Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 878–884
Marseille, 11–16 May 2020

The DReaM Corpus: A Multilingual Annotated Corpus of Grammars for the World's Languages

Shafqat Mumtaz Virk[1], **Harald Hammarström**[2], **Markus Forsberg**[1], **Søren Wichmann**[3]
[1]Språkbanken Text, Department of Swedish, University of Gothenburg
[2]Department of Linguistics and Philology, University of Uppsala
[3]Leiden University Centre for Linguistics, Leiden University
[3]Laboratory of Quantitative Linguistics, Kazan Federal University
[3]Beijing Advanced Innovation Center for Language Resources, Beijing Language University
[1]{shafqat.virk, markus.forsberg}@gu.se
[2]{harald.hammarstrom}@lingfil.uu.se
[3]{wichmannsoeren}@gmail.com

Abstract

There exist as many as 7000 natural languages in the world, and a huge number of documents describing those languages have been produced over the years. Most of those documents are in paper format. Any attempts to use modern computational techniques and tools to process those documents will require them to be digitized first. In this paper, we report a multilingual digitized version of thousands of such documents searchable through some well-established corpus infrastructures. The corpus is annotated with various meta, word, and text level attributes to make searching and analysis easier and more useful.

Keywords: corpus, natural languages, grammatical descriptions, world's languages

1. Introduction

The diversity of the 7000 languages of the world represents an irreplaceable and abundant resource for understanding the unique communication system of our species (Evans and Levinson, 2009). All comparison and analysis of languages departs from *language descriptions* — publications that contain facts about particular languages. The typical examples of this genre are grammars and dictionaries (Hammarström and Nordhoff, 2011).

Until recently, language descriptions were available in paper form only, with indexes as the only search aid. In the present era, digitization and language technology promise broader perspectives for readers of language descriptions. The first generation of enhanced search tools allow searching across many documents using basic markup and filters, and modern natural language processing (NLP) tools can take exploitation arbitrarily further. In this paper we describe the collection, digitization, management and search infrastructure so far developed for a comprehensive collection of language descriptions.

The paper is organized as follows: Section 2. describes the collection and digitization process, while the statistics of the corpus are given in Section 3. The methods applied to do post-OCR corrections are explained in Section 4. Section 5., briefly describes the two corpus infrastructures to be followed by Section 6., which explains how the corpus can be accessed using those two infrastructures.

2. Collection and digitization

Enumerating the extant set of language descriptions for the languages of the world is a non-trivial task. Thanks to the Glottolog project, this task is now complete in the sense that the most extensive description for *every* language is known (Hammarström and Nordhoff, 2011). These references, along with a large body for further entries for most languages, are included in the open-access bibliography of (Hammarström et al., 2019).

A core subset of about 30,000 publications — including the most extensive description for 99% of the worlds languages — has been digitized or obtained in born-digital form for the present project. Each item has been manually annotated with the language(s) described in it (the object-language), the language used to describe it (the meta-language), the number of pages and its type (e.g., grammar, dictionary, phonology, sociolinguistic study, overview etc.). The set of digital documents has been subjected to optical character recognition (OCR) to recognize the meta-language. For approximately 1% of the documents, OCR was not possible (poor quality, handwriting, script not available for OCR and similar reasons).

3. Corpus statistics

Table 1 shows the number of documents collected and digitized. For space reasons, only the top 10 (with respect to the number of documents in the collection) document types (e.g. grammatical description, word list, dictionary, etc.), and meta languages are listed in the table. Table 2 shows statistics about the type and number of documents per natural language. Again, only the top 10 languages with respect to the number of documents and document types are shown.

As mentioned previously, for each document, a BibTex entry with fields for various text-level attributes (i.e. title of the document, author, publisher, publishing year, language code, etc.) is maintained. An example BibTex entry is shown below:

```
@book{g:Lichtenberk:Manam,
author = {Frantisek Lichtenberk},
title = {A Grammar of Manam},
publisher = {Honolulu: University
```

Type/Language	English	French	German	Spanish	Portuguese	Russian	Indonesian	Dutch	Italian	Chinese	Total
overview	5949	830	748	477	182	83	106	43	52	11	8481
grammar_sketch	4542	970	591	610	157	314	93	322	116	41	7756
comparative	4588	471	464	264	103	73	37	26	34	10	6070
grammar	3199	773	350	284	96	218	49	8	52	21	5050
ethnographic	2881	477	713	347	247	10	207	13	53	2	4950
wordlist	2664	404	448	415	140	17	109	43	40	13	4293
dictionary	1379	402	150	378	95	127	61	43	55	16	2706
specific_feature	2122	206	83	129	56	31	10	47	8	1	2693
phonology	1146	288	44	94	97	8	5	36	6	6	1730
minimal	1179	132	141	120	34	15	22	25	29	5	1702
Total	29649	4953	3732	3118	1207	896	699	606	445	126	45431

Table 1: Number of documents by type and meta-language.

Type/Language	Swahili [swh]	Hausa [hau]	Namibia [naq]	Mongo [lol]	English [eng]	Koongo [kng]	Ewe Anglo [ewe]	Semai [sea]	Zande [zne]	Hadza [hts]	Total Total
overview	50	35	40	17	3	25	24	30	29	28	281
wordlist	46	16	8	6	2	7	8	29	19	8	149
comparative	41	24	22	7	1	11	16	31	7	15	175
ethnographic	4	5	17	19	0	7	7	4	20	24	107
grammar_sketch	16	11	15	41	59	10	15	1	6	3	177
grammar	16	13	6	7	17	13	15	0	3	0	90
minimal	1	3	2	1	4	3	2	1	2	2	21
specific_feature	5	2	3	4	5	7	3	0	5	2	36
dictionary	7	5	1	8	3	9	6	2	3	0	44
socling	7	4	1	0	15	1	2	1	0	3	34
Total	193	118	115	110	109	93	98	99	94	75	1114

Table 2: Collections Statistics for the top 10 (object-)languages with the most documents digitized.

```
            of Hawaii Press},
series = {Oceanic Linguistics
         Special Publication},
volume = {18},
pages = {xxiii+647},
year = {1983},
glottolog_ref_id = {55327},
hhtype = {grammar},
inlg = {English [eng]},
isbn = {9780824807641},
lgcode = {Manam [mva]},
macro_area = {Papua}
}
```

We selected the subset of documents providing grammatical description, i.e., of the types 'grammar_sketch', 'grammar' and 'specific_feature', for corpus infrastructure support (detailed in Section 5.). The remaining types, e.g., word list and dictionary, are not primarily prose descriptions. For the future, we have plans to use Karp (another infrastructure tool developed by Språkbanken) for storing and exploring lexical data types such as dictionaries, wordlists, etc.

This set was further divided into two subsets (English and non-English): one with documents written in English and the other in the remainder of the meta-languages. The first set is to be annotated with various word level (POS tag, lemma, etc.), text-level (document title, author, production year, etc) annotations, and syntactic parsing. The other subset is to have only text-level annotations in addition to POS tagging, but no syntactic parsing. This is mainly because of the unavailability of appropriate annotation and parsing tools for languages other than English.

According to the copyright status of individual document, each of the two subsets (English and non-English) were further divided into open and restricted sets. The former consists of all those documents that are at least a century old and/or do not have any copyrights, while the latter set contains documents which have copyrights and can not be released as in an open-access corpus. Table 3 shows some statistics of the both the open and restricted parts of the corpus language wise. The open-access subset is being released together with this paper, and all the search examples shown in the next sections are limited to this set, while the other set is to be used only for the internal research purposes.

4. Post-OCR Corrections

Even though there has been a lot of progress in the area of OCR (a survey of available tools and techniques can be found in (Islam et al., 2016)), the available techniques and tools are expected to fail at times and make errors. This is true, especially, if the image quality is poor, the document is very old, or it has a complex page structure. In our collection, there are many documents which are more than a century old, meaning that they were not digitally born, and hence, the OCR'd version is expected to have errors.

The field of post-OCR corrections deals with the corrections of OCR errors, and a number of techniques have been proposed for this purpose (a survey of those techniques and subsequent work can be found in (Niklas, 2010; Reffle and Ringlstetter, 2013)). More recently, a deep learning based approach was introduced by (Mokhtar et al., 2018), which claims to outperform previously reported state-of-the-art results significantly. Since this system is not made available, we used a simple and readily available system (Hammarström et al., 2017) for post OCR corrections of the corpus. The system we used is lightweight in the sense that it does not require any manual labeling, training of models, and tuning of parameters. Rather, it is based on a simple idea of observing the presence of words, which are similar in form and are also distributionally more similar than expected. Such words are deemed OCR variants and hence corrected. The evaluation results show that this simple technique can capture and correct many OCR errors, although the accuracy is lower than the state-of-the-art. The language and genre independence of the system makes it suitable for us, hence, we used it for the post-OCR corrections of our data-set.

5. The corpus infrastructure

Recent years have seen a dramatic increase in the production of digital textual data (i.e. corpora), and conversion from non-digital to digital textual form. This has, in parallel, necessitated the development of efficient ways of storing and exploring those large volumes. As a consequence, the technology has improved from simple string-matching search approaches to the development of corpus infrastructures with advanced query based search, comparison, and visualization options. The following sections briefly introduce two such corpus infrastructure tools: Korp and Strix. These tool provide various options to explore, compare, and visualize the corpus and related statistics at the sentence and document levels respectively.

5.1. Korp

Korp[1] (Borin et al., 2012) is a system in the corpus infrastructure developed and maintained at Språkbanken[2] (the Swedish language bank). It has separate backend and frontend components to store and explore a corpus. The backend is used to import the data into the infrastructure, annotate it, and export it to various other formats for downloading. For the annotations, it has an annotation pipeline that can be used to add various lexical, syntactic, and semantic annotations to the corpus using internal as well as external annotation tools. The frontend provides basic, extended, and advanced search options to extract and visualize the search hits, annotations, statistics, and more. Some examples are given in Section 6.

5.2. Strix

Strix[3] is another system in Språkbanken's corpus infrastructure. It is similar to Korp in that it allows for search and exploration of a text collection and its annotations, but it differs – hence complements Korp – in that a search hit in Strix is a document but not an occurrence. Some additional

[1] https://spraakbanken.gu.se/korp/#?stats_reduce=word

[2] https://spraakbanken.gu.se/

[3] https://spraakbanken.gu.se/strix

	Open-Access			Restricted		
Meta-Language	# Documents	# Sentences	# Words	# Documents	# Sentences	# Words
English	462	2208184	28468332	8757	25434214	558540669
German	270	1482053	18622901	463	2072342	35773389
French	176	859140	11235961	1244	4287114	94478231
Spanish	128	709255	9065547	744	1730437	37100519
Dutch	45	270688	4654236	104	371844	6759210
Italian	30	166935	1883211	100	357792	7157718
Russian	15	56227	990637	441	1647647	37401127

Table 3: Statistics on the number of documents, sentences and word tokens in the corpus, organized by meta-language.

examples of differences are that Strix has support for metadata filtering and text similarity and it provides a reading mode with annotation highlighting.

6. The corpus infrastructure in use

This section contains a detailed description of the process that we followed to annotate the open-access part of the data and make it available through Korp and Strix. As mentioned in the previous section, Språkbanken's corpus infrastructure has a pipeline architecture to annotate the data. Using that pipeline we have annotated the English data with the following lexical, syntactic, and text-level attributes. The non-English subset was annotated with only some lexical and text-level attributes for the reason mentioned previously.

- Lexical Annotations
  - Part of Speech (POS) tag
  - Lemma
- Structural Annotations
  - Dependency Parse
- Text-Level Annotations
  - Title of the Document
  - Year of Production
  - Type of Document (e.g. grammar, overview, specific feature, etc.)
  - Language Code

The text level annotations were taken from the corresponding BibTex entry (shown in the previous section). Most of the fields from those BibTex entries have been imported as text level attributes to Korp, and hence can be used for filtering and searching through the corpus (as will be shown later in this section). For other lexical and syntactical annotations, we have used the Stanford Dependency parser for English, which is part of the Stanford's CoreNLP toolkit (Manning et al., 2014). Only sentence and word segmentation were done for the non-English data. Figure 1 shows a screenshot of Korp frontend.

It shows the hits of a basic free-text search, when searched for the string 'tone'. The left-hand pane shows the sentences retrieved from all documents in the corpus which contained the string 'tone', while the right hand pane shows the text-level as well as the word-level attributes of the selected word (i.e. 'tone' highlighted with black background). A simple, yet very useful use-case of such a search could be to retrieve all sentences from all the documents in the corpus which contains the term 'tone', and then analyze them in a quest to know which languages do or do not have tones.
The search can be restricted (or expanded) to various word and text level attributes using the 'Extended' search tab (e.g. search only through a single document, search for a particular POS, or any combination of the attributes, etc.) Figure 2 shows the results of restricting the search to only one document by its title i.e. 'A progressive grammar of the Telugu language', and search for the word 'tone' again. This time, only a couple of sentences from the searched document are found and returned as shown. Further, to meet other particular needs the front end also provides 'Advanced' search option, where a search query could be designed using the CQP query language (Christ, 1994). Apart from this, there are many other interesting features provided by the frontend (e.g. displaying the context i.e. a few sentences before or after the searched term), which could come handy while exploring the corpus. Due to the space limitations, it is not possible to explain all features of Korp here, so we refer the reader to visit the `https://spraakbanken.gu.se/korp/?mode=dream` to explore the corpus and experience various search options.

As can be noticed from the given screenshots, in Korp each search hit is restricted to be 'a sentence' (or a few sentences if the context visualization is turned on). An alternative is to return the documents containing the searched terms as search hits (as opposed to sentences), and then provide an option to view the full document in reading mode. This is exactly, what the Strix is designed for (as already mentioned in the previous section). If we search for the term 'tone' through the Strix interface, a list of documents from the collection containing the search term will be displayed as shown in Figure 3.

There are 78 documents in our open-access collection which contained the term 'tone'. This list can be filtered

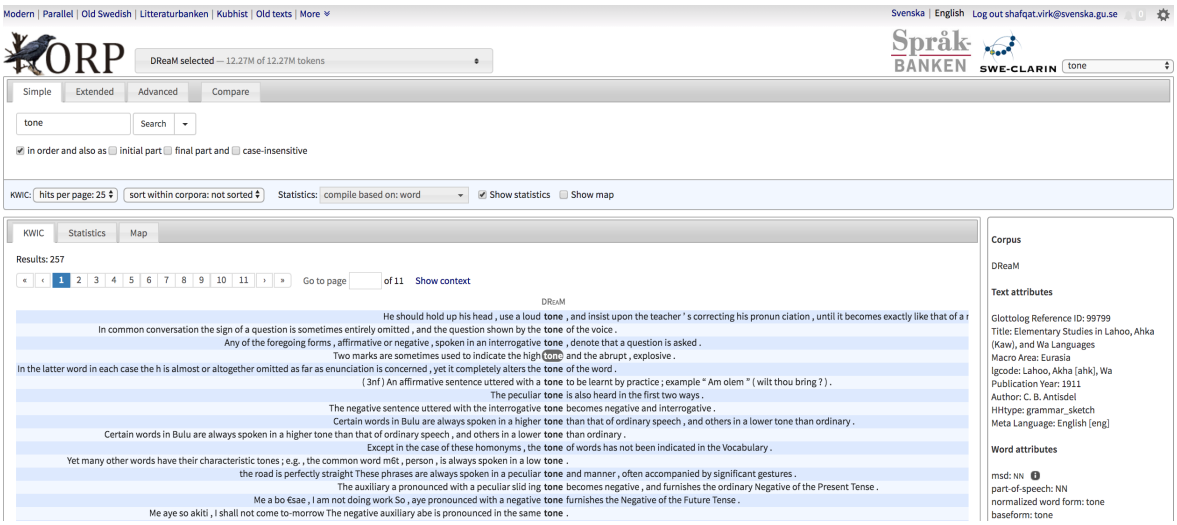

Figure 1: Screenshot of Korp frontend 'Basic' search

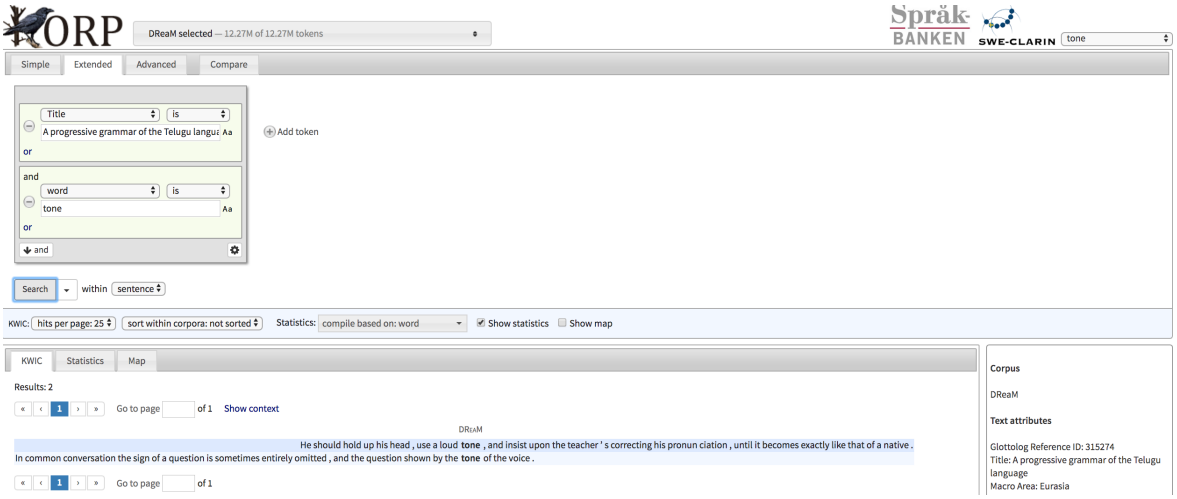

Figure 2: Screenshot of Korp frontend 'Extended' search

further based on various text-level attributes (e.g. author, document type etc.) using the given metadata filtering options in the left-hand side pane.
Clicking on any document will open the full document in text mode as shown in Figure 4.

Further, a list of related documents (based on a separately computed semantic relatedness measure) is displayed in the left hand side pane, while various text and word-level attributes of the selected text are displayed in the right hand side. Also note that the selected document can be further searched using the 'Search the current document' search box on top. Again due to the space limitations, it is not possible to explain all searching and exploring options provided by Strix, and we refer the reader to Språkbanken for further details.

6.1. Resource URL's

The following url can be used to access the open-access part of the data through Korp:

`https://spraakbanken.gu.se/korp/?mode=dream`

Once opened, a particular corpus can be selected using the 'corpora tag' before making the search as shown in Figure 5.

First in the list is the English corpora (labeled 'DReaM') followed by the German (DReaM-de-open), Spanish (DReaM-es-open), French (DReaM-fr-open), Italian (DReaM-it-open), Dutch (DReaM-nl-open), and Russian (DReaM-ru-open).
And the following is the url to access the DReaM data through Strix:

`https://spraakbanken.gu.se/strix/?lang=eng`

Use the filter on the left to select one or more corpora from the list: 'DReaM-English', 'DReaM-German', 'DReaM-Spanish', 'DReaM-French', 'DReaM-Dutch', 'DReaM-Italian', 'DReaM-Russian'.

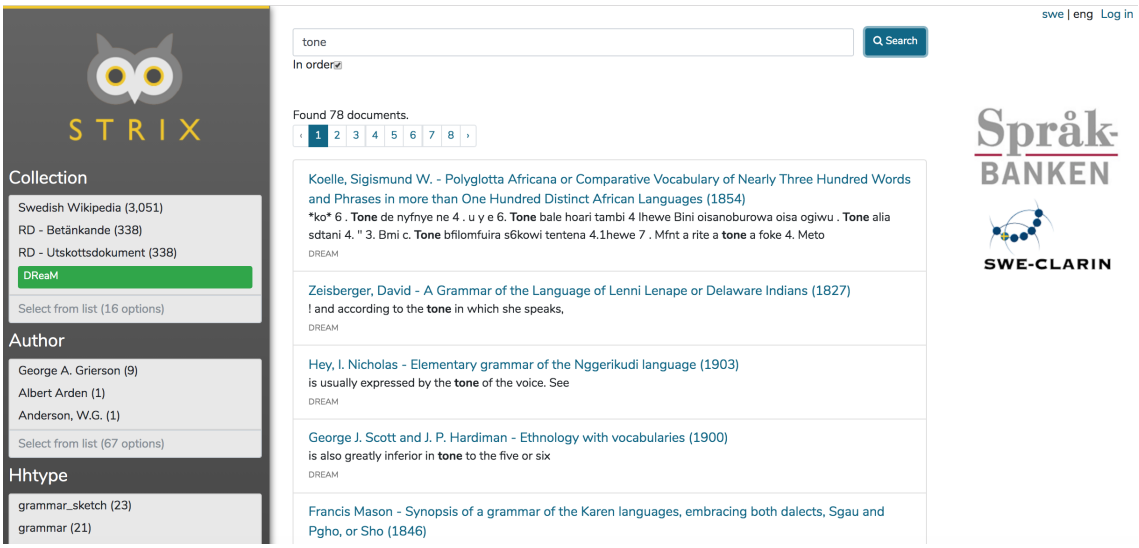

Figure 3: Screenshot of Strix

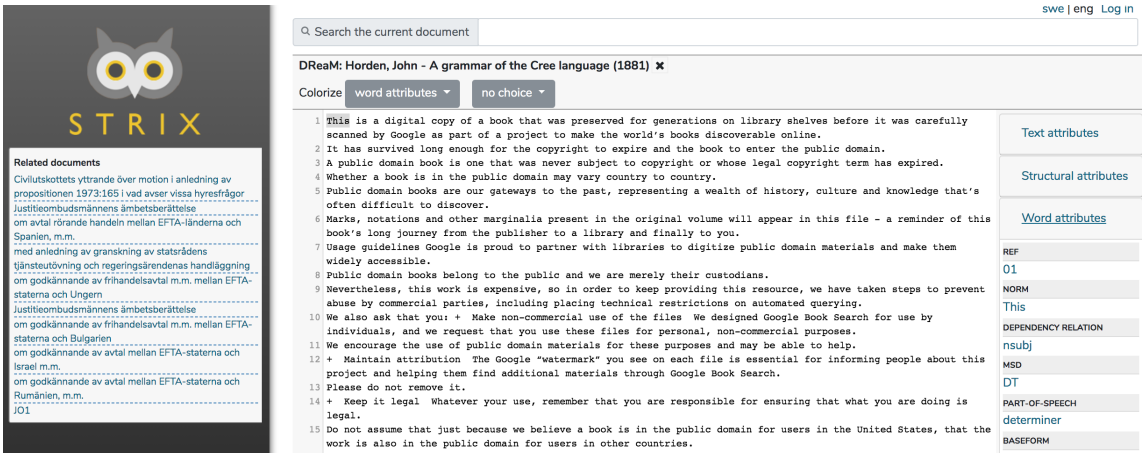

Figure 4: Strix Document View

7. Conclusions

Descriptive linguistic documents contain within them very valuable knowledge about world's natural languages and their characteristics, which in turn contains keys to many unanswered questions concerning limits on human communication, human prehistorical population movements, and cultural encounters. We have collected, scanned, digitized a large size multilingual corpus of the world's language descriptions, and have made them explorable through a couple of corpus infrastructures. We have also annotated the data with text-level as well token level attributes to make the searching, filtering, and exploration much easier and useful. We believe such a collection is a useful resource for deeper analysis of world's natural languages to find answers to some of the above raised questions.

8. Acknowledgements

The work presented here was funded by (1) the *Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage* (DReaM) Project awarded 2018–2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden, (2) the *From Dust to Dawn: Multilingual Grammar Extraction from Grammars* project funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2007.0105, Uppsala University, and (3) the University of Gothenburg, its Faculty of Arts and its Department of Swedish, through their truly long-term support for the Språkbanken research infrastructure.

9. References

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp — the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 474–478, Istanbul, Turkey, May. European Languages Resources Association (ELRA).

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *ArXiv*, abs/cmp-lg/9408005.

Evans, N. and Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492.

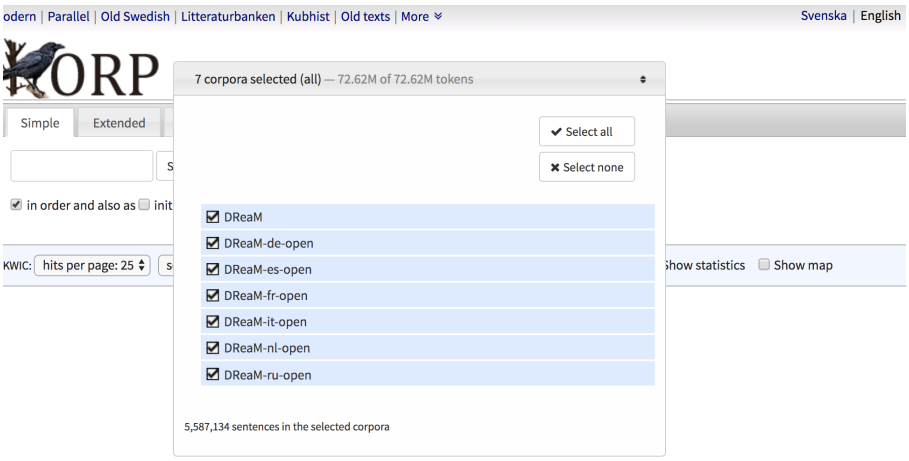

Figure 5: List of available corpora through Korp

Hammarström, H. and Nordhoff, S. (2011). Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language*, 3(2):31–43.

Hammarström, H., Virk, S. M., and Forsberg, M. (2017). Poor man's ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *DATeCH*.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). Glottolog 4.0. Jena: Max Planck Institute for the Science of Human History. Available at `http://glottolog.org`. Accessed on 2019-09-12.

Islam, N., Islam, Z., and Noor, N. (2016). A survey on optical character recognition system. *ITB Journal of Information and Communication Technology*, 12.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014*, pages 55–60, Baltimore. ACL.

Mokhtar, K., Bukhari, S., and Dengel, A. (2018). Ocr error correction: State-of-the-art vs an nmt-based approach. In *The 13th IAPR Workshop on Document Analysis Systems, DAS18, Vienna Austria, 2018*, pages 429–434, 04.

Niklas, K. (2010). Unsupervised post-correction of ocr errors. Master's thesis, Leibniz Universität Hannover Fakultät für Elektrotechnik und Informatik Institut für verteilte Systeme Fachgebiet Wissensbasierte Systeme Forschungszentrum L3S.

Reffle, U. and Ringlstetter, C. (2013). Unsupervised profiling of ocred historical documents. *Pattern Recogn.*, 46(5):1346–1357, May.