, which are perceived as deficiencies today, met the standards or were even innovative at the moment of its creation. Therefore, its comparability with modern corpora in terms of composition and size are limited”.[7] Betz decides not to subject *CORLEC* to an evaluative comparison with similar modern corpora, but to concentrate on its own characteristics. *Europarl*, in contrast, has been updated several times since it was first published in 2001 - its latest version dating from 2012, also due to successive funding periods. Neuefeind explains that the corpus has grown steadily with regards to the number of languages and words. In the case of *Europarl*, all the different versions are still accessible. Like *CORLEC*, also *Europarl* is still in use: “Although this is an older corpus, it was prepared in a way that still meets the usual standards for the creation of parallel corpora für machine translation. The *Europarl* corpus is therefore still often used, for example as a baseline or benchmark in the development and evaluation of classical systems of statistical machine translation (SMT) [...]”.[8] The ongoing interest in the *Europarl* corpus is evidenced by the fact that it is also included into other aggregated corpora, for example *InterCorp*, reviewed in this same issue of RIDE, where it is part of the so-called “collections”: texts integrated into the core corpus from external resources. Such integration of external sources, or more specifically the aggregation of various independent resources is in itself a point worth of discussion, because it poses special challenges for the editors and administrators of a resource, for its users and also for reviewers.

Aggregated resources

7 When DTCs consist completely or partly of external resources, several questions arise such as: When were different parts integrated into the collection and how does this affect its integrity? How stable are the sources of a collection over time? The last two questions are especially relevant in the case of the Czech parallel corpus *InterCorp* and papyrological portal *Papyri.info* which rely partly or entirely on aggregated texts from external resources.

8 *InterCorp* is partly based on aggregated resources, in that it consists of a core corpus which has been manually aligned, and an extended collection of aggregated pre-existing resources which have been automatically aligned. In qualitative terms, the aggregated texts differ from the corpus because of their automatic alignment. As such, according to Kim, they have many more incorrect alignments.[9] In addition, the metadata of the aggregated resources is often missing or incomplete. *Papyri.info* deals with similar challenges. It is a platform in which resources from several different origins are aggregated and at the same time it offers a research environment for collaborative work on the joined materials. Because of this, updates of the sources challenge the synchronisation of the integrated resources and sometimes the transparency of the data's origin is jeopardized.[10]

9 The described challenges are a strong case for the relevance of a review. In the case of *InterCorp*, the reviewer learned about the problems and peculiarities of aggregated resources through the resource's own documentation. For *Papyri.info* much of the information regarding the state of the data and the update procedures was obtained by the reviewer only through personal communication with those responsible for the resource. For better transparency and reproducibility, however, it would be helpful for users if this was documented on the website. Generally, aggregated resources can be a complex set-up and the reviews can help to disentangle the dependencies for a novice user.

Text and Tools

10 Some of the DTCs in this issue are strongly connected and/or dependent on the use of certain research tools. *ShakespearePlaysPlus*, for example, has been designed explicitly for the WordSmith Tools which is reflected in its data model, its primary usage possibilities, and also its place of publication on the WordSmith Tools website.[11] *InterCorp* is accessible via the corpus manager KonText,[12] "an advanced corpus query interface and corpus data integration middleware built around corpus search engine Manatee-open"[13], developed by the Institute of the Czech National Corpus and used for several different corpora. Europarl is not dependent on third-party tools, but published together with a set of scripts for sentence splitting and alignment which enable the user to create a corpus of the desired language pairs on the fly. It can thus rather be considered a toolkit for a parallel corpus than a finished product.[14]

11 There will rarely be any DTC which does not rely on any tool for its development or use. So far, in our criteria for reviewing DTCs the role of tools is addressed at several points. For instance, it is asked whether tools are provided to analyse the DTC further and if the data can be examined with other tools easily. However, all these questions always refer strongly to the textual data, less to the tool as an independent research object, which in turn would have to be subject to it's very own criteria of reviewing.

Rethinking Classification?

12 In the earlier editorials,[15] we spent some thoughts on how DTCs could be classified further. Our first approach of a classification scheme was based on a terminology relying primarily on the kinds of materials that are collected in a resource (collection of literary texts, letters, historical documents, inscriptions etc.) and the research aims of a collection (e.g. establishment of a canon, comparison, evolution over time).[16] Terminologies used by creators of DTCs and by scholars discussing them pointed out other important factors for a classification such as aspects of preserving, technical and presentational organization (archive, library, database, portal). Still, the reviews on DTCs so far have revealed that our current scheme of classification may not be enough to capture significant nuances of different types of DTCs.

13 With the current issue, for example, it became clear to us that resources can differ significantly in aspects with minor importance in traditional taxonomies of DTCs, even if they usually are considered to be of the same type. For instance, *InterCorp* and *Europarl* are both parallel corpora by definition, but the resources differ in a significant point: their research paradigm and their “target group”. *InterCorp* was created in the first place as a human-readable corpus. Kim stresses in her review that “*InterCorp* [...] explicitly addresses human users and tries to meet their needs.”[17] *Europarl*, on the other hand has “no dedicated user interface, since *Europarl* is primarily meant to support research on machine translation, which has to rely on parallel texts”, as Neuefeind explains.[18] From a user’s and reviewer’s point of view, the quality of these two DTCs is strikingly different which shows that kinds of interfaces, accessibility and usage scenarios are aspects which also should be taken into account in a typology of DTCs.

14 One possibility would be to identify and define DTCs on more levels but this would certainly lead to a typology which is farther away from known categories. Already now, reviewers tend to assign the resources to a range of types which shows that the resources are either not easily classifiable, that the categories are not very well-

established, or that they may not be easily integrated into a typological system. In the current issue, for instance, Vannini identifies *Papyri.info* as “General purpose collection”, “Corpus”, and “Reference corpus”[19] while Mahler identifies *ShakespearePlaysPlus* as “Corpus”, “Canon”, and “Complete works”.[20] Another approach would be to rethink the architecture of the classification scheme as a whole. So far we built the typology “top-down” by defining types of DTCs and assigning them characteristics. Thinking “bottom-up”, types of DTCs would be build only by the combination of characteristics. Such an approach where characteristics are not bound to certain classes would probably be more suitable in the long term to sufficiently capture all the nuances of DTCs which emerge in the still growing and fast changing field of digital (text) technologies. It could also make it easier for the reviewers to decide what characteristics a resource has instead of assigning it to a predefined concept. It can be assumed that such a bottom-up system based on many different characteristics will lead to a “fine threaded carpet” of text collections with relations and similarities on many different levels. Therefore, it seems reasonable to combine both strategies and not to give up the idea of a typology integrating existing categories entirely. Certainly, in the future we will consider both options and evaluate modifications of our taxonomy.

15 To sum up, as editors we found that all the reviews together provide an excellent empirical basis for digging into the ground of digital scholarship which we hope will grow and be explored by the readers of RIDE. Not to forget the valuable discussion of an individual DTC by each review as well as its contribution to the critical discourse on digital scholarly resources, which it helps to keep alive and to mature.

16 We thank the authors of this issue and the peer reviewers involved in the creation of this issue for their enthusiasm, commitment, and also patience. Special thanks to Bernhard Assmann for the “typesetting” and to all of the IDE.

Enjoy the ride!

The editors, Ulrike Henny-Krahmer and Frederike Neuber, November 2018.

Notes

1. Vannini, Lucia. "Review of papyri.info." RIDE 9 (2018). doi: 10.18716/ride.a.9.4. https://ride.i-d-e.de/issue-9/papyri-info/, accessed: November 2, 2018.

2. Kim, Agnes. "Review of 'InterCorp – Ein mehrsprachiges Parallelkorpus des Tschechischen Nationalkorpus (Český národní korpus)'." RIDE 9 (2018). doi: 10.18716/ride.a.9.1. https://ride.i-d-e.de/issue-9/intercorp/, accessed: November 2, 2018.

3. Mahler, Katharina. "Review of “ShakespearePlaysPlus Text Corpus”." RIDE 9 (2018). doi: 10.18716/ride.a.9.3. https://ride.i-d-e.de/issue-9/shakespeare-plays/, accessed: November 2, 2018.

4. Neuefeind, Claes. "Rezension von Europarl." RIDE 9 (2018). doi: 10.18716/ride.a.9.2. https://ride.i-d-e.de/issue-9/europarl/, accessed: November 2, 2018.

5. Betz, Katrin. "Rezension des „Corpus Oral de Referencia de la Lengua Española Contempóranea“." RIDE 9 (2018). doi: 10.18716/ride.a.9.5. https://ride.i-d-e.de/issue-9/corlec/, accessed: November 2, 2018.

6. Daniel Kozák evaluated *PHI Latin Texts* in the previous issue of RIDE in a similar way by saying that even if the methods used by *PHI* do not meet today’s standard, the DTC is widely used in practice and therefore deserves a positive overall evaluation.

7. “Deshalb soll hier vorweggenommen werden, dass bestimmte Merkmale und Eigenschaften des *CORLEC*, die heute als Mangel wahrgenommen werden, zu der Zeit seiner Erstellung wohl dem Standard entsprachen oder aber innovativ waren. Das Korpus ist daher in seiner Zusammensetzung und Größe nur bedingt mit modernen Korpora [...] vergleichbar” (translated into English by the editors). Betz 2018, § 4, https://ride.i-d-e.de/issue-9/corlec/#p4.

8. “Wenngleich es sich um ein älteres Korpus handelt, so entspricht diese Art der Aufbereitung nach wie vor den üblichen Standards zur Erstellung von Parallelkorpora für die maschinelle Übersetzung. Das *Europarl*-Korpus wird dementsprechend noch immer häufig eingesetzt, etwa als *baseline* bzw. *benchmark* in der Entwicklung und Evaluation klassischer SMT-Systeme [...]” (translated into English by the editors). Neuefeind 2018, § 13, https://ride.i-d-e.de/issue-9/europarl/#p13.

9. Cf. Kim 2018, § 10, https://ride.i-d-e.de/issue-9/intercorp/#p10.

10. In her review, Vannini describes how the update process on *Papyri.info* is organized for those source collections that still exist independently. On the other hand, one of the source collections ceded to exist on its own when it was integrated into the portal and yet

another one still exists but its changes are not reflected in *Papyri.info*. Cf. Vannini, § 11f., https://ride.i-d-e.de/issue-9/papyri-info/#p11.

11. Cf. Mahler 2018, https://ride.i-d-e.de/issue-9/shakespeare-plays/.

12. Cf. Kim 2018, § 32ff., https://ride.i-d-e.de/issue-9/intercorp/#p32.

13. https://web.archive.org/web/20181102152657/https://github.com/czcorpus/kontext.

14. Cf. Neuefeind 2018, § 8, https://ride.i-d-e.de/issue-9/europarl/#p8.

15. See https://ride.i-d-e.de/issues/issue-6/editorial-reviewing-digital-text-collections/ and https://ride.i-d-e.de/issues/issue-8/editorial/.

16. Currently, the typology includes: General purpose collection, Corpus, Collection of records, Canon, Complete works/œuvre, Reference corpus, Contrastive corpus, Parallel corpus, and Diachronic corpus. Cf. https://ride.i-d-e.de/issues/issue-6/editorial-reviewing-digital-text-collections/.

17. “Das *InterCorp* richtet sich [...] explizit an menschliche NutzerInnen und versucht deren Bedürfnissen nachzukommen” (translated into English by the editors). Kim 2018, § 2, https://ride.i-d-e.de/issue-9/intercorp/#p2.

18. Neuefeind 2018, abstract, https://ride.i-d-e.de/issue-9/europarl/#abstract.

19. Cf. https://ride.i-d-e.de/issue-9/papyri-info/factsheet/.

20. Cf. https://ride.i-d-e.de/issue-9/shakespeare-plays/factsheet/.