What's New in EuReCo?
Interoperability, Comparable Corpora, Licensing

Marc Kupietz, Eliza Margaretha, Nils Diewald, Harald Lüngen, Peter Fankhauser
Leibniz-Institut für Deutsche Sprache
R5 6–13, 68161 Mannheim, Germany
{kupietz|margaretha|diewald|luengen|fankhauser}@ids-mannheim.de

Abstract

This paper reports on the latest developments of the European Reference Corpus EuReCo and the German Reference Corpus in relation to three of the most important CMLC topics: interoperability, collaboration on corpus infrastructure building, and legal issues. Concerning interoperability, we present new ways to access DeReKo via KorAP on the API and on the plugin level. In addition we report about advancements in the EuReCo- and ICC-initiatives with the provision of comparable corpora, and about recent problems with license acquisitions and our solution approaches using an indemnification clause and model licenses that include scientific exploitation.

1 Interoperability

At the last CMLC workshop on the special topic of interoperability, we presented our general concept of how to make DeReKo data available as comprehensively and freely as possible, taking into account all legal restrictions (Kupietz et al., 2018). In this context, we had defined four levels at which an improvement of accessibility of corpus data was desirable and feasible:

1. data level
2. API level
3. plugin level
4. source code level

In the meantime, there have been relevant developments, especially with regard to the API level, about which we would like to report in the following section.

1.1 API Level

Authorization

The notorious problem with language resources as research data is that in the vast majority of cases, they are not free from the rights of third parties that do not belong to the scientific community (Kupietz et al., 2010). Thus, often the required rights of use have to be transferred from the right holder via a corpus provider to the end user by signing licence agreements. In the scenario of web corpus management tools, these agreements are then referred to, when a user is authenticated. In browsers this procedure is comparatively unproblematic, as the authorization can be handled in a common session management flow. For programmatic access to corpora via Web service APIs, the problem in the area of language resources, even in the context of the large CLARIN initiative, is, however, in practice unsolved. It is necessary to tackle some significant challenges including authentication of users and external user applications, as well as explicit authorization from users for the applications to access their data and the corpora on their behalf. In a complex scenario involving a system of multiple chains of independent applications like the CLARIN Federated Content Search, there is also an issue of delegating authorization from one application to another. In KorAP, to provide API access to DeReKo, we make use of the OAuth 2.0[1] protocol dealing particularly with authorization procedures.

Public Metadata Requests

A particularly simple approach to dealing with legal obstacles, which could even do without authentication and authorization, is to limit the disclosure of data to those that are not protected by copyright, provided that the origin and creation of such data is legitimate, e.g. through licensing agreements and/or copyright exceptions.

In the case of DeReKo, this approach can probably be used for a fairly broad range of applications, covering the analysis of frequency distributions in

[1] https://oauth.net/2/

relation to metadata variables such as publication date, publication location and subject area, thus including application scenarios such as diachronic analysis and comparison of language variants. In the context of the automatically processed part of such investigations, the output of textual, copyrighted data can usually be completely dispensed with; only the hits and their metadata are required.

KorAP manages user access to copyrighted and otherwise restricted data by using a query rewrite mechanism (Bański et al., 2014) that restricts access only to available resources according to user agreements and access location. Public metadata requests allow performing actions such as query search involving restricted data, but only the public metadata of each result are returned as output. The actual text snippet of the matches and non-public metadata are omitted. To provide public metadata requests of all resources to unauthenticated users, we introduce an additional request parameter `access-rewrite-disabled=true` allowing KorAP to disable this particular rewrite. Nonetheless, the rewrite is not disabled for requests requiring user authentication or authorization, such as requesting non-public metadata, requesting not sufficiently licensed corpora, and requesting metadata of virtual corpora restricted to a user or a group.

Listing 1 shows the JSON response to a simple query for the keyword 'Monnemer'. It comprises the generated operator tree for the `query`, the rewritten operator tree for the metadata constraints (`collection`), and the actual `matches`. As shown in Table 1, the current API provides only the metadata for every search hit. The advantage of such an unaggregated output is that it keeps the API simple and lets the user freely analyse any combination of metadata variables.

Listing 2 shows complete functions to query the DeReKo/KorAP API in R for (1) the size of the (virtual) corpus and (2) a search term. Note that the search function also provides a link to a corresponding albeit restricted request[3] to the KorAP web user interface (line 19), so that query results can be validated and analysed also manually. The result of a simple query for 'Hatespeech' is shown in Table 1. Apart from the support for multiple

[2] via http://korap.ids-mannheim.de/api/v1.0/search?ql=poliqarp&q=Monnemer&access-rewrite-disabled=true

[3] As usual, this requires a login – and the user to be authorized to access the requested data including the primary data.


```
{
  "@context": "http://korap.ids-mannheim.de/ns/
      KoralQuery/v0.3/context.jsonld",
  "meta": {
     "fields": ["textSigle","title","availability"],
     ...
  },
  "query": {
     "@type": "koral:token",
     "wrap": {
        "@type": "koral:term",
        "match": "match:eq",
        "layer": "orth",
        "key": "Monnemer",
        "foundry": "opennlp"
     }
  },
  "collection": {
     "operands": [{
           "@type": "koral:doc",
           "match": "match:eq",
           "type": "type:regex",
           "value": "CC-BY.*",
           "key": "availability"
        },{
           "operands": [{
                 "@type": "koral:doc",
                 "match": "match:eq",
                 "type": "type:regex",
                 "value": "ACA.*",
                 "key": "availability"
              },{
                 "operands": [{
                       "@type": "koral:doc",
                       "match": "match:eq",
                       "type": "type:regex",
                       "value": "QAO-NC",
                       "key": "availability"
                    },{
                       "@type": "koral:doc",
                       "match": "match:eq",
                       "type": "type:regex",
                       "value": "QAO.*",
                       "key": "availability"
                    }
                 ],
                 "@type": "koral:docGroup",
                 "operation": "operation:or"
              }
           ],
           "@type": "koral:docGroup",
           "operation": "operation:or"
        }
     ],
     "@type": "koral:docGroup",
     "operation": "operation:or",
     "rewrites": [{
        "@type": "koral:rewrite",
        "src": "Kustvakt",
        "operation": "operation:insertion",
        "scope": "availability(ALL)"
     }]
  },
  "matches": [
     {
        "matchID": "match-WDD17/M00/35548-p730-731",
        "textSigle": "WDD17/M00/35548",
        "availability": "CC-BY-SA",
        "title": "Diskussion:Mannheim"
     },{
        "matchID": "match-WDD17/M00/35548-p777-778",
        "textSigle": "WDD17/M00/35548",
        "availability": "CC-BY-SA",
        "title": "Diskussion:Mannheim"
     },{
        "matchID": "match-HMP18/FEB/00566-p153-154",
        "textSigle": "HMP18/FEB/00566",
        "availability": "QAO-NC",
        "title": "Der Rockstar unter den Comedians"
     }
  ]
}
```


Listing 1: Shortened JSON result of the query for 'Monnemer'[2].

```r
library(jsonlite)
korapurl <- "https://korap.ids-mannheim.de/"
apiurl <- paste0(korapurl, 'api/v1.0/')

fields <- c("corpusSigle", "textSigle", "pubDate", "pubPlace",
            "availability", "textClass")

derekoStats <- function(vc='') {
  return(fromJSON(paste0(apiurl, 'statistics?cq=',
                         URLencode(vc, reserved=TRUE))))
}

derekoQuery <- function(query, vc="", ql="poliqarp") {
  page <- 1
  results <- 0
  request <- paste0('?q=', URLencode(query, reserved=TRUE),
                    ifelse(vc != '', paste0('&cq=', URLencode(vc, reserved=TRUE)), ''),
                    '&ql=', ql);
  print(paste0("corresponding KorAP-UI request: ", paste0(korapurl, request)))
  repeat {
    res <- fromJSON(paste0(apiurl, 'search', request,
        '&count=50&fields=', paste(fields, collapse = ","),
        '&access-rewrite-disabled=true&page=', page))
    if (res$meta$totalResults == 0) { return(data.frame()) }
    for (field in fields) {
      if (!field %in% colnames(res$matches)) {
        res$matches[, field] <- NA
      }
    }
    currentMatches <- res$matches[fields]
    factorCols <- colnames(subset(currentMatches, select=-c(pubDate)))
    currentMatches[factorCols] <- lapply(currentMatches[factorCols], factor)
    currentMatches$pubDate = as.Date(currentMatches$pubDate, format = "%Y-%m-%d")
    if (page == 1) {
      allMatches <- currentMatches
      expectedResults <- res$meta$totalResults
    } else {
      allMatches <- rbind(allMatches, currentMatches)
    }
    print(paste0("Retrieved page: ", page, "/",
                 ceiling(expectedResults / res$meta$itemsPerPage)))
    page <- page + 1
    results <- results + res$meta$itemsPerPage
    if (results >= expectedResults) {
      break
    }
  }
  return(allMatches)
}
```

Listing 2: R sample functions to query the DeReKo / KorAP API. `derekoStats` returns the size of a (virtual) corpus and `derekoQuery` returns the results of a search for some term as a data frame.

textClass	textSigle	pubPlace	availability	pubDate	corpusSigle
staat-gesellschaft biographien-interviews	SOL13/SEP/01462	Hamburg	QAO-NC	2013-09-14	SOL13
politik ausland politik inland	T15/AUG/00332	Berlin	QAO-NC	2015-08-04	T15
politik inland staat-gesellschaft familie-geschlecht	S15/SEP/00251	Hamburg	QAO-NC	2015-09-19	S15
politik inland staat-gesellschaft familie-geschlecht	S15/SEP/00251	Hamburg	QAO-NC	2015-09-19	S15
politik inland staat-gesellschaft familie-geschlecht	S15/SEP/00251	Hamburg	QAO-NC	2015-09-19	S15
kultur literatur	SOL15/SEP/02745	Hamburg	QAO-NC	2015-09-30	SOL15
staat-gesellschaft familie-geschlecht	RHZ15/NOV/03331	Koblenz	QAO-NC	2015-11-05	RHZ15
staat-gesellschaft familie-geschlecht	RHZ15/NOV/03331	Koblenz	QAO-NC	2015-11-05	RHZ15
staat-gesellschaft familie-geschlecht wissenschaft populaerwissenschaft	T15/NOV/02335	Berlin	QAO-NC	2015-11-24	T15
technik-industrie edv-elektronik wissenschaft populaerwissenschaft	T15/DEZ/00520	Berlin	QAO-NC	2015-12-05	T15
politik inland	T15/DEZ/01762	Berlin	QAO-NC	2015-12-17	T15
politik inland	T15/DEZ/01762	Berlin	QAO-NC	2015-12-17	T15
politik inland	T15/DEZ/01762	Berlin	QAO-NC	2015-12-17	T15
politik inland	T15/DEZ/01762	Berlin	QAO-NC	2015-12-17	T15
staat-gesellschaft biographien-interviews	SOL16/JAN/01169	Hamburg	QAO-NC	2016-01-14	SOL16
...					

Table 1: Output of the first 15 results from `derekoQuery("Hatespeech")` using the R function from Listing 2, sorted by publication date.

query languages (Bingel and Diewald, 2015), the API also supports the restriction of searches to virtual sub-corpora based on metadata properties. A more complex example query, that involves a more complex search referring to multiple POS and lemma annotations as well as a virtual corpus definition is shown in Listing 3.

In the near future we will provide libraries to access the DeReKo/KorAP API for different programming languages, starting with R. In order to comply with license agreements and/or the § 60d UrhG text and data mining exception, the access will be limited to academic, non-commercial use.

A current and more detailed documentation of the API can be found in the Wiki of the KorAP component Kustvakt on KorAP's github page.[4]

1.2 Plugin Level

The KorAP user interface provides several entry points to embed results and configuration options for plugins (Diewald et al., to appear). Views can be embedded in so-called *panels* in the user interface, currently available for views on a) the virtual corpus, b) the search result, and c) matches. These entry points are still in an early stage and interactions with the user interface are initially limited. They are planned to be cautiously extended on demand, mainly for security reasons. For example, embedded plugins can already send messages to the global notification system of the user interface, but cannot alter query strings or virtual corpus definitions yet. In case a plugin requires access to the corpus data (for example to provide specific data visualisations), it can communicate via the API, authorized using OAuth 2.0. The first plugins under preparation focus on export capabilities embedded in the search result panel and communicate with the search API.

[4] https://github.com/KorAP/Kustvakt/wiki

2 Comparable Corpora

As discussed at the penultimate CMLC workshop, IDS participates in two essentially complementary initiatives to build comparable corpora: 1) the European Reference Corpus EuReCo (Kupietz et al., 2017) and 2) the International Comparable Corpus ICC (Kirk and Čermáková, 2017; Kirk et al., 2018).

2.1 EuReCo

Within the EuReCo initiave, the first pilot project *DRuKoLA* for the development of a German-Romanian corpus was completed in 2018 (Kupietz et al., to appear(a)). In this context, first virtual comparable corpora based on DeReKo and the Romanian reference corpus CoRoLa were defined and already used for first linguistic investigations (Kupietz et al., to appear(b)). In addition, parts of the Hungarian National Corpus were integrated into EuReCo framework within the 2nd pilot project *DeutUng* (Kupietz et al., to appear(b)).

```
> derekoQuery('[orth="[dw]as" & (tt/p=PRELS | opennlp/p=PRELS)] [tt/p=ADJA] [tt/l=
    sein]', 'corpusTitle="Der Spiegel" & pubDate since 2017-01-01')
[1] "Retrieved page: 1/1 (2.146951205 s)"
        textSigle    pubDate                                  textClass
1  S17/SEP/00243 2017-09-16 staat-gesellschaft biographien-interviews
2  S18/APR/00171 2018-04-14 staat-gesellschaft biographien-interviews
3  S18/MAR/00397 2018-03-24         wissenschaft populaerwissenschaft
4  S18/SEP/00396 2018-09-22 staat-gesellschaft biographien-interviews
5  S18/JUL/00234 2018-07-21                              politik inland
6  S17/AUG/00322 2017-08-26   biographien-interviews kultur literatur
7  S18/SEP/00156 2018-09-08                              politik inland
8  S18/AUG/00281 2018-08-18                              sport fussball
9  S17/MAI/00359 2017-05-27     staat-gesellschaft familie-geschlecht
10 S18/MAI/00069 2018-05-05              freizeit-unterhaltung reisen
11 S18/JUN/00010 2018-06-02                            politik ausland
12 S18/APR/00342 2018-04-28                           kultur literatur
13 S18/JAN/00285 2018-01-20                                kultur film
```

Listing 3: Complex query for 'das' (the/that) or 'was' (what) annotated as relative pronoun by TreeTagger or by the OpenNLP tools, followed by an attributive adjective and a form of 'sein' (to be), according to the TreeTagger annotations, in a virtual sub-corpus restricted to issues of the news magazine Der Spiegel published since 1st January 2017.

2.2 ICC

While EuReCo rather uses a primordial sample design approach (Kupietz et al., 2010) and wants to enable users to define a virtual comparable corpus based on the underlying individual language corpora, depending on the task and language domain investigated, the composition of the target corpus of the ICC initiative is determined from the outset to mimic the one of the *International Corpus of English* (ICE), with a few exceptions. The ICC plan is to complete at least the written linguistic corpus parts for some languages by 2019.

3 Legal issues and Licensing

The German reference corpus DeReKo relies on and continuously acquires licenses for the scientific use of text content, mostly from publishing companies. Many newspaper publishers are prepared to grant a free license for the use of their latest content in the DeReKo scenario (i. e. performing query and analysis via the dedicated corpus research interface that displays results only as text snippets in a KWIC format, or querying metadata and deriving statistics via the new KorAP API described above). Book publishers (both of fiction or specialised books), however, have on average not been not so generous, i. e. many do not reply to our acquisition campaigns in the first place, and those who do, grant licenses only for a limited number of titles most of the time. We attribute this firstly to some reluctance to make content available that is still actively being marketed, and secondly to the much higher need of time and effort to select and provide book content to external archives because it is simply not part of their established workflows (Kupietz and Lüngen, 2014). Since 2018, we have come across the new phenomenon that book publishers told us that they appreciated our project and would be willing to grant licenses, but they could not say whether they could grant rights for scientific use of their books in DeReKo, not being able to know whether they actually are in a legal position to do so. They would have to look into each particular author contract to assess this, which would be (too) costly to do (given that DeReKo would like to get the licenses for free). The risk of being sued for a breach of intellectual property rights by an author if they still granted us licenses was indeed considered low, however seemed not worth to be taken by the publisher if they have no gain from the deal. Another publisher had sought a legal opinion which stated that the type of use of their content in DeReKo was not at all covered by the model contract for authors provided by the Publishers and Booksellers Association that they generally use.

The main reason for these apparently new problems and the deterioration in the acquisition of books was that §§ 31a UrhG "Contracts for Unknown Types of Use" and 137l UrhG "Transitional Provisions for New Types of Use", which entered the German Copyright Act on 1 January 2008 and which essentially state that older author contracts automatically permit electronic exploitation unless

the author objects, initially made the acquisition of rights much easier. Subsequently, it seemed that publishers first reviewed their author contracts, including also newer ones that were no longer actually affected by the amendment, with regard to electronic rather than scientific exploitation. This seems to have changed by now.

As a first reaction to the problem, we have added a new indemnity clause against third party intellectual property claims to our standard agreements with the help of our legal experts. The idea is that the risk will be taken by the IDS, or that explicit licenses will subsequently be acquired directly from the authors. A second measure will be to approach the Publishers Association and ask them to explicitly include the scientific type of use of linguistic analysis in their model contract. In doing so, we hope to make book publishers more willing to grant free licenses for the scientific use of text content in the DeReKo scenario.

4 Conclusions

Apart from the reports on progress in the provision of comparable corpora in European languages and the long-term consequences of a 10-year-old amendment to the German Copyright Act, this paper has above all shown new ways in which, despite legal and ultimately economic hurdles, large corpora can be opened up for programmatic frequency analyses without infringing on the interests or rights of right holders and without incurring great technical expense.

The method we have presented here and implemented for DeReKo and KorAP basically follows our motto borrowed from Jim Gray (2003):

> *If the data is to big or not allowed to move, put the computation near the data.*
> (cf. Kupietz et al., 2010, 2014, 2018)

with the addendum:

> *If not all computation can be put near the data, move just such data that is allowed and required to move.*

References

Piotr Bański, Nils Diewald, Michael Hanl, Marc Kupietz, and Andreas Witt. 2014. Access Control by Query Rewriting: the Case of KorAP. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik. European Language Resources Association (ELRA).

Joachim Bingel and Nils Diewald. 2015. KoralQuery – a General Corpus Query Protocol. In *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*, Vilnius, Lithuania.

Nils Diewald, Verginica Barbu Mititelu, and Marc Kupietz. to appear. The KorAP user interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique*.

Jim Gray. 2003. Distributed Computing Economics. Technical Report MSR-TR-2003-24, Microsoft Research.

John Kirk and Anna Čermáková. 2017. From ICE to ICC: The new International Comparable Corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017*, pages 7 – 12. IDS. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6249.

John Kirk, Anna Čermáková, Signe Oksefjell Ebeling, Jarle Ebeling, Michal Kren, Karin Aijmer, Vladimir Benko, Radovan Garabik, Rafal Gorski, Jarmo Jantunen, Marc Kupietz, Maria Simkova, Thomas Schmidt, and Oliver Wicher. 2018. Introducing the International Comparable Corpus. In *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition), CECL Papers 1*, pages 96 – 97, Louvain-la-Neuve. Université catholique de Louvain.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, page 1848–1854, Valletta, Malta. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Marc Kupietz, Anda Cosma, and Andreas Witt. to appear(a). The DRuKoLA project. *Revue Roumaine de Linguistique*.

Marc Kupietz, Ruxandra Cosma, Dan Cristea, Nils Diewald, Beata Trawinski, Dan Tufis, Tamás Váradi, and Angelika Wöllstein. to appear(b). Recent developments in the European Reference Corpus (EuReCo). In *Proceedings of UCCTS 2018*, Louvain-la-Neuve.

Marc Kupietz, Nils Diewald, and Peter Fankhauser. 2018. How to get the computation near the data: improving data accessibility to, and reusability of analysis functions in corpus query platforms. In *Proceedings of the LREC 2018 Workshop "Challenges in the Management of Large Corpora (CMLC-6)" 07 May 2018 – Miyazaki, Japan*, pages 20 – 25, Paris. European language resources association (ELRA).

Marc Kupietz and Harald Lüngen. 2014. Recent developments in DeReKo. In *Proceedings of the ninth conference on international language resources and evaluation (LREC'14)*, pages 2378–2385, Reykjavik, Iceland. ELRA.

Marc Kupietz, Harald Lüngen, Piotr Bański, and Cyril Belica. 2014. Maximizing the potential of very large corpora: 50 years of big language data at IDS Mannheim. In *Proceedinigs of the LREC-2014-workshop challenges in the management of large corpora (CMLC2)*, pages 1 – 6, Reykjavik / Paris. ELRA.

Marc Kupietz, Andreas Witt, Piotr Bański, Dan Tufiş, Dan Cristea, and Tamás Váradi. 2017. EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017*, pages 15 – 19. IDS. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6258.