CURLICAT

Curated Multilingual Language Resources for CEF.AT

Agreement number: INEA/CEF/ICT/A2019/1926831

Action No: 2019-EU-IA-0034

CEF programme

Deliverable 2

Extending multilingual corpora and ensuring IPR clearance

Version 1.0

2022-08-31

Document Information

Activity:	Activity 2: Additional collection and IPR clearance
Deliverable number:	D2
Deliverable title:	Extending multilingual corpora and ensuring IPR clearance
Indicative submission date:	2022-06-30
Actual submission date of deliverable:	2022-08-31
Main Author(s):	Radovan Garabík
Participants:	Bence Nyéki, Elena Irimia, Svetla Koeva, Simon Krek, Andraž Repar, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Vanja Štefanec, Marko Tadić
Version:	v1.0

History of versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.1	2021-11-19	Completed	JÚĽŠ SAV	Bence Nyéki, Elena Irimia, Svetla Koeva, Simon Krek, Andraž Repar, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Marko Tadić	
V1.0	2022-08-29	Completed	JÚĽŠ SAV	Bence Nyéki, Elena Irimia, Svetla Koeva, Simon Krek, Andraž Repar, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Vanja Štefanec, Marko Tadić	

EXECUTIVE SUMMARY

This deliverable provides a documentation of the domain distribution of text data in the CURLICAT corpus for seven consortium languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovene. A detailed analysis of metadata values for each of the subcorpora is provided together with the description of newly acquired data and its IPR status and a description of the metadata and domain distribution of the second version of the corpus.

Table of contents

1. Introduction

The Activity was successfully completed by all partners, the subcorpora have been extended by newly acquired texts, meeting the contractually stipulated sizes. The metadata have been updated and the data validated and improved (where applicable). The second version of the corpus consisting of the following datasets (Table 1):

Dataset language	Size in sentences	Size in tokens
Bulgarian	3 164 359	89 902 991
Croatian	1 720 652	40 311 047
Hungarian	2 815 314	60 245 090
Polish	2 421 154	59 301 782
Romanian	3 557 812	94 925 454
Slovak	4 805 678	66 891 145
Slovenian	2 003 626	43 481 563

Table 1. Overview of the corpora.

Detailed analysis of the domains present in the subcorpora is provided for each of the languages in the corpus. The description of the metadata has been used in Activity 5, Metadata harmonisation, and metadata fields in the second version of the corpus are harmonized among the subcorpora, described in detail in Deliverable 5.1 *Metadata Harmonisation in CURLICAT*.

2. The Bulgarian corpus

The **Bulgarian CURLICAT corpus** consists of texts from different sources, provided with appropriate licences for distribution. We used three general types of sources with regard to the metadata extraction:

- [1] Sources with very rich metadata structure, such as the Bulgarian National Corpus (provided that they have redistributable licencing terms);
- [2] Sources with a shallow metadata structure, such as some public repositories with open and copyright free data;
- [3] Sources with no metadata structure but with extractable metadata values, such as blogs with redistributable licenses, open content websites, etc.

Concerning the sources from which the documents were obtained and the way of providing the free distribution (either by the licence attached to the document or by an agreement with the copyright holder) we can distinguish four groups:

- [1] Documents form the Bulgarian National Corpus (provided that they have redistributable licences);
- [2] Documents additionally obtained from the internet (if they satisfy conditions for copyright and domain affiliation);
- [3] Documents additionally obtained after an agreement with their providers (if they satisfy conditions for copyright and domain affiliation);
- [4] Documents from some open domain repositories (provided that they have redistributable licences).

2.1. IPR overview

For the CURLICAT corpus, we selected documents under redistributable licences such as: Universal Public Domain Dedication (CC0 1.0); Creative Commons Attribution 4.0 International (CC BY 4.0), Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0), Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0), Creative Commons Attribution-ShareAlike 3.0 (CC BY-SA 3.0), Thus, we are avoiding copyright encumbered material, which might limit the use of our dataset only for academic purposes.

We have chosen documents from government websites (ministries, government agencies, municipalities, European Union bodies and institutions), academic websites and their repositories granting access to full-text journal articles, other research articles, monographs, books, dissertations; the Bulgarian Portal for Open Science repository, media and NGOs (blogs, media, political bodies) websites which grant access to texts which are released under Creative Commons licences, including: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0), Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0), Attribution-NonCommercial- -ShareAlike 2.5 Bulgaria (CC BY-NC-SA 2.5 BG), Attribution 3.0 Unported (CC BY 3.0), Attribution-ShareAlike 3.0 IGO (CC BY-SA 3.0 IGO).

2.2. Additional data acquisition

As for some domains the documents from the Bulgarian National Corpus were insufficient in number of sentences, we had to find additional sources of data. IPR clearance was successful with several providers such as University of National and World Economics and some other blog publishers. One university did not refuse but postponed the agreement and we expect some more documents to be added.

In order to ensure sufficient size of the corpus of redistributable documents we identified several new sources: libraries of scientific texts (books and PhD theses) and several other websites and blogs providing texts from required thematic domains and copyright.

One of the resources for new documents is the Bulgarian Portal of Open Science: a platform providing open access to full scientific texts. The portal allows filtering by different criteria, among with we used language (to select documents in Bulgarian), type of access (open access), type of the documents (to exclude such text as scientific report, reviews, etc.), and scientific domain (to select CURLICAT domains). We had to filter open access documents for which appropriate CCS licences were not specified.

To assure the correct document classification according to the CURLICAT domains, an extensive process of automatic and manual validation was performed (see Deliverable 5.1 *Metadata Harmonisation in CURLICAT*, Section 1.3.2).

2.3. Domain distribution analysis

The following table shows the distribution of domains in the current version of the Bulgarian CURLICAT corpus after collecting new documents and cleaning the data.

Domain	Source	Number of sentences	Number of tokens	% from the Corpus
Culture	[1], [2], [3]	197446	4706160	5.23
Economics	[1], [2], [3], [4]	171864	5631287	6.26
Education	[1], [2], [3], [4]	123818	4474074	4.98
European Union	[1], [2], [3], [4]	4270	157275	0.17
Energy	[1], [2], [4]	209458	6980827	7.76
Finance	[1], [2], [3], [4]	259003	10006469	11.13
Health	[1], [2]	41539	1049916	1.17
Industry	[1], [2], [3], [4]	20938	657003	0.73
Politics	[1], [2], [3], [4]	577697	16527181	18.38
Science	[1], [2], [3], [4]	1554064	39565239	44.01
Trade	[1], [2], [3], [4]	4262	147560	0.16
Total	[1], [2], [3], [4]	3 164 359	89 902 991	100

Table 2.3.1. Distribution of domains in the current version of the Bulgarian CURLICAT corpus.

Source	Number of sentences	Number of words	% from the Corpus
[1]	2 615 138	68 501 283	76.2
[2]	339 952	12 602 458	14.02
[3]	77 939	2 588 738	2.88
[4]	131 330	6 210 512	6.9
Total	3 164 359	89 902 991	100

Table 2.3.2. Distribution of sources in the current version of the Bulgarian CURLICAT corpus.

The corpus meets the CURLICAT requirement concerning the total number of sentences (at least 2M per language) but some domains are still underrepresented compared to the other domains (especially to the domain Science). This issue may be resolved by collecting further data and discarding some of the texts initially selected among the documents at the Bulgarian National Corpus.

The corpus consists of 116634 documents. The shortest document is 51 tokens long; the longest one 423701 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 770.81 tokens, the median is 339.0 tokens and the standard deviation is 3722.32.

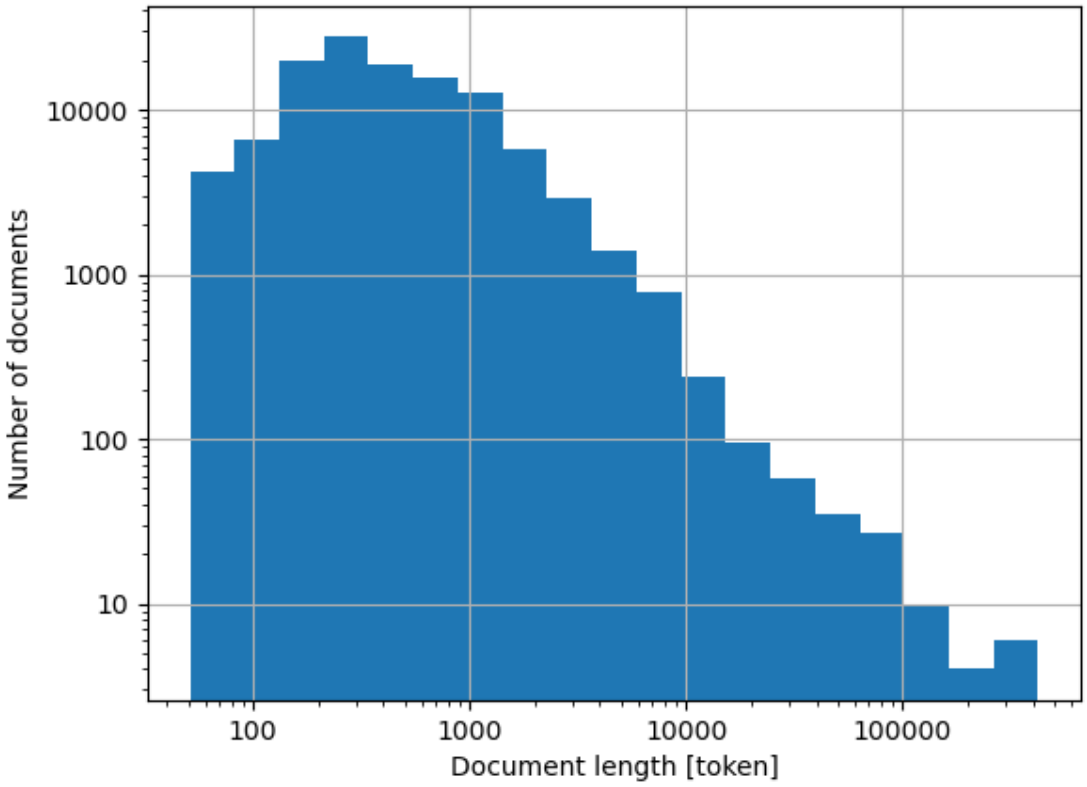


Figure 2.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 1 sentence long; the longest one 16928 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 27.13 sentences, the median is 13.0 sentences and the standard deviation is 111.83.

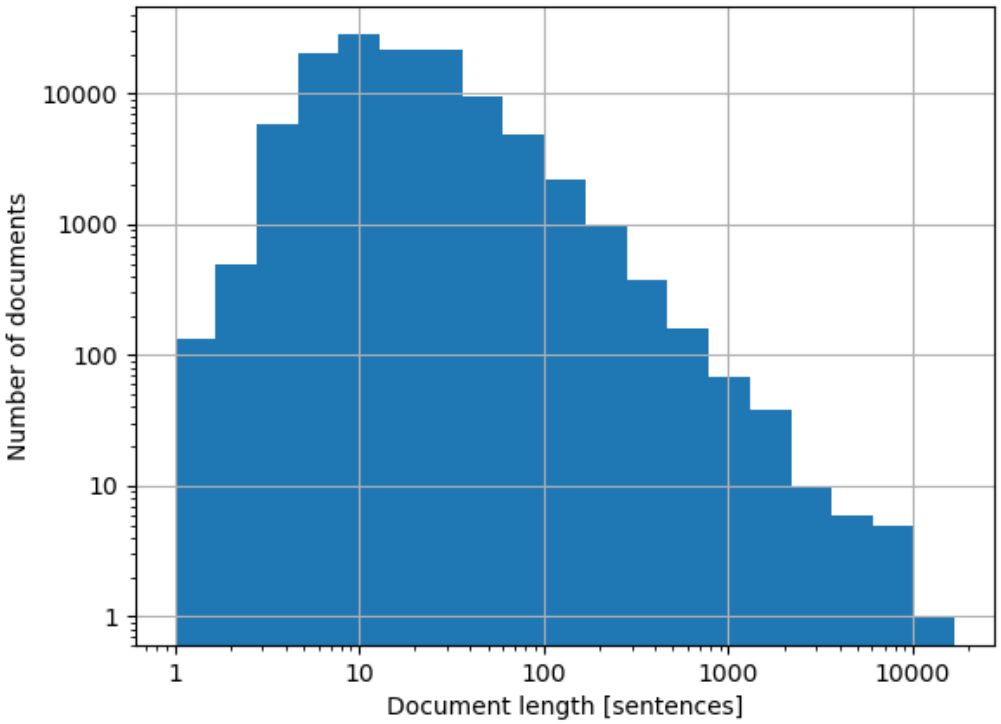


Figure 2.4.2: Distribution of document lengths in sentences, log-log axes.

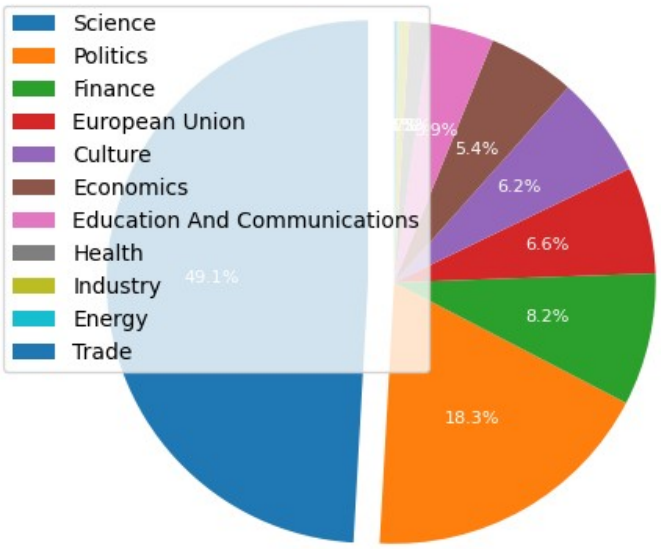


Figure 2.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	4706160	197446	5.23	6.24
Economics	5631287	171864	6.26	5.43
Education And Communications	4474074	123818	4.98	3.91
Energy	157275	4270	0.17	0.13
European Union	6980827	209458	7.76	6.62
Finance	10006469	259003	11.13	8.19
Health	1049916	41539	1.17	1.31
Industry	657003	20938	0.73	0.66
Politics	16527181	577697	18.38	18.26
Science	39565239	1554064	44.01	49.11
Trade	147560	4262	0.16	0.13

Table 2.4.1: Distribution of domains.

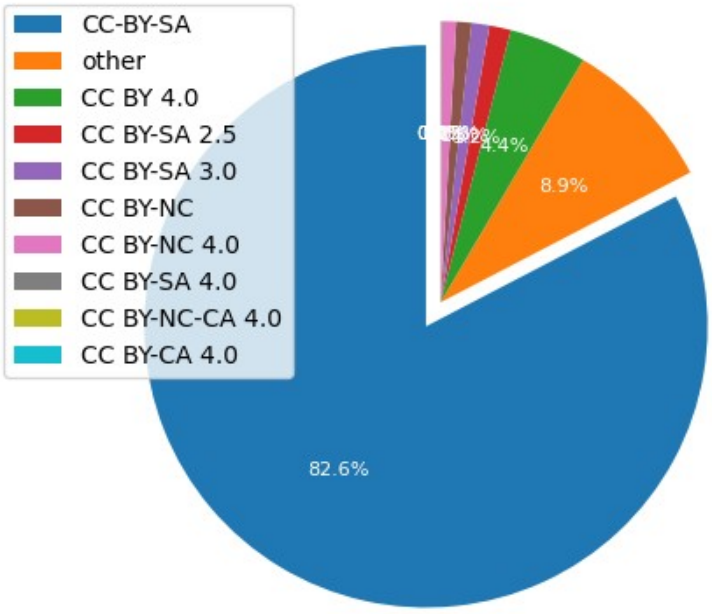


Figure 2.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY 4.0	140385	4.44
CC BY-NC	27672	0.87
CC BY-NC 4.0	25626	0.81
CC BY-NC-CA 4.0	800	0.03
CC BY-SA 2.5	38826	1.23
CC BY-SA 3.0	32428	1.02
CC BY-SA 4.0	1238	0.04
CC BY-CA 4.0	403	0.01
CC-BY-SA	2615138	82.64
other	281843	8.91

Table 2.4.2: Distribution of licenses.

2.5. Data delivery

The Bulgarian CURLICAT corpus is not publicly available at the moment. The delivery is planned as soon as some of the selected documents are supplied with an official permission for redistribution by their owners (for some of which we still are waiting for) .

3. The Croatian corpus

3.1. IPR overview

Since the text from which the Croatian National Corpus (HNK) has been compiled were not available in their full extent to any users apart from the online concording interface, the IPR had to be sorted out. For the largest part of the culture domain, the negotiations with the publisher of the newspaper "Vijenac" are on their way, but since it is a nationally funded main cultural institution – Matica hrvatska (Matrix Croatica) – we expect their full understanding. Texts from the Vjesnik newspaper, published by the state-owned publisher, covering domains of both culture and economy, we consider as falling under a PSI Directive, and thus not subject to a non-permissive licence.

Besides that, additional contracts with publishers of two other newspapers (Dubrovački vjesnik and Glas Slavonije), covering also domains of culture and economy, will have to be signed to explicitly allow redistribution of texts.

3.2. Additional data acquisition

As for other domains texts from HNK were either IPR protected or of insufficient cumulative size, we had to find additional sources of data.

For the health domain we first included texts from the MedCorA corpus (Kocijan et al. 2020) that consists overall of 2,232 texts with 71,911,667 tokens in total. Since this corpus is composed from the publicly available data from the Agency for Medicinal Products and Medical Devices of Croatia (HALMED) which comprise only of the instructions for drug usage, we selected only 3,113,034 tokens from this source in order to preserve the balance of different topics in this domain and avoid the monotony of texts from that source. Additional data in this domain were selected from the papers in the scientific field of biomedicine (see the description of the source below).

As a third source of data, we used the central portal of Croatian scientific journals – Hrčak (http://hrcak.srce.hr). Hrčak offers access to the journals following the Open Access Initiative and covers 510 journals, with 244,159 articles in full text from all scientific fields. However, as the primary source of texts we considered only 172,165 articles published in Croatian. The papers from Hrčak, covering the domains of science, health and economy, were downloaded in the PDF format, converted into plain text, manually checked and cleaned where the unnecessary parts of documents were discarded. The texts were then processed with an usual pipeline for processing Croatian.

3.3. Domain distribution analysis

The corpus consists of 43563 documents. The shortest document is 5 tokens long; the longest one 151561 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 925.35 tokens, the median is 494 tokens and the standard deviation is 1793.36.

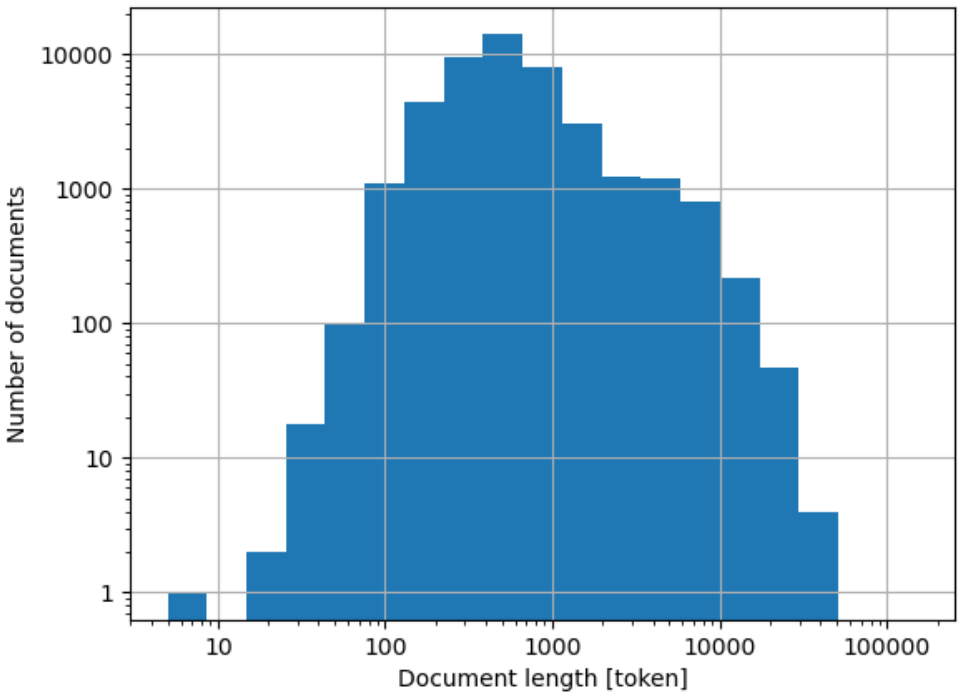


Figure 3.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 1 sentence long; the longest one 3590 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 39.50 sentences, the median is 22 sentences and the standard deviation is 62.38.

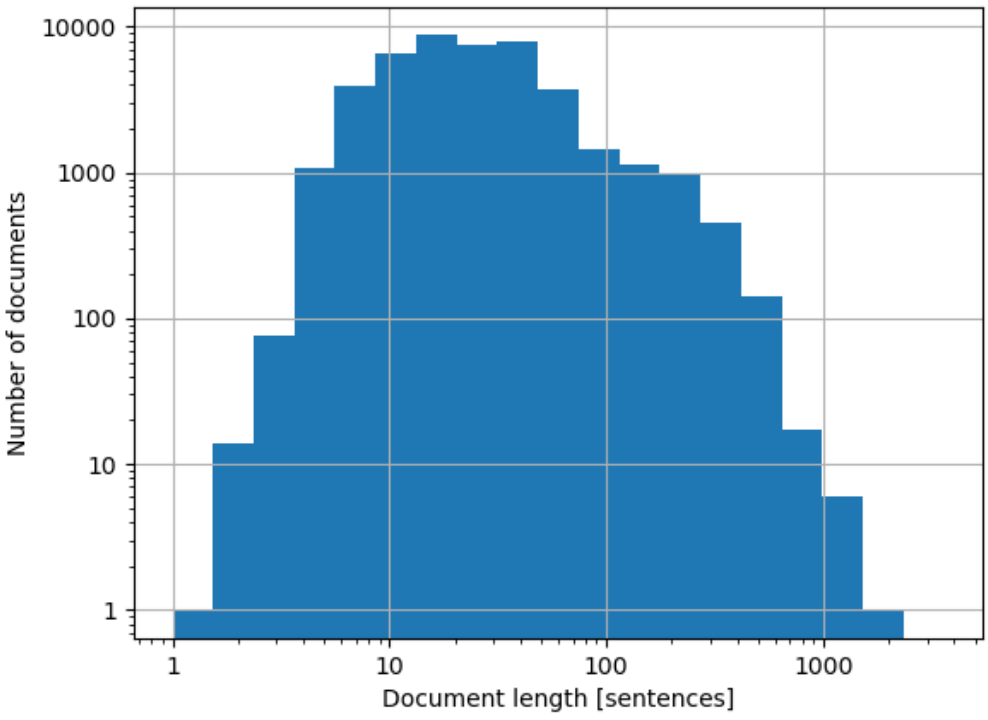


Figure 3.4.2: Distribution of document lengths in sentences, log-log axes.

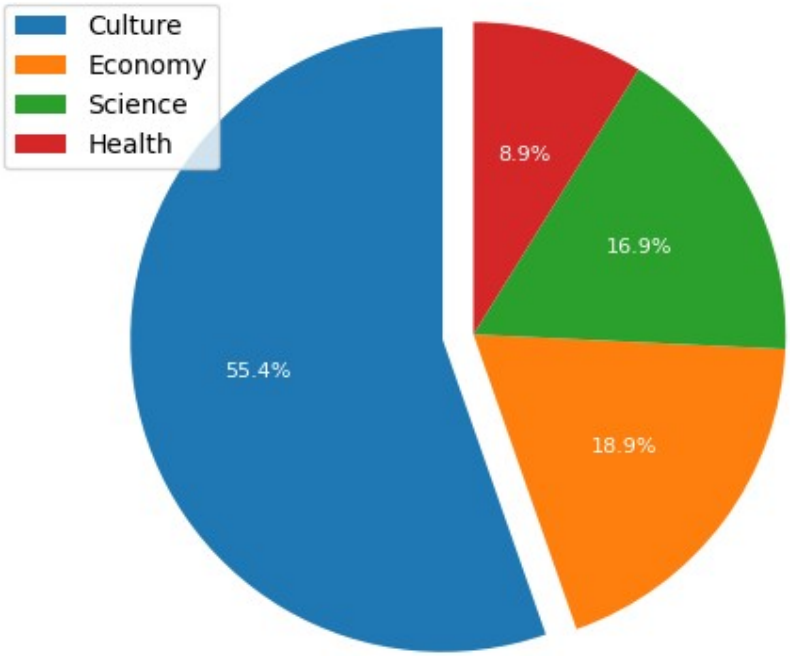


Figure 3.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	19348740	952556	48.00	55.36
Economy	8792036	325083	21.81	18.89
Health	3573336	152711	8.86	8.88
Science	8596935	290302	21.33	16.87

Table 3.4.1: Distribution of domains.

.

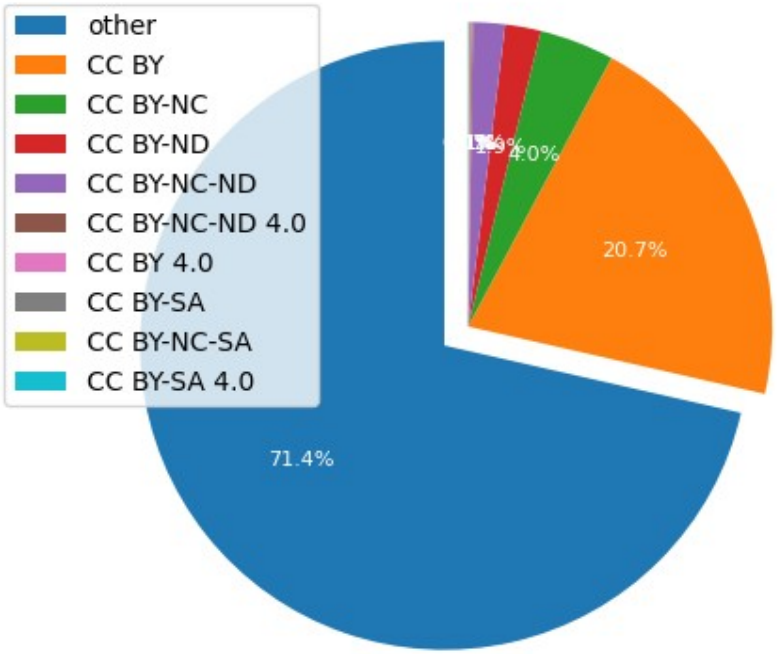


Figure 3.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY	356914	20.74
CC BY 4.0	1265	0.07
CC BY-NC	68230	3.97
CC BY-NC-ND	28673	1.67
CC BY-NC-ND 4.0	1593	0.09
CC BY-NC-SA	427	0.02
CC BY-ND	32676	1.90
CC BY-SA	1159	0.07
CC BY-SA 4.0	415	0.02
other	1229300	71.44

Table 3.4.2: Distribution of licenses.

3.4. Data delivery

The Croatian CURLICAT corpus is not publicly available at the moment. The delivery is planned as soon as some of the selected documents are accompanied with an official permission for redistribution by their IPR owners. After that the full Croatian corpus will be made available at ELRC-SHARE.

4. The Hungarian corpus

4.1. IPR overview

The legal terms of use and redistribution concerning the documents originally selected from the Hungarian National Corpus[1] (Oravecz et al. 2014) are highly restrictive. Only the Wikipedia articles collected from the scientific subcorpus can be freely redistributed. Texts provided by the Digital Literature Academy, mentioned in Deliverable 1.1., had to be discarded from the Hungarian CURLICAT corpus.

Negotiations on the legal status of the documents downloaded from the Hungarian Electronic Library[2] also failed. As a result, the Hungarian CURLICAT corpus currently does not comprise any documents from this source.

However, IPR clearance was successful with Arcanum,[3] which is now the most important data provider for the Hungarian CURLICAT corpus. The legal documents permitting the use and redistribution of the texts acquired from Arcanum are expected to be ratified shortly.

4.2. Additional data acquisition

Arcanum has provided digitized texts published by Akadémiai Kiadó and Osiris Kiadó. Furthermore, a small amount of additional data was downloaded from the REAL-J repository of the Library and Information Centre, Hungarian Academy of Sciences.[4] These are digitized journals available under the licence *Creative Commons Attribution Non-commercial Share Alike* or *Creative Commons Attribution*.

The documents belonging to the publishers Akadémiai Kiadó and Osiris Kiadó were delivered in JSON and/or plain text format. These source files contained text data that had been obtained using OCR technology. The documents from the REAL-J repository were downloaded in PDF format and then converted to plain text with a Python port of the Apache Tika library.[5]

4.3. Domain distribution analysis

The following tables show the distribution of domains in the current version of the Hungarian CURLICAT corpus by the source.

Domain Source	Science	Economy	Culture	Social_issues	total	%
HNC	386984	0	0	0	386984	13.746
Akadémiai Kiadó	168695	175991	660212	63205	1068103	37.939
Osiris Kiadó	565814	162759	333174	228192	1289939	45.819

1 http://corpus.nytud.hu/mnsz/index_eng.html
2 https://mek.oszk.hu/indexeng.phtml
3 https://www.arcanum.com/en/
4 http://real-j.mtak.hu/
5 https://github.com/chrismattmann/tika-python

REAL-J	24051	0	0	46237	70288	2.497
total	1145544	338750	993386	337634	2815314	100
%	40.690	12.032	35.285	11.993	100	

Table 4.3.1. Sentence counts in the Hungarian CURLICAT corpus

Domain Source	Science	Economy	Culture	Social_issues	total	%
HNC	4201693	0	0	0	4201693	6.974
Akadémiai Kiadó	4145596	4569318	13080280	1717579	23512773	39.029
Osiris Kiadó	13722413	3920623	7242273	5755651	30640960	50.861
REAL-J	592439	0	0	1297225	1889664	3.137
total	22662141	8489941	20322553	8770455	60245090	100
%	37.617	14.092	33.733	14.558	100	

Table 4.3.2. Token counts in the Hungarian CURLICAT corpus

The corpus meets the CURLICAT requirement concerning the total number of sentences (at least 2M per language) but the domains ‘Economy’ and ‘Social issues’ are still underrepresented compared to the other two domains. This issue may be resolved by collecting further data and discarding some of the texts classified into the domains ‘Science’ and ‘Culture’.

The corpus consists of 441 documents. The shortest document is 439 tokens long; the longest one 4201693 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 136920.66 tokens, the median is 97146.5 tokens and the standard deviation is 218614.20.

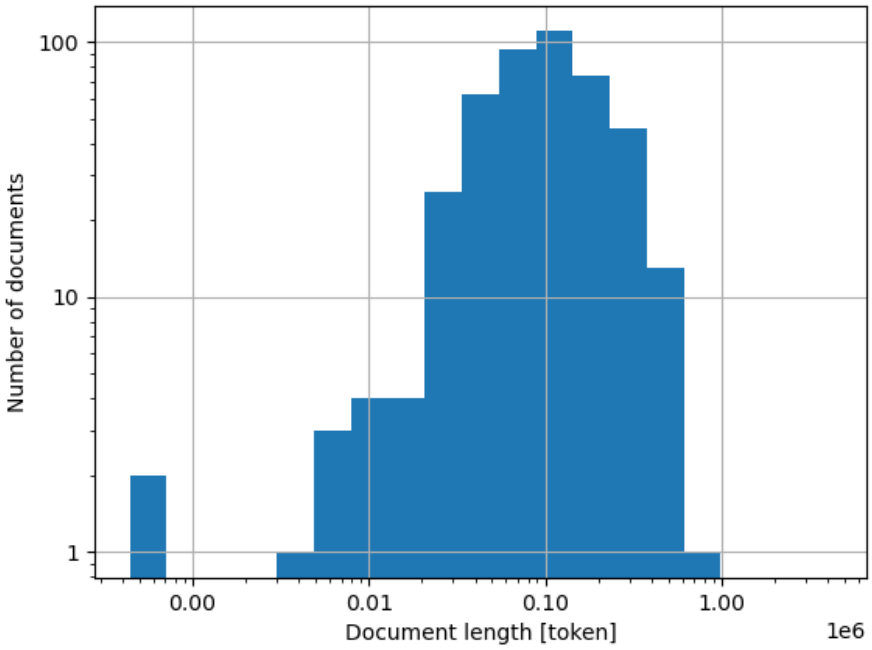


Figure 4.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 28 sentences long; the longest one 386984 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 6398.44 sentences, the median is 4101.5 sentences and the standard deviation is 18772.77.

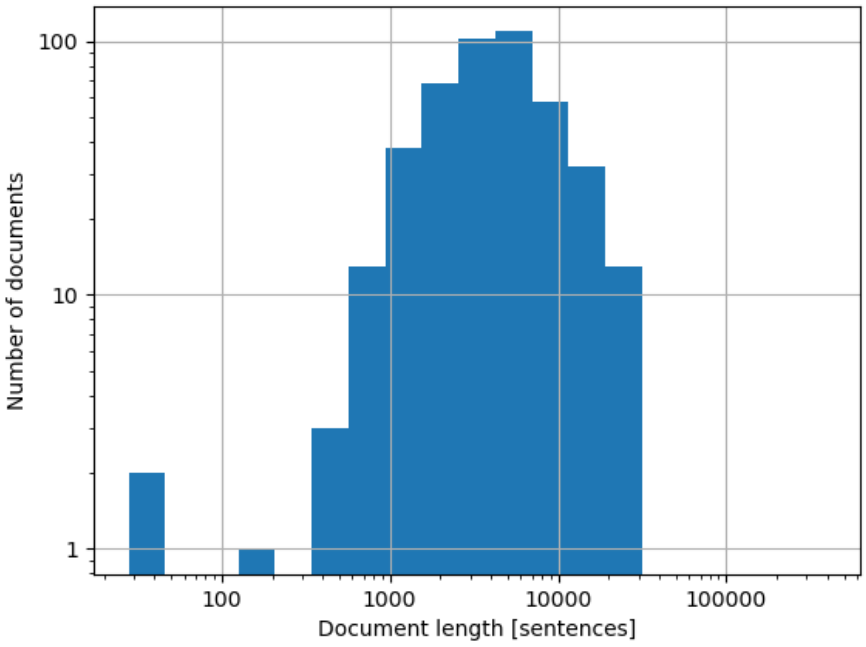


Figure 4.4.2: Distribution of document lengths in sentences, log-log axes.

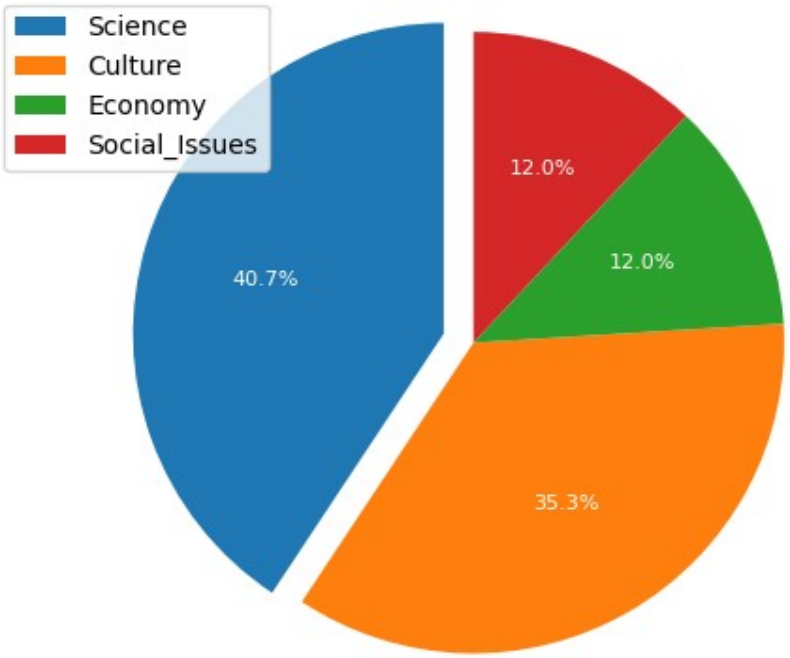


Figure 4.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	20322553	993386	33.73	35.29
Economy	8489941	338750	14.09	12.03
Science	22662141	1145544	37.62	40.69
Social_Issues	8770455	337634	14.56	11.99

Table 4.4.1: Distribution of domains.

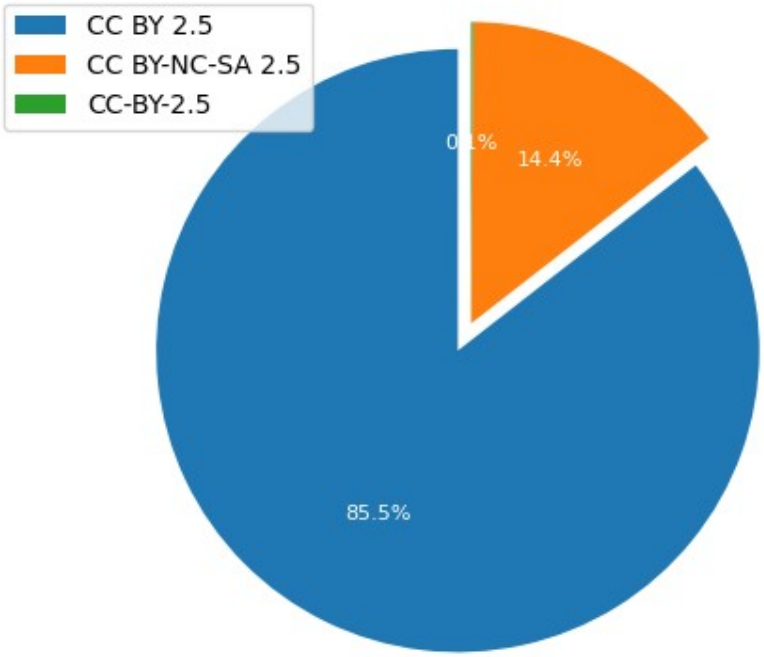


Figure 4.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY 2.5	2407927	85.53
CC BY-NC-SA 2.5	404150	14.36
CC-BY-2.5	3237	0.11

Table 4.4.2: Distribution of licenses.

4.4. Data delivery

The Hungarian CURLICAT corpus is not publicly available at the moment. The delivery is planned as soon as the legal documents that officially permit the redistribution of the Arcanum documents are signed.

5. The Polish Corpus

5.1. IPR overview

For clarification of the intellectual property rights, CURLICAT representatives consulted Jakub Szprot, Head of the Open Science Platform at the Interdisciplinary Centre for Mathematical and Computational Modelling, who confirmed that the articles which are flagged in their metadata headers as available on CC licences could be used without any additional consent from their authors and publishers. Moreover all the metadata (including titles and abstracts) is available on the Creative Commons 0 license.

5.2. Additional data acquisition

A genuinely new source of valuable corpus data was identified in the early stages of the project in the form of the Library of Science (https://bibliotekanauki.pl/), a platform providing open access to full texts of articles published in Polish scientific journals and full texts of selected scientific books together with reach bibliographic metadata. It is one of the results of the Platform of Polish Scientific Publications, a project co-financed by the European Regional Development Fund. The Library of Science is run at the Interdisciplinary Centre for Mathematical and Computer Modelling at the University of Warsaw within the Open Science Platform. The initial investigation showed that several thousand publications with Polish as the main language were available in the Library of Science. Having confirmed a sufficient level of representation of various topics and scientific disciplines in this corpusthe corpus was selected as our primary source of data.

The original content of the Library of Science is imported from five thematic databases, corresponding to domains relevant to CURLICAT:

- AGRO (journals in agricultural sciences and science and life sciences)
- BazTech (journals in engineering sciences and science and life sciences)
- CEJSH (journals in social sciences and humanities)
- DML-PL (journals in engineering sciences)
- PSJD (journals in science and life sciences and in medical and health sciences).

The integrated platform also contains full texts of scientific books previously collected in the Open Book service and other books made available by publishers cooperating with the Library of Science. Articles are available as PDF files and books in EPUB, MOBI and XML formats. One of the limitations of the Library of Science is the format of the full text articles, which are almost invariably PDF documents without any explicit structural annotation of the published texts.

For the purposes of the project 318 088 scientific publications were acquired over the programmatic interface endpoints provided by the Library of Science platform. They were mostly articles and scientific studies and less frequently reviews from 45 disciplines and 8 fields of science published by more than 400 different publishers in more than 1000 scientific journals. The data was initially imported at the metadata level into a relational database using a collector tool (http://git.nlp.ipipan.waw.pl/Marcell/collector).

Although the minimum size of the data (in tokens) to be delivered was almost reached by simply including titles and abstracts (available on the Creative Commons 0 license) we had to extract sentences from the full text PDF documents to meet threshold requirements for the number of

sentences. From over 19k of full texts available with CC-BY and CC-BY-SA licences, after extracting text (using Apache Tika) and applying a series of cleaning rules and scripts we obtained over 45M additional tokens and over 1.7M sentences. Exact numbers of tokens and sentences acquired from abstracts (and titles) and from full texts are presented in the table below.

Type of text	**Sentences**	**Tokens**
Abstracts and titles	683 986	14 168 753
Full texts	1 737 168	45 133 029
Total	**2 421 154**	**59 301 782**

Table 5.2.1. Sizes of abstracts & titles and full texts.

5.3. Domain distribution analysis

Detailed description of mapping from **ScientificDiscipline(s)** (from Library of Science) to CURLICAT domains can be found in the report 5.1 *Metadata Harmonisation in CURLICAT*.

The corpus consists of 115465 documents. The shortest document is 57 tokens long; the longest one 57496 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 513.59 tokens, the median is 138 tokens and the standard deviation is 1147.19.

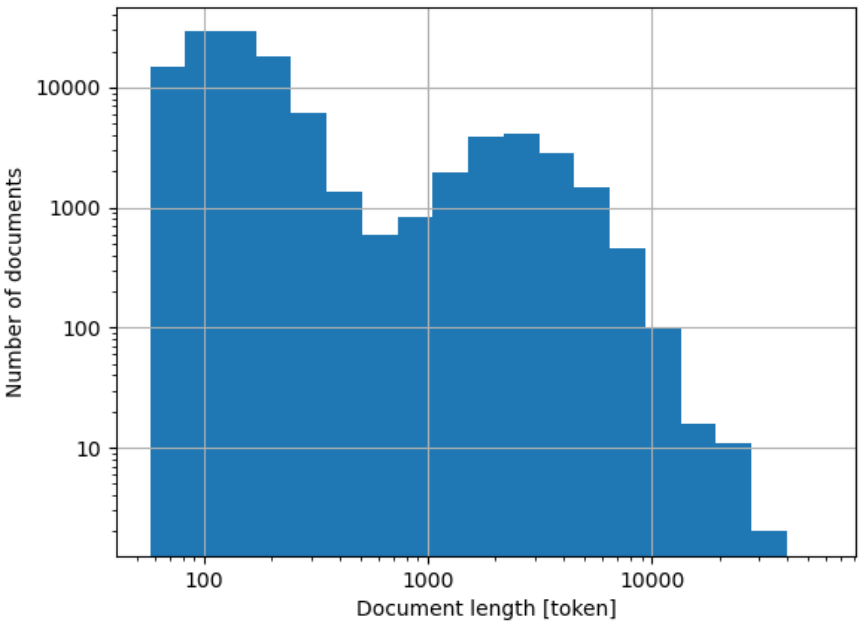


Figure 5.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 1 sentence long; the longest one 2148 sentences long. As expected, the length distribution is skewed to the right, having a tail of a

few long documents. The average length is 20.97 sentences, the median is 7 sentences and the standard deviation is 44.22.

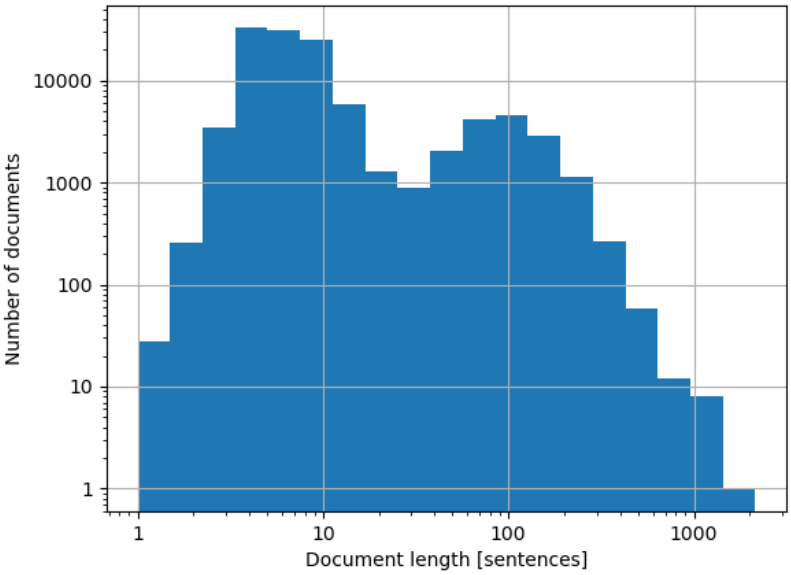


Figure 5.4.2: Distribution of document lengths in sentences, log-log axes.

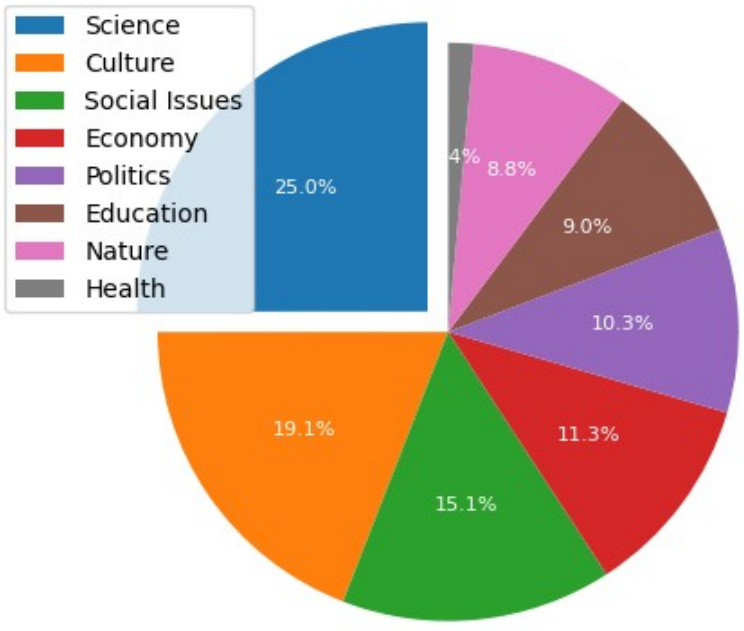
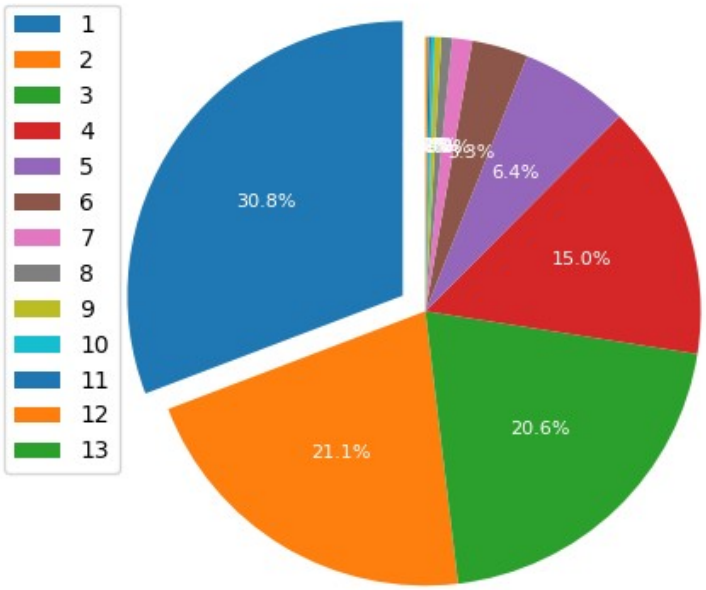


Figure 5.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	11813446	462400	19.92	19.10

Economy	6587872	273888	11.11	11.31
Education	5471707	217913	9.23	9.00
Health	788312	34292	1.33	1.42
Nature	5138481	212872	8.66	8.79
Politics	6651233	249340	11.22	10.30
Science	13521660	605423	22.80	25.01
Social Issues	9329071	365026	15.73	15.08

Table 5.4.1: Distribution of domains.



1–CC BY - Creative Commons Uznanie Autorstwa 4.0; 2–CC0; 3–CC BY - Creative Commons Uznanie Autorstwa 3.0 PL; 4–CC BY-SA Creative Commons Uznanie Autorstwa - Na tych samych warunkach 4.0; 5–CC BY-SA Creative Commons Uznanie Autorstwa - Na tych samych warunkach 3.0 PL; 6–CC BY-NC-ND Creative Commons-Uznanie autorstwa - Użycie niekomercyjne - Bez utworów zależnych 3.0 PL; 7–CC BY-NC-ND Creative Commons-Uznanie autorstwa - Użycie niekomercyjne - Bez utworów zależnych 4.0; 8–CC BY-NC Creative Commons Uznanie Autorstwa - Użycie niekomercyjne 4.0; 9–CC BY-NC Creative Commons Uznanie Autorstwa - Użycie niekomercyjne 3.0 PL; 10–CC BY-NC-SA Creative Commons Uznanie autorstwa - Użycie niekomercyjne - Na tych samych warunkach 4.0; 11–CC BY-ND Creative Commons Uznanie autorstwa - Bez utworów zależnych 4.0; 12–CC BY-ND Creative Commons Uznanie autorstwa - Bez utworów zależnych 3.0 PL; 13–CC BY-NC-SA Creative Commons Uznanie autorstwa - Użycie niekomercyjne - Na tych samych warunkach 3.0 PL;

Figure 5.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY - Creative Commons Uznanie Autorstwa 3.0 PL	498 939	20.61

CC BY - Creative Commons Uznanie Autorstwa 4.0	745 853	30.81
CC BY-NC Creative Commons Uznanie Autorstwa - Użycie niekomercyjne 3.0 PL	9 350	0.39
CC BY-NC Creative Commons Uznanie Autorstwa - Użycie niekomercyjne 4.0	15 362	0.63
CC BY-NC-ND Creative Commons-Uznanie autorstwa - Użycie niekomercyjne - Bez utworów zależnych 3.0 PL	79 833	3.30
CC BY-NC-ND Creative Commons-Uznanie autorstwa - Użycie niekomercyjne - Bez utworów zależnych 4.0	28 517	1.18
CC BY-NC-SA Creative Commons Uznanie autorstwa - Użycie niekomercyjne - Na tych samych warunkach 3.0 PL	653	0.03
CC BY-NC-SA Creative Commons Uznanie autorstwa - Użycie niekomercyjne - Na tych samych warunkach 4.0	4 992	0.21
CC BY-ND Creative Commons Uznanie autorstwa - Bez utworów zależnych 3.0 PL	3 565	0.15
CC BY-ND Creative Commons Uznanie autorstwa - Bez utworów zależnych 4.0	3 759	0.16
CC BY-SA Creative Commons Uznanie Autorstwa - Na tych samych warunkach 3.0 PL	155 463	6.42
CC BY-SA Creative Commons Uznanie Autorstwa - Na tych samych warunkach 4.0	364 200	15.04
CC0	510 668	21.09

Table 5.4.2: Distribution of licenses.

5.4. Data delivery

The local copy of the new corpus was made available at:
http://curlicat.nlp.ipipan.waw.pl/download/latest/

6. The Romanian corpus

6.1. IPR overview

In this phase of the project, we obtained the written agreements from some data providers who did not answer in the previous rounds of negotiations. Comparing the data in Table 6.1 with the similar one from *D1. Collection of multilingual corpora* (Table 6.2, Section 6.2), there are 9 additional text providers who gave a positive answer to our request. As the approached text providers were among those who offered large amounts of data, even after the adjusting of the corpus due to the correction of data and metadata described in section 6.2., our distributable data size greatly exceeds the 2 million sentences target.

Targeted domain	**# selected text providers**	**# text providers without restrictions**	**# sent letters**	**# positive answers**
culture	11	1	10	**5**
economy	9	1	8	**3**
education	8	1	7	**3**
nature	4	1	3	**3**
health	14	1	13	**5**
politics	7	0	7	**5**
science	14	0	14	**4**

Table 6.1. Final no. of selected text providers, sent letters and positive letters.

6.2. Additional data acquisition

The following tables show the final statistics for the Romanian CURLICAT corpus, after collecting new documents and cleaning the data (in the same way as described in *D1, Collection of multilingual corpora,* Section 6.4.) and the metadata (as described below).

To assure the correct document classification according to the CURLICAT domains, an extensive process of manual and automatic validation and correction was done with the metadata files. For all the data acquired from the providers (3042 documents) in table 6.1., the validation was done completely manually. For 27,640 documents coming from Romanian Wikipedia corpus, the classification according to 7 CURLICAT domains was done automatically (see *Deliverable 5.1 Metadata Harmonisation in CURLICAT*, Section 5.3.2). Following the correction process, some documents proved to be not suitable for distribution in CURLICAT, since their newly attributed domain was outside CURLICAT's target. The new dataset – after acquiring new documents, cleaning them, correcting the metadata for all documents and filtering them according to the domain – contains **3,557,812** sentences.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Culture	Arts and Culture	Art History	6 115	176 443
		Literature	179 634	4 992 304
		Film	83 372	2 254 859
		Other	24 206	633 065
		Music	59 133	1 666 469
		Architecture	34 577	869 421
		Painting and Drawing	22 400	581 895
		Sculpture	1 211	34 159
		Theatre	1 433	39 003
		Dance	363	11 147
		Design	169	4 225
TOTAL			**412 613**	**11 262 990**

Table 6.2.1. New statistics for Culture domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Health	Science	Medicine	187 023	5 091 800
	Society	Health	26 689	477 727
	Science	Pharmacology	3 333	78 846
TOTAL			**217 045**	**5 648 373**

Table 6.2.2. New statistics for Health domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Nature	Nature	Environment	6 180	233 428
	Science	Agronomy	5 496	129 142
	Science	Biology	36 142	855 368
	Science	Astronomy	17 901	475 457
	Science	Chemistry	12 083	303 979
	Nature	Other	185	4 907
	Nature	Natural Disasters	214	5 238
TOTAL			**78 201**	**2 007 519**

Table 6.2.3. New statistics for Nature domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Politics	Science	Political Sciences	387 563	11 141 698
	Society	Politics	161 876	2 654 851
TOTAL			**549 439**	**13 796 549**

Table 6.2.4. New statistics for Politics domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Education	Science	Pedagogy	54 968	1 722 092
	Society	Education	83 578	2 394 872
TOTAL			**138 546**	**4 116 964**

Table 6.2.5. New statistics for Education domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of words
Economy	Science	Economy	143748	4034547
	Society	Economy	19366	364149
	Science	Geography	180157	4748094
	Society	Tourism	1602	63403
TOTAL			**344 873**	**9 210 193**

Table 6.2.6. New statistics for Economy domain.

CURLICAT domain	COROLA domain	COROLA subdomain	No. of sentences	No. of tokens
Science	Science	History	683 852	18 260 775
		Psychology	215 364	5 546 382
		Sociology	281 491	7 863 743
		Philosophy	189 113	5 138 868
		Philology	195 336	5 268 347
		Anthropology	27 096	776 337
		Linguistics	102 589	2 837 054

		Other	11 571	260 420
		Mathematics	12 230	315 126
		Religious Studies and Theology	11 946	339 608
		Juridical Sciences	20 964	569 011
		Informatics	26 360	669 896
		Archeology	1 604	44 400
		Technics/technology	12 139	332 246
		Physics	17 140	467 698
		Logics	174	4 822
		Ethnology	561	15 642
		Constructions	304	8 926
		Military Science	4 364	72 721
		Enology	2 897	90 844
Total			**1 817 095**	**48 882 866**

Table 6.2.7. New statistics for Science domain.

CURLICAT DOMAIN	**No. of sentences**	**No. of tokens**	**No. of tokens in percent**
Science	1 817 095	48 882 866	51.49605711
Politics	549 439	13 796 549	14.53408798
Culture	412 613	11 262 990	11.86508942
Economy	344 873	9 210 193	9.702553543
Health	217 045	5 648 373	5.950324978
Education	138 546	4 116 964	4.337049576
Nature	78 201	2 007 519	2.114837397
TOTAL	**3 557 812**	**94 925 454**	**100**

Table 6.2.8. New statistics for the selected data.

6.3. Domain distribution analysis

The corpus consists of 26477 documents. The shortest document is 8 tokens long; the longest one 465202 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few

long documents. The average length is 3585.20 tokens, the median is 677 tokens and the standard deviation is 18693.02.

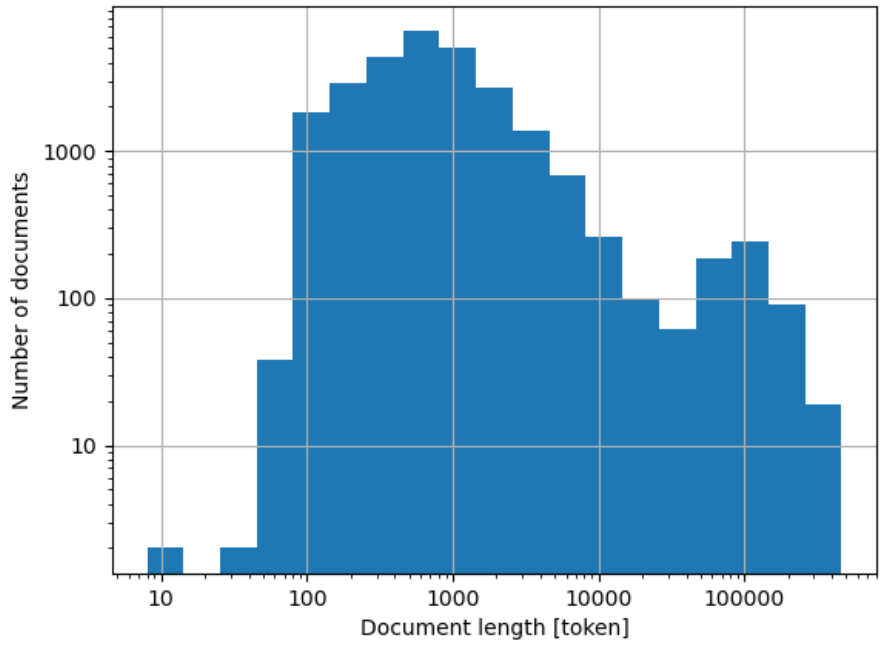


Figure 6.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 1 sentence long; the longest one 27914 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 134.37 sentences, the median is 25 sentences and the standard deviation is 754.49.

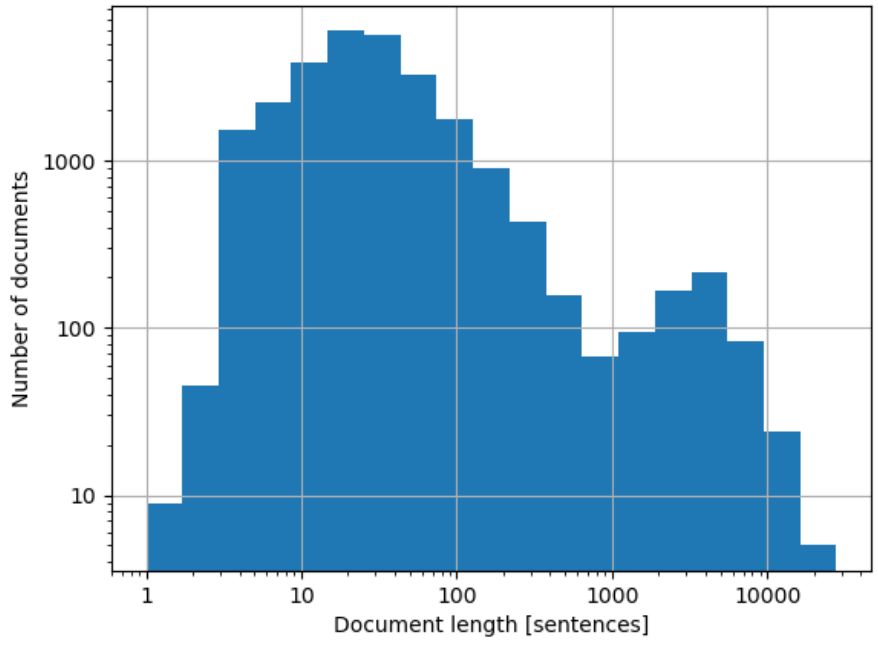


Figure 6.4.2: Distribution of document lengths in sentences, log-log axes.

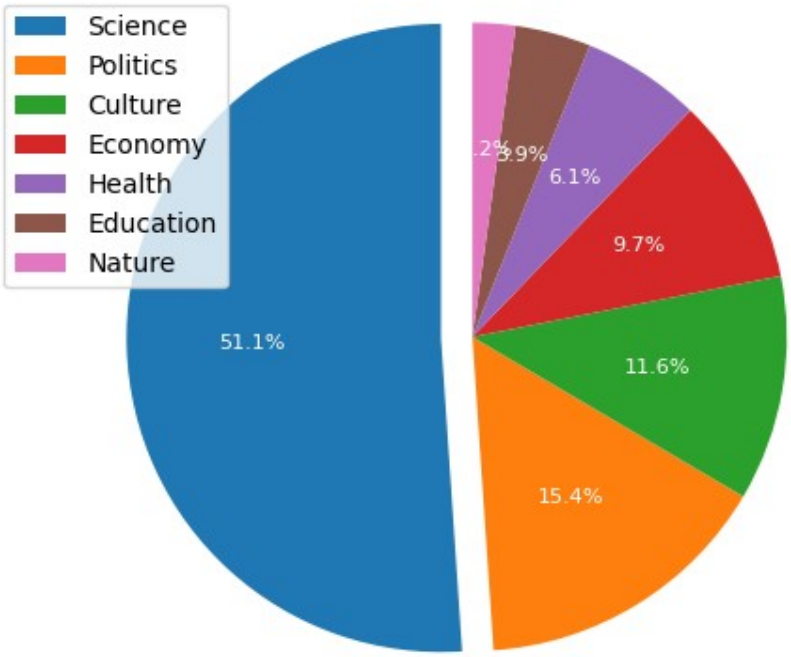


Figure 6.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	11262990	412613	11.87	11.60
Economy	9210193	344873	9.70	9.69
Education	4116964	138546	4.34	3.89
Health	5648373	217045	5.95	6.10
Nature	2007519	78201	2.11	2.20
Politics	13796549	549439	14.53	15.44
Science	48882866	1817095	51.50	51.07

Table 6.4.1: Distribution of domains.

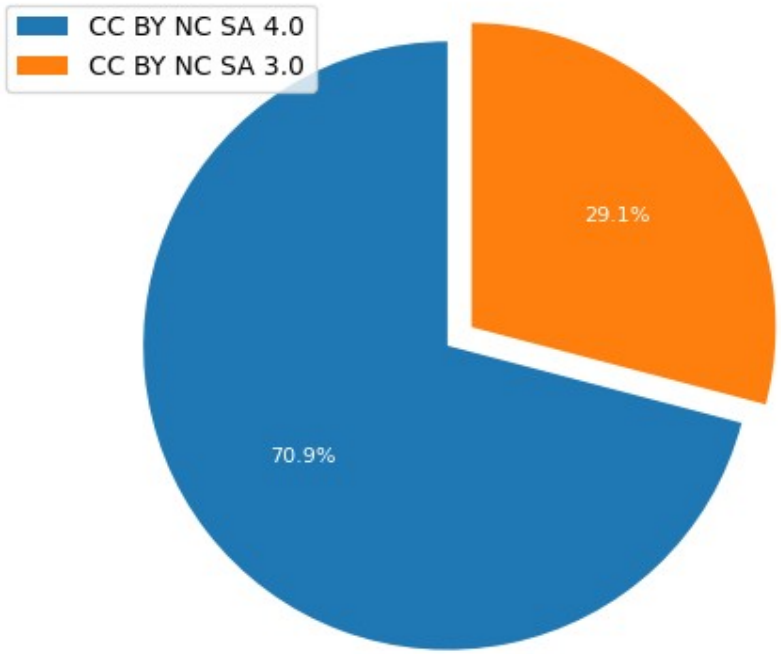


Figure 6.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY NC SA 3.0	1034942	29.09
CC BY NC SA 4.0	2522870	70.91

Table 6.4.2: Distribution of licenses.

6.4. Data delivery

The local copy of the new corpus was made available at: https://relate.racai.ro/index.php?path=corpus/list (see CURLICAT_Anonymsed)

7. The Slovak corpus

7.1. IPR overview

The licenses of documents in the Slovak National Corpus (SNK) *prim-9.0-juls-all* were investigated to select a redistributable subset of the corpus; the redistributable text size is 1718184 tokens. This includes selected articles from Slovak Wikipedia; however, at the time of their inclusion in the corpus these articles were hand picked to contain complete texts, spelling errors were fixed (in Wikipedia, prior to acquiring the texts), templates have been expanded and annotated where appropriate. Therefore we opted for keeping these texts in the redistributable subset.

The corpus *od-justice-1.0* is reasonably big, but the text (court decisions) are rather repetitive and schematic and the domain is very narrow (civil and criminal law). Therefore we did not include the corpus data in the CURLICAT Slovak corpus, but opted to have it available (without any copyright restrictions) separately. Most of the person names in this corpus are already anonymized at the source (i.e. not following the CURLICAT anonymization scheme).

Corpus	Size [tokens]	Size [sentences]	License
subset of *prim-9.0-juls-all*	1 718 184	93 184	varies
wiki-2019-08	47 334 899 (50 619 991)	4 058 505 (4 454 062)	CC BY-SA 3.0
wiki-2018-03 (*Necyklopédia* subset)	1 050 560 (1 056 661)	70 305 (73 075)	CC BY-SA 3.0
od-justice-1.0 (deduplicated)	1 319 172 174	39 817 842	exempt from copyright[6]

Table 7.1. Summary of redistributable corpora and subcorpora of the Slovak National Corpus, available for the CURLICAT project, after improved additional filtering. Numbers in parentheses show the size of the original corpus, not just the selected cleaned up portion.

7.2. Additional data acquisition

For the purposes of the project we identified several potential sources of redistributable texts.

The Greenie library[7] is an online library of freely accessible e-books. SInce the licensing of the books in the library varies and most of the licenses do not allow redistribution, we separated the books with redistributable licenses. We identified 140 books with redistributable licenses (note that the size of the "books" varies – some of them are only one or a few pages long). After removing duplicates (also texts already present in the subset of *prim-9.0-juls-all* corpus), unrecoverable texts with irregular formatting, and books in Czech or English, the final number of documents in the second version of the CURLICAT corpus is 127.

6 § 5 b) of the Copyright Law (185/2015 Z. z.) of the Slovak Republic
7 https://greenie.elist.sk/

Additionally, we downloaded some volumes of several journals published by the various research institutions – recent issues of some of the journals are often published under variants of CreativeCommons licenses, but we had to carefully verify the start date of such licensing and download only the appropriate issues; also taking into account the propensity of publishing in English during recent years. The journals come in various formats and various URL structures and each of them was processed and converted individually.

7.3. Domain distribution analysis

For documents that were not annotated manually by the domain, we used automatic domain classification. The classifier has been trained on the balanced representative corpus of Slovak language *prim-9.0-public-vyv*[8] of 453 594 173 tokens. The classifier uses Stochastic Gradient Descent method and reaches an accuracy of 87.1% on test data. Although the classifier is able to assign multiple domains (and their likelihood) to the document, in order to remain compatible with other data we select only the most likely domain.

7.3.1. Domain distribution in the prim-9.0-juls-all subset

The distribution of domains in the redistributable subset of the *prim-9.0-juls-all* corpus is summarized in Table 7.3.1. The *SNK domain* is the value of the domain metadata field, as encoded in the SNK annotation scheme, with *Domain description* a human-readable description of the value. Detailed description of SNK annotation and the mapping to the CURLICAT metadata can be found in the report 5.1 *Metadata Harmonisation in CURLICAT*.

SNK domain	Domain description	Tokens	%
hum	Humanities	1 126 048	65.54
ars	Arts	307 002	17.87
ins	Interdisciplinary	137 031	7.98
tec	Engineering, technical	63 969	3.72
nat	Natural sciences	31 708	1.85
plt	Politics	16 842	0.98
law	Law	16 504	0.96
ecn	Economics, management	11 171	0.65
lif	Life style	7 909	0.46
Total		**1 718 184**	**100.00**

Table 7.3.1. Domain composition of the redistributable subset of the *prim-9.0-juls-all* corpus.

7.3.2. Domain distribution in newly acquired texts

Newly acquired texts have been annotated using a common CURLICAT metadata scheme. The distribution of the domains is summarized in Table 7.3.2; the distribution is better balanced compared to the SNK data.

8 https://korpus.sk/korpusy-a-databazy/korpusy/struktura-korpusu/verejne-pristupne-korpusy-snk/

Domain	Tokens	%
Science	7 352 904	38.28
Culture	2 907 168	15.14
Nature	2 511 954	13.08
Politics	2 250 736	11.72
General	1 813 863	9.44
Education	1 313 719	6.84
Law	919 369	4.79
Medicine	127 711	0.66
Religion	10 600	0.06
Total	19 208 024	**100.00**

Table 7.3.2. Distribution of newly acquired texts by the domain (in tokens).

7.3. Analysis of the distribution

The corpus consists of 224474 documents. The shortest document is 10 tokens long; the longest one 536246 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 297.99 tokens, the median is 67.0 tokens and the standard deviation is 1792.08.

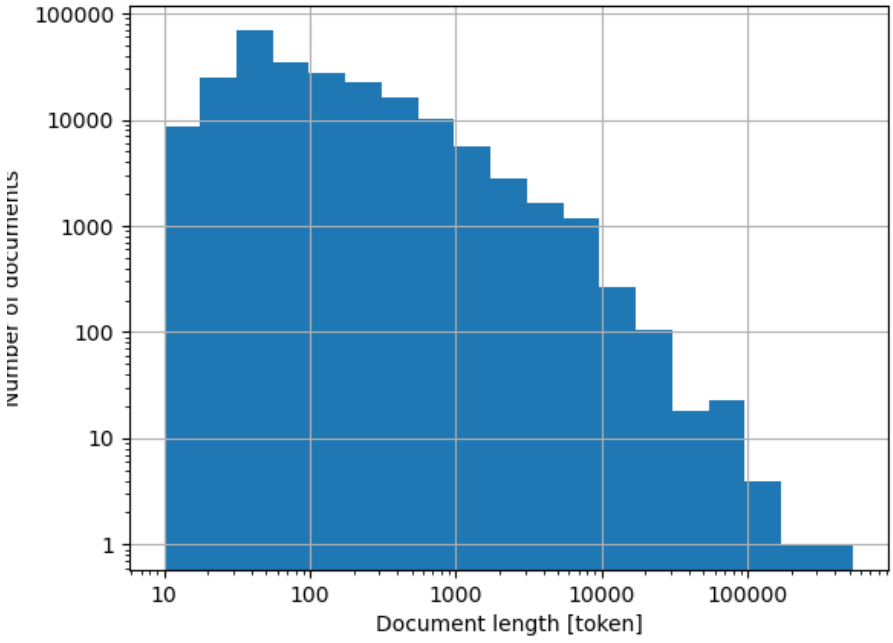


Figure 7.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 1 sentence long; the longest one 26539 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 21.41 sentences, the median is 7.0 sentences and the standard deviation is 107.30.

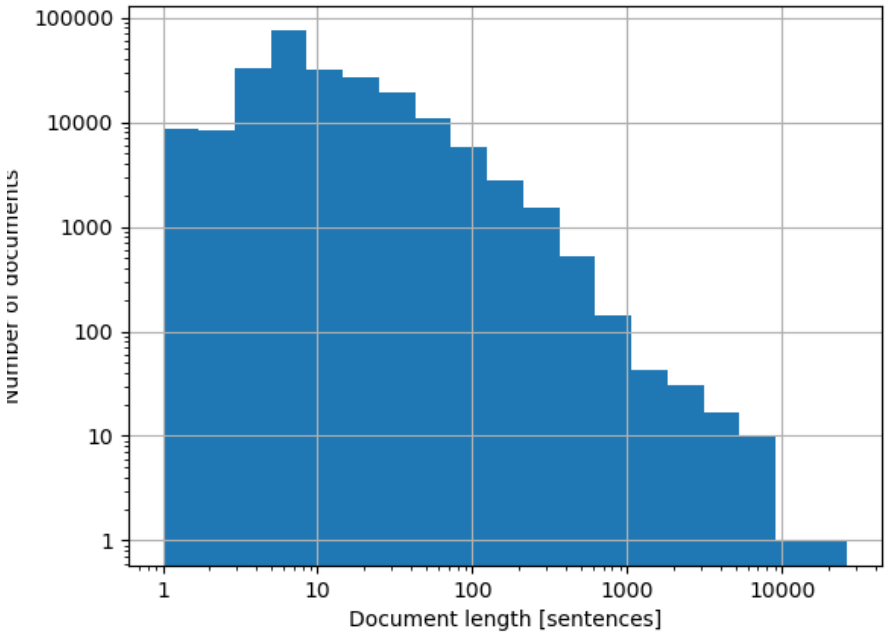


Figure 7.4.2: Distribution of document lengths in sentences, log-log axes.

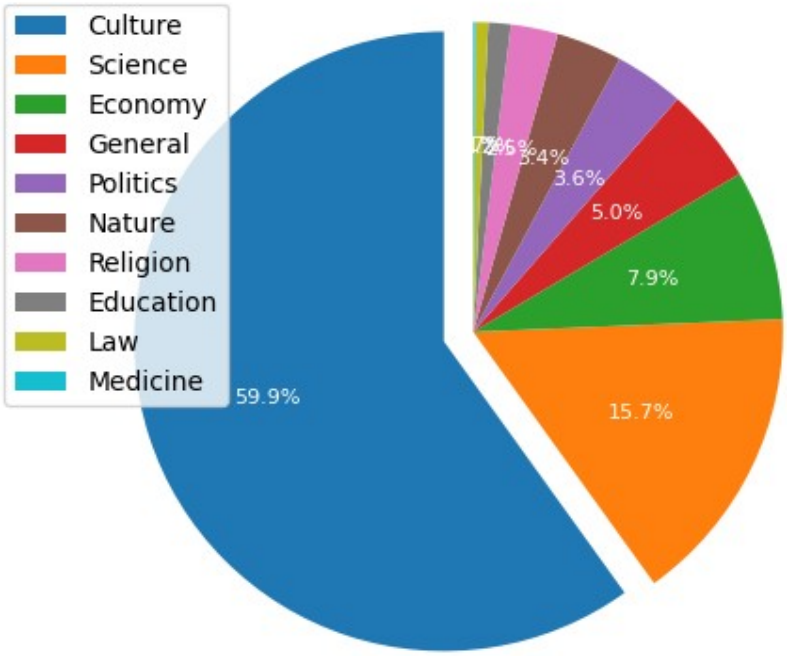


Figure 7.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	33321206	2879967	49.81	59.93
Economy	4771646	378023	7.13	7.87
Education	1313719	55880	1.96	1.16
General	2527313	241019	3.78	5.02
Law	935885	31814	1.40	0.66
Medicine	127711	5489	0.19	0.11
Nature	3562340	165377	5.33	3.44
Politics	3906072	174057	5.84	3.62
Religion	1655998	118325	2.48	2.46
Science	14769255	755725	22.08	15.73

Table 7.4.1: Distribution of domains.

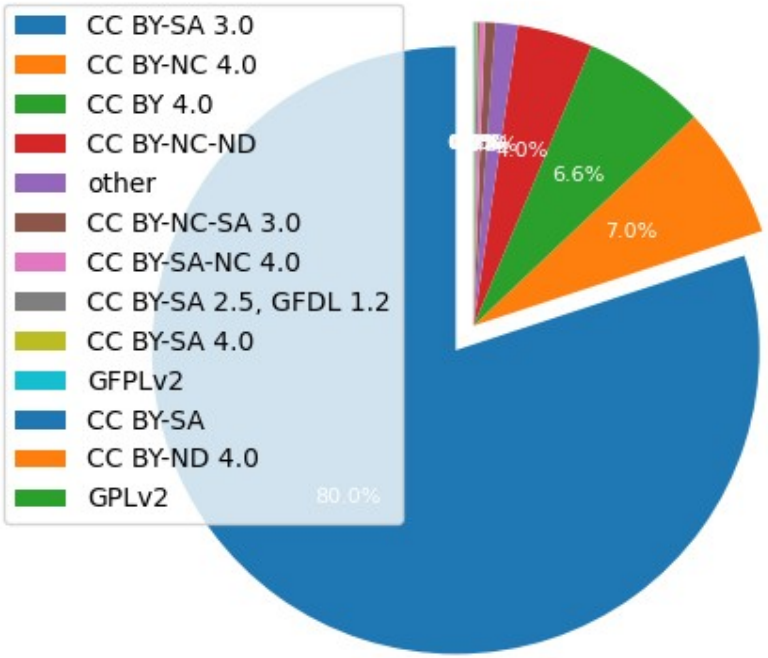


Figure 7.4.4: Distribution of licenses (by number of sentences).

License	sentences	%
CC BY 4.0	316630	6.59
CC BY-NC 4.0	336854	7.01
CC BY-NC-ND	192705	4.01
CC BY-NC-SA 3.0	26539	0.55
CC BY-ND 4.0	1072	0.02
CC BY-SA	1253	0.03
CC BY-SA 2.5, GFDL 1.2	7380	0.15
CC BY-SA 3.0	3846777	80.05
CC BY-SA 4.0	3204	0.07

CC BY-SA-NC 4.0	13034	0.27
GFPLv2	2117	0.04
GPLv2	803	0.02
other	57308	1.19

Table 7.4.2: Distribution of licenses.

7.4. Data delivery

The local copy of the second version corpus is available via the partner's CURLICAT local webpage at https://www.juls.savba.sk/curlicat.html.

The corpus *od-justice-1.0* is available at https://www.juls.savba.sk/justicecorp.html.

8. The Slovenian corpus

8.1. IPR overview

Under the Gigafida 2.0 text provision agreement, 10% of the corpus can be shared under the Creative Commons Attribution-ShareAlike 4.0 (CC-BY 4.0 SA) license which allows users to freely use and redistribute the data provided that appropriate attribution is made. Hence there was no need for additional IPR clearance.

8.2. Additional data acquisition

All the data provided to the CURLICAT project corpus comes from the Gigafida 2.0 corpus maintained by the Centre for Language Resources and Technologies, of which the JSI Institute is a member. The corpus data is readily available for export in various formats. In Activity 1, we were able to identify a large amount of appropriate texts for the CURLICAT corpus. To meet the promised amount of 2 million sentences, we contacted existing Gigafida text donors in the domains of culture, economics, health and politics and acquired additional texts allowing us to meet the requirements of the CURLICAT project.

8.3. Domain distribution analysis

The corpus consists of 790 documents. The shortest document is 35 tokens long; the longest one 2762310 tokens long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 55039.95 tokens, the median is 877.5 tokens and the standard deviation is 219466.89.

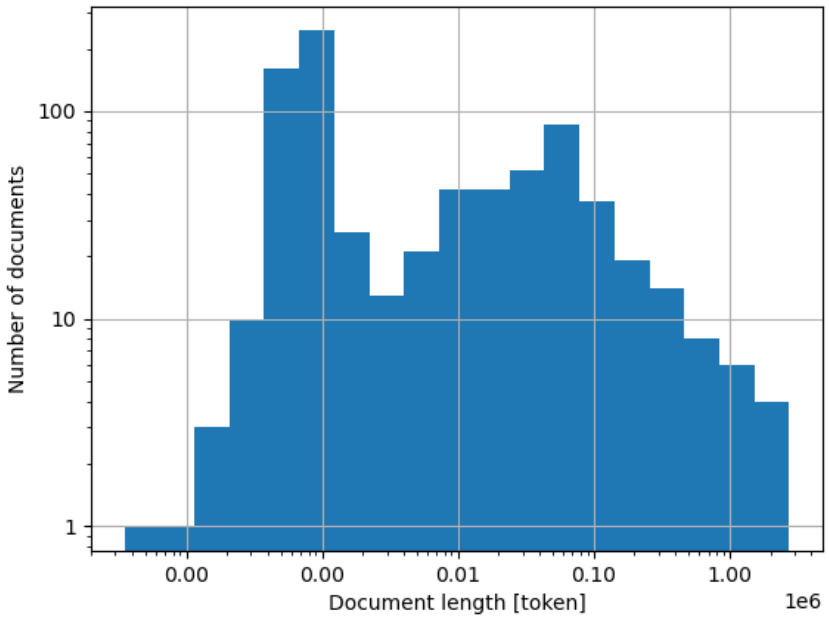


Figure 8.4.1: Distribution of document lengths in tokens, log-log axes.

By the length measured in sentences, the shortest document is 2 sentences long; the longest one 120738 sentences long. As expected, the length distribution is skewed to the right, having a tail of a few long documents. The average length is 2536.24 sentences, the median is 48.0 sentences and the standard deviation is 9950.14.

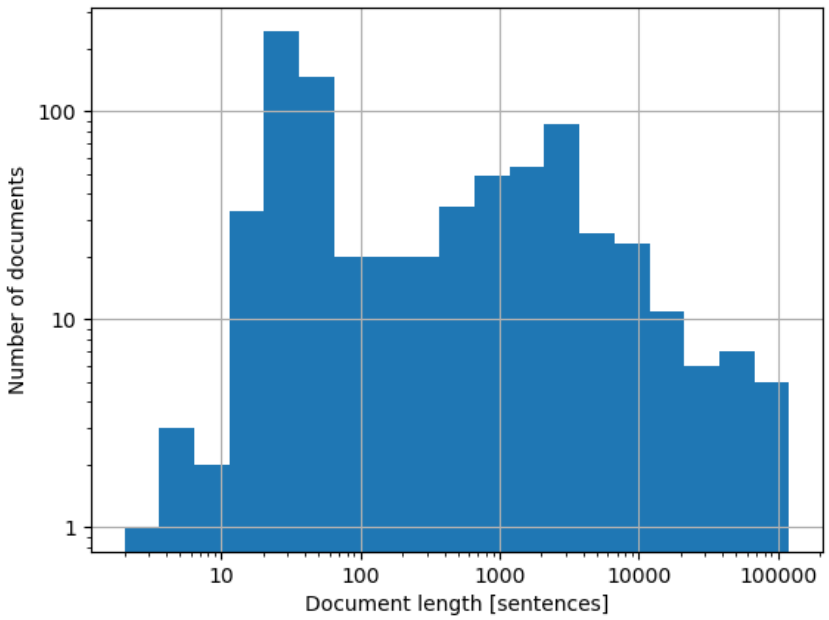


Figure 8.4.2: Distribution of document lengths in sentences, log-log axes.

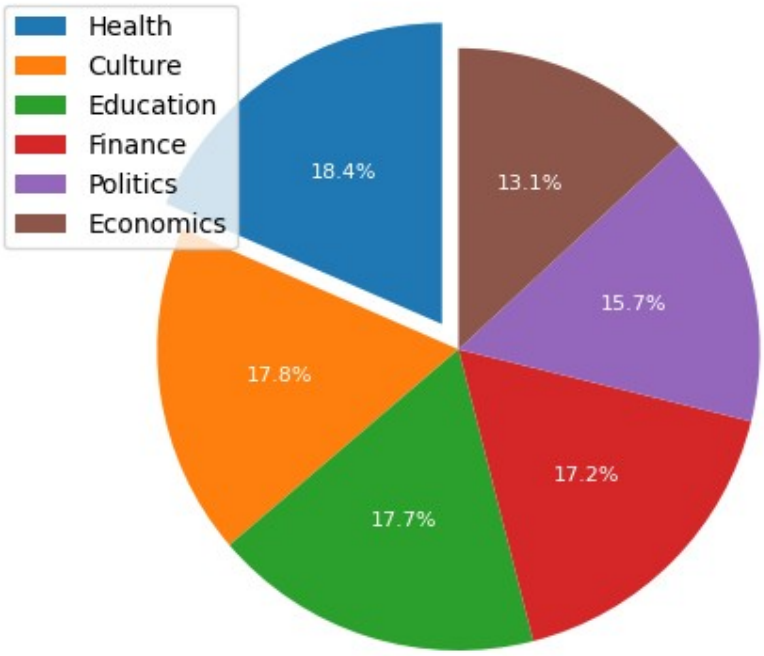


Figure 8.4.3: Distribution of domains (by number of sentences).

domain	tokens	sentences	tokens [%]	sentences [%]
Culture	7268772	356910	16.72	17.81
Economics	6380019	262418	14.67	13.10
Education	7274946	355300	16.73	17.73
Finance	7888640	343956	18.14	17.17
Health	7342310	369606	16.89	18.45
Politics	7326876	315436	16.85	15.74

Table 8.4.1: Distribution of domains.

8.4. Data delivery

The Slovene CURLICAT corpus is not publicly available at the moment. The delivery is planned as soon as the technical formatting of the corpus data is harmonized with the rest of CURLICAT corpora.

9. Bibliographical references

Garabík R. (2010). Slovak National Corpus tools and resources. In: Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies (WIKT 2010). Laclavík, M., Hluchý, L (eds.). Bratislava, ISBN 978-80-970145-2-0, pp. 2 – 7.

Koeva S., Stoyanova I., Leseva S., Dekova R., Dimitrova T., Tarpomanova E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. Journal of Language Modelling (1), pp. 65–110. https://doi.org/10.15398/jlm.v0i1.33

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede.

Mititelu V. B., Tufiș D., Irimia E., Păiș V., Ion R., Diewald N., Mitrofan M., Onofrei M. (2019). Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian. Revue Roumaine de Linguistique, No./Issue 3.

Oravecz C., Váradi T., Sass B. (2014). The Hungarian Gigaword Corpus. In Proceedings of LREC 2014.

Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. (2012). Narodowy Korpus Języka Polskiego. PWN Scientific Publishers.

Tadić, M., (2009) New Version of the Croatian National Corpus. In Hlaváčková D., Horák A., Osolsobě K., Rychlý P. (eds.) After Half a Century of Slavonic Natural Language Processing, pp. 221-228, Tribun EU, Brno.