# The German Reference Corpus DeReKo:
# New Developments – New Opportunities

**Marc Kupietz**[*], **Harald Lüngen**[*], **Paweł Kamocki**[*†], **Andreas Witt**[*]

[*]Institut für Deutsche Sprache,
R5, 6-13; D-68161 Mannheim
[†]European Language Resources Distribution Agency
9 rue des Cordelières; F-75013 Paris
{kupietz|luengen|witt}@ids-mannheim.de, kamocki@elda.org

### Abstract

This paper discusses current trends in DeReKo, the German Reference Corpus, concerning legal issues around the recent German copyright reform with positive implications for corpus building and corpus linguistics in general, recent corpus extensions in the genres of popular magazines, journals, historical texts, and web-based football reports. Besides, DeReKo is finally accessible via the new corpus research platform KorAP, offering registered users several news features in comparison with its predecessor COSMAS II.

**Keywords:** reference corpus, very large corpus, diversity, intellectual property rights, copyright reform, corpus analysis, comparable corpora, collocation analysis, word embeddings

## 1. Introduction

The German Reference Corpus DeReKo is presumably the largest archive of German language texts designed for linguistic research (Kupietz et al., 2010). As of 2018 (Institut für Deutsche Sprache, 2018), it contains more than 42 billion tokens, comprising a multitude of genres such as newspaper text, fiction, or specialised text, with a current growth rate of 3.1 billion word per year. Besides the constant acquisition of new newspaper sources, one focus of corpus extension in the past four years has been on the curation of content from sources of computer-mediated communication (CMC), cf. Margaretha and Lüngen (2014); Schröck and Lüngen (2015); Lüngen et al. (2016).

In 2017, several other new genres that had previously not been available in DeReKo, have been acquired and included in the latest release (Institut für Deutsche Sprache, 2018). These are discussed in Section 3. Likewise as of 2017, the bulk of DeReKo can be accessed online via the new corpus research platform KorAP, which offers several new corpus query features, which are discussed in Section 4. Actual and envisioned improvements in DeReKo's legal status are discussed in Section 2.

## 2. Legal and Licensing Situation

### 2.1. License Types

The Institut für Deutsche Sprache (IDS), DeReKo's host institution, is not the owner of DeReKo's content. Rather, we have more than 200 license agreements with rights holders, mostly publishers, granting the use of texts for non-commercial, scientific research by registered users and strictly within the query-and-analysis-only framework (cf. Kupietz/Lüngen, 2014) offered through the corpus research interfaces of COSMAS II and KorAP (see Section 4.).

Due to the fact that the range of lexicographically edited dictionaries and lexicons on the German language offered by commercial publishers has deteriorated considerably in the last few years, IDS is interested in acquiring additional licensing that would allow for building, editing, and (semi-)commercially publishing fundamentals for comprehensive print and on-line dictionaries based on statistical and lexicographic analyses and evaluations of (parts of) DeReKo, in the future. Such a license extension towards commercial use could also open new application fields e. g. for DeReKo-based language models in commercial contexts and also in academic contexts whenever copies of DeReKo(-parts) are required which are currently prohibited by its license terms. Thus, this way the re-usability and productivity of DeReKo could be increased considerably (presented in Kupietz and Belica, 2013), the efforts required for such license renegotiations and their outcome are, however, not yet fully clear.

### 2.2. 2017 Copyright Reform

To address the new possibilities of distributing and using copyright-protected content that have emerged due to the development of digitisation and the web in the past 20 years, a reform of parts of the German *Urheberrechtsgesetz* (Copyright Act) has been passed by the German legislator in July 2017. The new "Act on Adapting Copyright to Current Requirements of the Knowledge Society" (UrhWissG, 2017) will enter into force on 1 March 2018.[1] It introduces new exceptions *(Schranke)* to exclusive rights, regulating how copyright-protected content may be used in the spheres of education and research, and within so-called knowledge institutions. From the perspective of DeReKo and the IDS corpus extension project, the paragraph § 60d on text and data mining in which the term "corpus" occurs several times, is most interesting. It states that available content may, without explicit permission from the copyright holders and for strictly non-commercial research purposes, be automatically reproduced, structured, and categorised for building a corpus which can then be exploited by and shared with a group of users for common research. This TDM exception

---

[1]The reform echoes the *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market* of 14 September 2016 (European Comission, 2016).

cannot be overridden by separate agreements. Upon completion of the research activities, however, the corpus and its copies must be deleted or handed over to an archive, library, or other educational institution. In the explanatory statement of the ministerial draft, it is pointed out that the new law does not imply a right to access copyright-protected material, but that it holds for cases where access is already given e. g. through a library or via the internet (Referentenentwurf, 2017). The said uses require equitable compensation to be paid to the collecting society VG WORT, which means that agreements will need to be negotiated.

While previously the prevailing legal opinion on the redistribution of web content that constitutes a copyright-protected work (*das Werk*) outside small projects was that an authorisation would have to be obtained form every single author (Beißwenger et al., 2017), the new legal situation will enable us to legally build linguistic corpora by scraping German language content from the web (be it original web genres, CMC, or any kind of documents that are offered openly via the web), and to process, and republish them in DeReKo for analysis by our registered, scientific user community.

## 3. Quantitative and Qualitative Extensions

The most recent DeReKo release (Institut für Deutsche Sprache, 2018) contains over 42 billion tokens. From most sources we continually get new material, and DeReKo's annual growth is currently at 3.1 billion tokens, see Fig. 1. After a focus on corpora of computer-mediated communication in the last few years (cf. Lüngen, 2017), recent additions to DeReKo comprise a press archive consisting of popular magazines, journals, and daily newspapers, as well as a Football Linguistics Corpus.
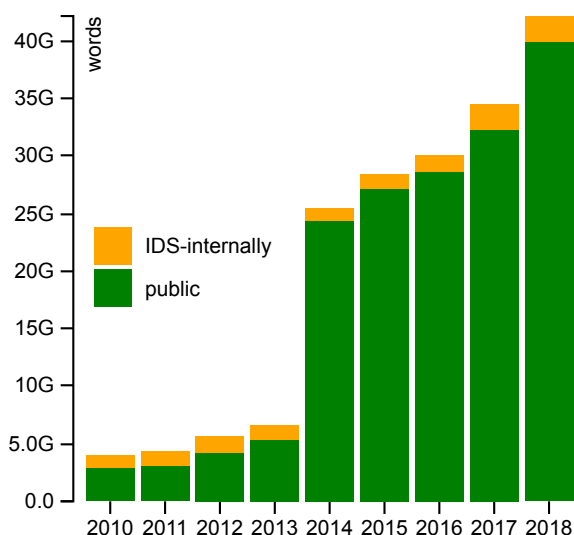


Figure 1: DeReKo-growth from 4 billion words in 2010 to 42 billion words in 2018.

As the bulk of DeReKo has always consisted of text data from daily newspapers, the new acquisitions have improved the genre dispersion in DeReKo in general. In the following, we give a brief description of the new corpora.

### 3.1. Press including journals and popular magazines

In an extended cooperation with a commercial news database provider, DeReKo has acquired new licenses for popular magazines. Most of them are weeklies such as *Stern, Brigitte,* or *Gala*, with the editions starting between 2007 and 2015, the earliest already in 1996. 23 of them, i.e. those that contained most data, have been included in the current release (Institut für Deutsche Sprache, 2018). They comprise altogether 90,272,352 tokens.

| magazine | start | tokens |
|---|---|---|
| Stern | 1996 | 50,167,964 |
| Brigitte | 2009 | 7,643,387 |
| Hörzu | 2007 | 5,466,757 |
| GEO | 2009 | 3,461,172 |
| Gala | 2015 | 2,467,552 |
| Essen und Trinken | 2015 | 1,814,979 |

Table 1: Examples of popular magazines newly included in DeReKo

Up to now, almost no popular magazines were available in DeReKo, as previously, unlike newspaper publishers, magazine publishers would frequently not hold digital exploitation rights for their content, such that agreements would have to be concluded with every author of an article in such a publication.

By the same cooperation, DeReKo has also acquired licenses for several newspapers previously not contained, mostly regional papers, but also e.g. the nation-wide *Die Welt*, as well as *Dolomiten*, the German-language daily with the highest circulation in Alto Adige, Italy. Most of the new titles start in the archive around the year 2000, the earliest one in 1992. Several papers from the north of the German-speaking area are also contained, such as *Neue Osnabrücker Zeitung* or *Ostseezeitung*, thus closing a certain gap (cf. Kupietz/Lüngen, 2014) in the regional dispersion of DeReKo. 53 new newspapers comprising altogether 6,248,900,215 tokens are newly included in the release (Institut für Deutsche Sprache, 2018). More (regional and local) newspapers from the same archive will be included in the releases to come.

| daily | start | tokens |
|---|---|---|
| Kölner Stadt-Anzeiger | 2000 | 551,722,542 |
| Die Welt | 1999 | 434,170,447 |
| Hamburger Abendblatt | 1999 | 346,807,799 |
| Dolomiten (Italy) | 2000 | 175,535,474 |
| Neue Osnabrücker Zeitung | 2012 | 94,276,568 |
| Nordwest-Zeitung | 2016 | 68,421,655 |
| Ostsee-Zeitung | 2016 | 37,221,069 |

Table 2: Examples of dailies newly included in DeReKo

Finally, the acquired archive contained a set of trade and technical journals, all starting in 2017. They include titles such as *allgemeine fleischer zeitung, Deutsches Ärzteblatt,* or *Technische Textilien*. 53 of them comprising altogether 1,183,122 tokens, are included in the latest release. Though

representing a smaller portion than the other extensions, they also improve the dispersion of genres in DeReKo.
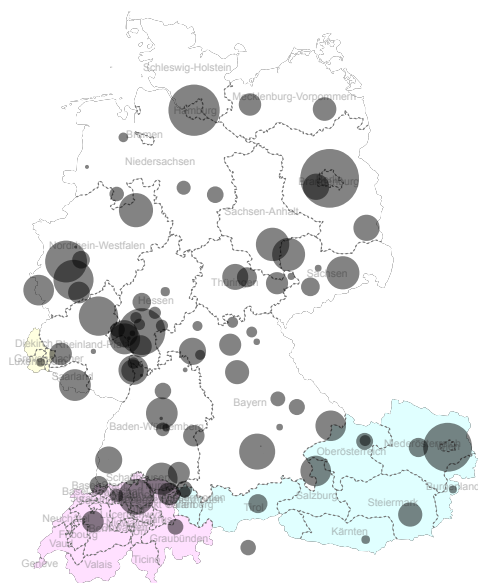


Figure 2: Geographical coverage with DeReKo press sources. The size of a circle corresponds to number of tokens.

The quantitative regional distribution of all (old and new) press sources in (Institut für Deutsche Sprache, 2018) is visualised in 2. To better judge the differences or similarities amongst the new and selected old corpora in DeReKo, we calculated the NMDS-projected distance map based on frequency lists shown in Fig. 3. In the map, DeReKo as a whole is shown in light blue font and as expected, the corpora of nationwide papers with the greatest extension in the archive (s=Der Spiegel; u=Süddeutsche; z=Die Zeit) lie close to it. Amongst the popular magazines, there is a main group shown in orange, with a cluster of yellow press/women's/family magazines (dgb, dak, gal, nwt, fis, edf, brg, ndo, elt), and a small cluster of interior design magazines (scw, cou). Ppm and art are a pop science and an art magazine, which for some reason get situated close to DeReKo as a whole, too. The blue group is clustered so closely, but its members lie next to each other and all represent food/cooking magazines (bee, eut, chk). The group shown in gold contains TV magazines (hrz, tvd, gng) and clusters nicely. Outliers different colours are loz (a corpus of fiction), wpd (wikipedia articles), and wdd (wikipedia talk pages). Ph, shown in green, is a monthly psychological magazine.[2]

### 3.2. Football Linguistics Corpora

Simon Meier of TU Berlin has kindly made available for DeReKo his *Football Linguistics Corpora* (Meier, 2017) (Meier, 2017), consisting of liveticker protocols and match reports from two different web publishers, from who we subsequently could also acquire the appropriate licenses.

---

[2]More detailed, interactive visualizations can be found on DeReKo's archive page under http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html.
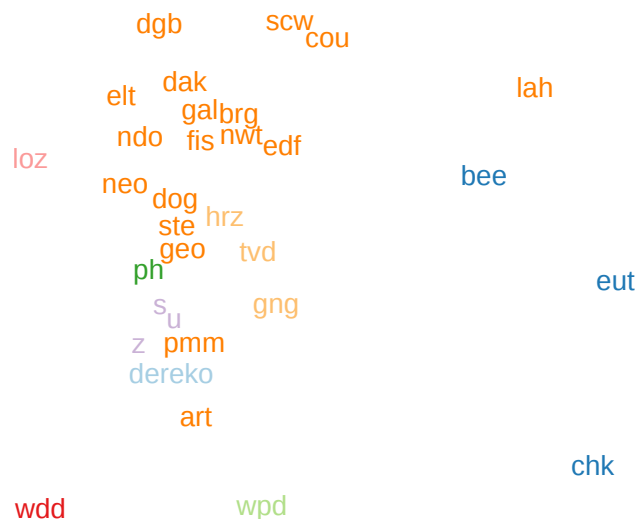


Figure 3: NMDS projection of the distances between different DeReKo sources. Based on Kilgarriff's (2001) *comparing corpora* approach (see Kupietz and Lüngen, 2014, for a detailed description of the procedure).

These reports and protocols cover all matches of the national football leagues First and Second Bundesliga, the European Champions and Europa Leagues as well as the Euro Cups, starting in 2006. With altogether 8.85 million tokens, they are part of the current release (Institut für Deutsche Sprache, 2018) and further strengthen the web genres and the sports domain within DeReKo.

### 3.3. Textgrid Digital Library

In 2017, we have also prepared an I5 version of the freely available TextGrid Digital Library which had previously been derived from the zeno.org Online Library.[3] It contains more than 2500 fictional and non-fictional texts (from domains such as fiction, fairy tales, (cultural) history, art, music, science, and philosophy) from the beginnings of printing until the first decades of the 20th century, amongst them nearly all German canonical literary texts for which copyright protection has expired (TextGrid, 2016). The TextGrid Digital Library corpus in its I5 incarnation contains almost 170 million tokens.

DeReKo is a corpus of contemporary German and has up to now contained only a few historical corpora, which are considered as "milestones" of the development of the German language i.e. still having influence on and relevance for the current shape of German, such the Goethe corpus or the Grimm corpus. We consider the TextGrid Digital Library relevant in that sense as well.

During the conversion to I5, it turned out that for up to 500 texts or so, only their 20th century publication date was available in a metadata field while the time of their actual creation was hidden in a bibliographic information string or not contained at all in the TEI source. Since I5 provides a dedicated metadata field for the date or period of creation and its specification seems essential for working with DeReKo (e.g. for any analysis of variation in time), we

---

[3]http://www.zeno.org

are currently running a post-processing campaign in which the creation dates are being derived from bibliographical information strings or, if not available, from external sources. For this reason, it is not contained in the release (Institut für Deutsche Sprache, 2018).

## 4. Querying and analyzing DeReKo

### 4.1. Now publicly accessible via KorAP

The corpus analysis platform KorAP (Bański et al., 2012; Diewald and Margaretha, 2016) that has been developed at the IDS as a successor to COSMAS II (Bodmer, 1996, 2005) since 2012 has been publicly available for querying DeReKo since May 2017.[4] Compared with COSMAS II, some essential features such as collocation analysis and sorting and aggregation of query hits are still under development. On the other hand, KorAP already offers some unique features for the analysis of DeReKo. One is that KorAP can efficiently query the complete DeReKo-collection with more than 40 billion words in one archive. Even though COSMAS II has no principle token number limitation related to 32 bit integers like many other query engines of its generation (COSMAS II was already designed in 1994), it is in effect limited by memory constraints to currently not much more than 7 billion words with one annotation layer per archive.

### 4.2. Multiple query languages

One more feature that distinguishes KorAP not only from COSMAS II, but probably from all currently existing corpus query systems is the support for multiple corpus query languages. This allows users coming from different research communities or traditions to readily use KorAP using the language they are accustomed to and enables the user to benefit from combining the advantages of different query languages. The supported query languages currently include the ANNIS Query Language (AQL; Rosenfeld 2010), Cosmas II (al Wadi, 1994), and Poliqarp (a CQP variant / extension; Przepiórkowski et al. 2010).

### 4.3. Multiple annotation layers

The support of an in principle unlimited number of linguistic annotation layers was one of the main reasons for the development of KorAP. The goal with regard to DeReKo is to add all annotations that are of interest to users or to tool providers and to make them queryable via KorAP. So far, the spectrum of queryable annotation layers was extended by dependencies and constituents (co-references, RST relations, etc. will follow). In addition, the set of different, competing annotations within one level was extended in order to improve the handling of erroneous and uncertain classifications. The use of linguistic annotations in search queries is often indispensable for the investigation of linguistic phenomena at an abstract level. If, however, they are needed for quantitative investigations, corpus-based hypothesis testing,

and even for exploring phenomena beyond the finding of examples, their use is not trivial, and some possible pitfalls have to be considered as they do not have the status of observations, but rather the status of interpretations (Belica et al., 2011). This fact is also relevant for part-of-speech annotations for which an accuracy of up to 97-98% are reported (Giesbrecht and Evert, 2009). One problem even at such low error rates is that errors are not distributed uniformly and with an unfavourable tendency (Belica et al., 2011, p. 466). For typical linguistic questions, lower accuracy values are to be expected since such questions are typically related to less clear or common phenomena and the accuracy of the results for a handful of particular search queries can hardly be derived from the reported average accuracy of the annotation tool. Rather, it must be taken into account that although most constructions are almost always correctly annotated, others are almost always wrong, so that in the case of search requests to the latter, a very low recall must be expected. Since false-negative hits are not visible and thus cannot be easily identified, such a situation can easily lead to misinterpretations.

In order to improve the manageability of annotations, KorAP offers the possibility to use and query arbitrarily many competing annotations. Depending on the task, by using disjunctive queries, recall can be maximized (and false negatives minimized), and by using conjunctive queries, precision can be maximized (and false positives minimized). As shown in table 3 (similar example query in figure 4), the results obtained this way can differ significantly from one another, yielding almost 2 million more results for the disjunctive query than for the conjunctive one.

Furthermore, depending on the phenomenon to be investigated and the language domain to be examined, the annotations which promise the best results can be used in a targeted manner. Ideally, the annotations are based not only on different algorithms, but also on different training data.

### 4.4. DeReKo-based distributional analysis

Having a long history in providing and using distributional models based on collocation analysis for paradigmatic and syntagmatic analysis (Keibel and Belica, 2007)[5], we now investigate the pros and cons of these models in comparison to models based on word embeddings with respect to different linguistic applications. For this purpose we have developed a pipeline for training structured skip-gram networks (Ling et al., 2015) and building a collocation frequency database from DeReKo-releases, as well as a web interface to explore and compare them for syntagmatic (see figure 5 and paradigmatic relations. The interface will be made availably shortly. For the publication of the underlying models, however, we first need to investigate the possible implications on additional license fees and compensations to the VG WORT (see section 2.2.).

### 4.5. Using DeReKo in contrastive studies

In order to make DeReKo (re-)usable also in contrastive and cross-linguistic studies DeReKo takes part in two initiatives

---

**KorAP interface:** `[orth=das & opennlp/p=PRELS | tt/p=PRELS]`

availability eq /CC-BY.*/
or availability eq /ACA.*/

in one Collection with [Poliqarp / Cosmas II / Annis QL / CQL v1.2]  ☑ Glimpse

ein Flugzeug, **das** in der Welt nicht seinesgleichen [hat] ... amerikanischen Journalisten Richard Cook im "Wallstreet Journal" über das sowjetische Turboprop-Flugzeug TU 114 das größte P... g der Welt, die TU 114, brachte nicht nur einen hohen Politiker, den Ersten Stellvertreter des Vorsitzenden

**Ein Flugzeug, das in der Welt nicht seinesgleichen hat** by Cook, Richard (1959-07-08)  [BZK/D59/00371]

| Foundry | Layer | ein | Flugzeug | das | in | der | Welt | nicht | seinesgleichen | hat |
|---|---|---|---|---|---|---|---|---|---|---|
| corenlp | p | ART | NN | PRELS | APPR | ART | NN | PTKNEG | VVPP | VAFIN |
| marmot | m | case:nom gender:ne... number:sg | case:nom gender:ne... number:sg | case:nom gender:ne... number:sg | | case:dat gender:fe... number:sg | case:dat gender:fe... number:sg | | case:* gender:* number:* | mood:ind number:... person:3 tense:pres |
| marmot | p | ART | NN | PRELS | APPR | ART | NN | PTKNEG | PIS | VAFIN |
| opennlp | p | ART | NN | PRELS | APPR | ART | NN | PTKNEG | VVINF | VAFIN |
| tt | l | eine | Flugzeug | die | in | die | Welt | nicht | seinesgleichen | haben |
| tt | p | ART | NN | PDS PRELS | APPR | ART | NN | PTKNEG | PIS | VAFIN |
| corenlp | c | | | | | | | | | |

ein    das    VP    hat

**DEREKO-VECTORS** | Influencer | [SEARCH] [Options]

Semantics (TSNE-map)  Semantics (SOM)  Syntagmatic (collocators)

| # | w' | max(a) | (a) | $\Sigma a/\Sigma w'$ | $\perp(a/c)$ | $\Sigma a/\Sigma w$ | collocator |
|---|---|---|---|---|---|---|---|
| 1 | | 0.996 | 0.661 | 9.339e-5 | 3.895e-4 | 4.876e-5 | Neudeutsch |
| 2 | | 0.995 | 0.751 | 9.325e-5 | 4.507e-4 | 5.432e-5 | Social |
| 3 | | 0.995 | 0.099 | 9.322e-5 | 9.322e-5 | 2.200e-5 | Sogenannte |
| 4 | | 0.992 | 0.666 | 6.001e-5 | 3.654e-4 | 5.358e-5 | Youtuber |
| 5 | | 0.989 | 0.193 | 9.270e-5 | 1.430e-4 | 4.449e-5 | Millennial |
| 6 | | 0.988 | 0.579 | 6.037e-5 | 3.140e-4 | 5.326e-5 | Online-Kampagnen |
| 7 | | 0.988 | 0.286 | 5.811e-5 | 1.567e-4 | 4.954e-5 | App-Entwickler |
| 8 | | 0.988 | 0.195 | 5.116e-5 | 9.825e-5 | 3.116e-5 | Facebook-Managerin |
| 9 | | 0.988 | 0.099 | 5.902e-5 | 5.902e-5 | 4.244e-5 | Abo-Modelle |
| 10 | | 0.987 | 0.385 | 5.154e-5 | 1.960e-4 | 4.343e-5 | Online-Kanäle |
| 11 | | 0.987 | 0.193 | 5.895e-5 | 1.110e-4 | 4.463e-5 | neudeutsch |
| 12 | | 0.987 | 0.099 | 9.256e-5 | 9.256e-5 | 1.945e-5 | Multi- |
| 13 | | 0.986 | 0.945 | 9.116e-5 | 5.873e-4 | 5.554e-5 | Follower |
| 14 | | 0.986 | 0.099 | 9.243e-5 | 9.243e-5 | 3.130e-5 | hipper |
| 15 | | 0.985 | 0.390 | 6.110e-5 | 2.156e-4 | 5.027e-5 | Evsan |
| 16 | | 0.985 | 0.189 | 9.237e-5 | 1.408e-4 | 4.578e-5 | Self-Publishing |
| 17 | | 0.984 | 0.377 | 5.874e-5 | 2.012e-4 | 4.204e-5 | Web-Angeboten |
| 18 | | 0.984 | 0.283 | 5.921e-5 | 1.567e-4 | 4.340e-5 | crossmedial |
| 19 | | 0.983 | 0.750 | 6.095e-5 | 4.047e-4 | 5.097e-5 | Myspacecom |
| 20 | | 0.983 | 0.675 | 9.435e-5 | 4.039e-4 | 5.431e-5 | Internet-Tagebücher |
| 21 | | 0.983 | 0.284 | 6.095e-5 | 1.680e-4 | 4.999e-5 | Business-Partner |
| 22 | | 0.981 | 0.285 | 6.086e-5 | 1.528e-4 | 3.602e-5 | Kommunikationsinstrument |
| 23 | | 0.981 | 0.098 | 9.196e-5 | 9.196e-5 | 1.608e-5 | sogenannter |
| 24 | | 0.981 | 0.098 | 9.196e-5 | 9.196e-5 | 3.072e-5 | Corporate |
| 25 | | 0.981 | 0.098 | 5.858e-5 | 5.858e-5 | 3.816e-5 | Content- |
| 26 | | 0.980 | 0.190 | 5.852e-5 | 1.094e-4 | 4.464e-5 | Web-Unternehmen |
| 27 | | 0.980 | 0.098 | 9.187e-5 | 9.187e-5 | 1.746e-5 | Successful |
| 28 | | 0.979 | 0.763 | 6.070e-5 | 4.109e-4 | 5.405e-5 | Pinterest |
| 29 | | 0.979 | 0.376 | 5.849e-5 | 2.109e-4 | 4.456e-5 | Online-Bewertungen |
| 30 | | 0.979 | 0.285 | 5.085e-5 | 1.428e-4 | 4.278e-5 | Onlinekommunikation |

| LLR | $PMI^3$ | nPMI | raw | collocator |
|---|---|---|---|---|
| 156 | -79.6 | 0.62 | 6 | Beeinflusser |
| 86 | -79.8 | 0.66 | 3 | Instagram-Konten |
| 150 | -80.4 | 0.60 | 6 | Brandnew |
| 261 | -80.8 | 0.54 | 12 | Followern |
| 820 | -81.3 | 0.42 | 55 | Marketing |
| 55 | -81.9 | 0.62 | 2 | Marketingwaffe |
| 54 | -82.2 | 0.61 | 2 | Umschmeichelung |
| 75 | -82.3 | 0.58 | 3 | Social-Media-Stars |
| 240 | -82.9 | 0.47 | 13 | seinquot |
| 52 | -82.9 | 0.59 | 2 | Hotelmarketing |
| 52 | -82.9 | 0.59 | 2 | PATTERSON |
| 162 | -83.0 | 0.50 | 8 | Einflüsterer |
| 52 | -83.0 | 0.59 | 2 | Influencer-Marketing |
| 331 | -83.8 | 0.40 | 22 | Social |
| 88 | -83.8 | 0.52 | 4 | Influencer |
| 164 | -84.1 | 0.46 | 9 | Follower |
| 27 | -84.1 | 0.60 | 1 | Campaignquot |
| 27 | -84.4 | 0.59 | 1 | Top-Blogger |
| 47 | -84.5 | 0.55 | 2 | Video-Blogger |
| 405 | -85.0 | 0.33 | 35 | sogenannten |
| 46 | -85.1 | 0.53 | 2 | Google-Suchanfragen |
| 26 | -85.1 | 0.57 | 1 | Marketern |
| 45 | -85.2 | 0.52 | 2 | Würzburg |
| 148 | -85.4 | 0.42 | 9 | Hunderttausenden |
| 79 | -85.4 | 0.47 | 4 | Youtuber |
| 25 | -85.4 | 0.56 | 1 | Stylight |
| 25 | -85.4 | 0.56 | 1 | Expo-Bühne |
| 336 | -85.5 | 0.32 | 29 | sogenannte |
| 78 | -85.5 | 0.47 | 4 | Informationsträger |
| 25 | -85.5 | 0.56 | 1 | bezahltFür |
| 25 | -85.5 | 0.56 | 1 | Filmakteure |

Filter: _____    Filter: _____

| # | Query (Poliqarp-syntax) | mio. hits |
|---|---|---|
| 1 | `[orth="das" & tt/p=PRELS]` | 6,75 |
| 2 | `[orth="das" & corenlp/p=PRELS]` | 6,49 |
| 3 | `[orth="das" & (tt/p=PRELS \| corenlp/p=PRELS)]` | 7,60 |
| 4 | `[orth="das" & (tt/p=PRELS & corenlp/p=PRELS)]` | 5,65 |

Table 3: Query results for 'das', annotated as relative pronoun by: (1) TreeTagger, (2) CoreNLP, (3) TreeTagger or CoreNLP, (4) TreeTagger and CoreNLP – with the respective number of hits in DeReKo-2016-I (Institut für Deutsche Sprache, 2016).

that aim at providing comparable corpora: (1) *The International Comparable Corpus (ICC)* (Kirk and Čermáková, 2017) and (2) *The European Reference Corpus (EuReCo)* (Kupietz et al., 2017). Both initiatives have in common that rather than building comparable corpora from scratch they try to re-use existing corpora and currently mostly rely upon national and reference corpora. However, while the ICC tries to mimic roughly the composition of the International Corpus of English (ICE) (Greenbaum, 1996), EuReCo takes a bottom up approach that relies on DeReKo's and KorAP's concept of *virtual corpora* or *virtual collections* (Kupietz et al., 2010). The key idea of the approach is to allow drawing virtual comparable corpora dynamically from tuples of the (monolingual) source corpora based on metadata categories like date of publication, genre, text type and topic domain, as well as mappings between the respective taxonomies of these metadata categories that are used in the source corpora. Currently, such a mapping is being built for DeReKo and the Reference Corpus of Contemporary Romanian Language (CoRoLa; Romanian Academy 2017, Mititelu et al. 2014) within the DRuKoLa-project (Cosma et al., 2016), and for the Hungarian National Corpus (HNC; Hungarian Academy of Sciences 2018, Oravecz et al. 2014), within the DeutUng project.[6]

## 5. Conclusion

DeReKo is a very large corpus resource for contemporary German that is continuously expanded and improved. Current developments concern new licensing opportunities, the integration of new text types, and new possibilities for querying and analyzing DeReKo via KorAP – in future also within contrastive research scenarios.

## 6. Bibliographical References

al Wadi, D. (1994). *COSMAS - Ein Computersystem für den Zugriff auf Textkorpora*. Institut für Deutsche Sprache.

Bański, P., Fischer, P. M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O., and Witt, A. (2012). The New IDS Corpus Analysis Platform: Challenges and Prospects. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul. European Language Resources Association (ELRA).

Beißwenger, M., Lüngen, H., Schallaböck, J., Weitzmann, J. H., Herold, A., Kamocki, P., Storrer, A., and Wildgans, J. (2017). Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Beißwenger, M., editor, *Empirische Erforschung internetbasierter Kommunikation*, volume 9 of *Empirische Linguistik/ Empirical Linguistics*, pages 7–46. de Gruyter, Berlin.

Belica, C., Kupietz, M., Lüngen, H., and Witt, A. (2011). The morphosyntactic annotation of DeReKo: Interpretation, opportunities and pitfalls. In Konopka, M., Kubczak, J., Mair, C., Šticha, F., and Wassner, U., editors, *Selected contributions from the conference Grammar and Corpora 2009*, pages 451–471, Tübingen. Gunter Narr Verlag.

Bodmer, F. (1996). Aspekte der Abfragekompononente von COSMAS-II. *LDV-INFO. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung*, 8:112–122.

Bodmer, F. (2005). COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3/2005:2–5.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D., and Witt, A. (2016). DRuKoLA - towards contrastive german-romanian research based on comparable corpora. Proceedings of the LREC-2016-Workshop Challenges in the Management of Large Corpora (CMLC4), pages 28 – 32, Paris. ELRA. `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-52256`.

Diewald, N. and Margaretha, E. (2016). Krill: KorAP search and analysis engine. *JLCL*, 31(1):73–90.

European Comission (2016). *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market.* `http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-593-EN-F1-1.PDF`.

Giesbrecht, E. and Evert, S. (2009). Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In Alegria, I., Leturia, I., and Sharoff, S., editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain. `http://purl.org/stefan.evert/PUB/GiesbrechtEvert2009_Tagging.pdf`.

Greenbaum, S., editor (1996). *Comparing English Worldwide: The International Corpus of English.* Clarendon Press, Oxford.

Keibel, H. and Belica, C. (2007). CCDB: A corpus-linguistic research and development workbench. In *Proceedings of the 4th Corpus Linguistics conference, Birmingham*. University of Birmingham. `http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf` (07.01.2008).

---

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Kirk, J. and Čermáková, A. (2017). From ICE to ICC: The new International Comparable Corpus. In Bański, P., Kupietz, M., Lüngen, H., Rayson, P., Biber, H., Breiteneder, E., Clematide, S., Mariani, J., Stevenson, M., and Sick, T., editors, *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`.

Kupietz, M. and Belica, C. (2013). Big language data for academic and commercial use. Presentation at the Innovation Days 2013, 4 December 2013, Berlin. `http://corpora.ids-mannheim.de/slides/Innovation-Days-2013_KupietzBelica_BigLanguageData.pdf`.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *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` (24.10.2013).

Kupietz, M. and Lüngen, H. (2014). Recent developments in dereko. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the ninth conference on international language resources and evaluation (LREC'14)*, pages 2378–2385, Reykjavik, Iceland.

Kupietz, M., Witt, A., Bański, P., Tufiş, D., Cristea, D., and Váradi, T. (2017). EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In Bański, P., Kupietz, M., Lüngen, H., Rayson, P., Biber, H., Breitenender, E., Clematide, S., Mariani, J., Stevenson, M., and Sick, T., editors, *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`.

Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Lüngen, H. (2017). DeReKo – Das Deutsche Referenzkorpus . *Zeitschrift für germanistische Linguistik*, 45(1):161–170. DOI: `https://doi.org/10.1515/zgl-2017-0008`.

Lüngen, H., Beißwenger, M., Erhardt, E., Herold, A., and Storrer, A. (2016). Integrating corpora of computer-mediated communication in clarin-d: Results from the cu-ration project chatcorpus2clarin. In *Proceedings of KONVENS 2016*, Bochumer Linguistische Arbeitsberichte, pages 156–164.

Margaretha, E. and Lüngen, H. (2014). Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics*, 29(2):59–82. `http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf`.

Meier, S. (2017). Korpora zur Fußballlinguistik – eine mehrsprachige Forschungsressource zur Sprache der Fußballberichterstattung. *Zeitschrift für germanistische Linguistik*, 45(2):345–349. DOI: `https://doi.org/10.1515/zgl-2017-0018`.

Mititelu, V. B., Irimia, E., and Tufiş, D. (2014). CoRoLa The Reference Corpus of Contemporary Romanian Language. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. ELRA.

Oravecz, C., Váradi, T., and Sass, B. (2014). The Hungarian gigaword corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. ELRA.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent Developments in the National Corpus of Polish. In Calzolari, N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Referentenentwurf (2017). *Entwurf eines Gesetzes zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissensgesellschafts-Gesetz – UrhWissG)*. Bundesministerium der Justiz und für Verbraucherschutz, Berlin. `http://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_UrhWissG.pdf;jsessionid=0C34528B94659CE7CF6BA15B7B3516E5.1_cid297?__blob=publicationFile&v=1`.

Rosenfeld, V. (2010). An implementation of the Annis 2 query language. Technical report, Humboldt-Universität zu Berlin.

Schröck, J. and Lüngen, H. (2015). Building and Annotating a Corpus of German-Language Newsgroups. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015)*, pages 17–22, Essen. `https://sites.google.com/site/nlp4cmc2015/program`.

UrhWissG (2017). Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissensgesellschaftsgesetz). *Bundesgesetzblatt*, Jahrgang 2017 Teil I Nr. 61:3346–3351. `http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl117s3346.pdf`.

## 7. Language Resource References

Hungarian Academy of Sciences (2018). Hungarian National Corpus.

Institut für Deutsche Sprache (2016). German Reference Corpus DeReKo. Deutsches Referenzkorpus, DeReKo-2016-I.

Institut für Deutsche Sprache (2018). German Reference Corpus DeReKo. Deutsches Referenzkorpus, DeReKo-2018-I. PID: `http://hdl.handle.net/10932/00-03B9-3C6A-6F80-6601-A`.

Romanian Academy (2017). Reference Corpus of Contemporary Romanian Language. Romanian Academy, CoRoLa.