Corpus of Decisions

Permanent Court of International Justice

(CD-PCIJ)

Codebook

Version 2021-11-23

DOI: 10.5281/zenodo.3840480

Title	Corpus of Decisions: Permanent Court of International Justice
Abbreviation	CD-PCIJ
Author	Seán Fobbe
Version	2021-11-23
Download	https://doi.org/10.5281/zenodo.3840480
License	CC0 1.0 Universal

Citation

Seán Fobbe (2021). Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ). Version 2021-11-23. Zenodo. DOI: 10.5281/zenodo.3840480.

Digital Object Identifiers: Concept DOI and Version DOI

This data set is uniquely identified via the Digital Object Identifier (DOI) system. DOIs are persistent identifiers that are globally unique and can be resolved as a link by entering a DOI into the web service at www.doi.org. The DOI given in this document is a 'Version DOI', which uniquely identifies version 2021-11-23. Academics and others who wish to enable replication analyses are strongly advised to cite the *version DOI* and the precise version of the data used. A 'Concept DOI' is available from the page of the Zenodo record under the heading 'Cite all versions?' and will always resolve to the latest version.

Public Domain Status

Disclaimer

This data set is a personal academic initiative and is not associated with or endorsed by the International Court of Justice or the United Nations.

Contents

1 Introduction

The **Permanent Court of International Justice (PCIJ)** was the primary judicial organ of the League of Nations, the ill-fated predecessor of the United Nations, which existed from 1920 to 1946.

Nonetheless, as the first international court with general thematic jurisdiction the PCIJ influenced international law in profound ways that are still felt today. Every lawyer who sets out on the path of international law encounters epoch-defining opinions such as the *Lotus* and *Factory at Chorzów* decisions, but the Court's lesser-known jurisprudence and the appended minority opinions offer many more ideas and legal principles which are seldom appreciated today.

The **Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)** collects and presents for the first time in human- and machine-readable formats all documents of PCIJ Series A, B and A/B. Among these are judgments, advisory opinions, orders, appended minority opinions, annexes, applications instituting proceedings and requests for an advisory opinion. The International Court of Justice, the successor of the PCIJ, has kindly made available these documents on its website.

This data set is designed to be complementary to and fully compatible with the *Corpus of Decisions: International Court of Justice (CD-ICJ)*, which is also available open access.[1]

The quantitative analysis of international legal data is still in its infancy, a situation which is exacerbated by the lack of high-quality empirical data. Most advanced data sets are held in commercial databases and are therefore not easily available to academic researchers, journalists and the general public. With this data set I hope to contribute to a more systematic and empirical view of the international legal system. In an international community founded on the rule of law the activities of the judiciary must be public, transparent and defensible. In the 21st century this requires quantitative scientific review of decisions and actions.

Design, construction and compilation of this data set are based on the principles of general availability through freedom from copyright (public domain status), strict transparency and full scientific reproducibility. The *FAIR Guiding Principles for Scientific Data Management and Stewardship* (Findable, Accessible, Interoperable and Reusable) inspire both the design and the manner of publication.[2]

[1] Corpus of Decisions: International Court of Justice (CD-ICJ). <https://doi.org/10.5281/zenodo.3826445>.

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.

2 Reading Files

The data are published in open, interoperable and widely used formats (CSV, TXT, PDF). They can be used with all modern programming languages (e.g. Python or R) and graphical interfaces. The PDF collections are intended to facilitate traditional legal research.

Important: Missing values are always coded as 'NA'.

2.1 CSV Files

Working with the CSV files is recommended. CSV[3] is an open and simple machine-readable tabular data format. In this data set values are separated by commas. Each column is a variable and each row is a document. Variables are explained in detail in section 5.

To read **CSV** files into R I strongly recommend using the fast file reader **fread()** from the **data.table** package (available on CRAN). The file can be read into **R** like so:

```
library(data.table)
pcij.en <- fread("./filename.csv")
```

2.2 TXT Files

The **TXT** files, including metadata, can be read into **R** with the package **readtext** (available on CRAN) thus:

```
library(readtext)
pcij.en <- readtext("EN_TXT_TESSERACT_FULL/*.txt",
                    docvarsfrom = "filenames",
                    docvarnames = c("court",
                                    "series",
                                    "seriesno",
                                    "shortname",
                                    "applicant",
                                    "respondent",
                                    "date",
                                    "doctype",
                                    "collision",
                                    "opinion",
                                    "language"),
                    dvsep = "_",
                    encoding = "UTF-8")
```

[3] The CSV format is defined in RFC 4180: <https://tools.ietf.org/html/rfc4180>.

3 Data Set Design

3.1 Description of Data Set

The **Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)** collects and structures in human- and machine-readable formats all documents of PCIJ Series A, B and A/B. Among these are judgments, advisory opinions, orders, appended minority opinions, annexes, applications instituting proceedings and requests for an advisory opinion.

It consists of a CSV file of the full data set, a CSV file with the metadata only, individual TXT files for each document and PDF files with an enhanced text layer generated by the LSTM neural network engine of the optical character recognition software (OCR) *Tesseract*.

Additionally, the raw PDF files and some intermediate stages of refinement are included to allow for easier replication of results and for production use in the event that even higher quality methods of optical character recognition (OCR) can be applied to the documents in the future.

3.2 Complementarity

This data set is intended to be complementary to and fully compatible with the *Corpus of Decisions: International Court of Justice (CD-ICJ)*, which is also available open access.[4]

3.3 Table of Sources

Data Source	Citation
Primary Data Source	https://icj-cij.org/en/pcij
Source Code	https://doi.org/10.5281/zenodo.4136956
Country Codes	https://doi.org/10.5281/zenodo.4136956
Names and Parties of Cases	https://doi.org/10.5281/zenodo.4136956

3.4 Data Collection

Data were collected with the explicit consent of the Registry of the International Court of Justice. All documents were downloaded via TLS-encrypted connections and cryptographically signed after data processing was complete. The data set collects all decisions and appended opinions issued by the Permanent Court of International Justice in Series A, B and A/B and which were published on the official website of the International Court of Justice on the day of compilation.

3.5 Source Code and Compilation Report

The full Source Code for the creation of this data set, the resulting Compilation Report and this Codebook are published open access and permanently archived in the scientific

[4] Corpus of Decisions: International Court of Justice (CD-ICJ). <https://doi.org/10.5281/zenodo.3826445>.

repository of CERN.

With every compilation of the full data set an extensive **Compilation Report** is created in a professionally layouted PDF format (comparable to this Codebook). The Compilation Report includes the Source Code, comments and explanations of design decisions, relevant computational results, exact timestamps and a table of contents with clickable internal hyperlinks to each section. The Compilation Report is published under the same DOI as the Source Code.

For details of the construction and validation of the data set please refer to the Compilation Report.

3.6 Limitations

Users should bear in mind certain limitations:

1. The data set contains only those documents which were published by the PCIJ and have been made available by the ICJ on its official website (*publication bias*)
2. While Tesseract yields high-quality OCR results, current OCR technology is not perfect and minor errors must be expected (*OCR bias*)
3. Lengthy quotations in foreign languages may confound analyses (*language blurring*)

3.7 Public Domain Status

According to written communication between the author and the Registry of the International Court of Justice the original documents are not subject to copyright.

To ensure the widest possible distribution and to promote the international rule of law I waive any copyright to the data set under a **Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication**. For details of the license please refer to the CC0 copyright notice at the beginning of this Codebook or visit the Creative Commons website for the full terms of the license.[5]

3.8 Quality Assurance

Dozens of automated tests were conducted to ensure the quality of the data and metadata, for example:

1. Auto-detection of language via analysis of n-gram patterns with the *textcat* package for R.
2. Strict validation of variable types via *regular expressions*.
3. Construction of frequency tables for (almost) every variable followed by human review to detect anomalies.
4. Creation of visualizations for many common descriptive analyses.

For results of each test and more information on the construction of the data set please refer to the Compilation Report or the 'ANALYSIS' archive included with the data set.

[5] Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. <https://creativecommons.org/publicdomain/zero/1.0/legalcode>.

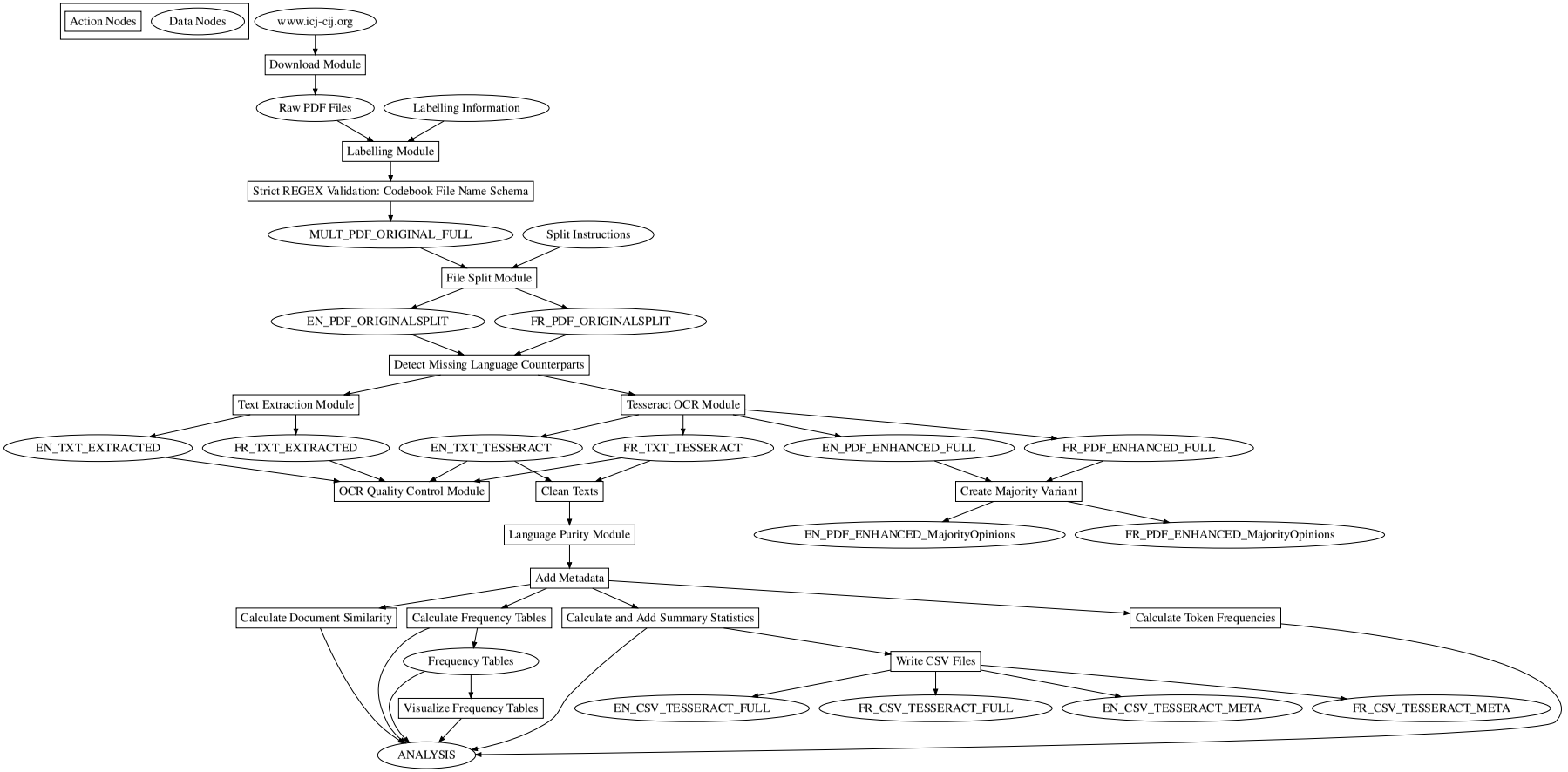


Figure 1: CD-PCIJ: Workflow Schematic

4 Variants and Primary Target Audiences

The data set is provided in two primary language versions (English and French), as well as several differently processed variants geared towards specific target audiences.

A reduced PDF variant of the data set containing only majority opinions is intended to assist practitioners.

Variant	Target Audience and Description
PDF_ENHANCED	**Traditional Legal Research (recommended)**. These PDF files contain the original document as a scan plus an enhanced text layer created with an LSTM neural network machine learning engine. Its main advantages are vastly improved local searches in individual documents via Ctrl+F and copy/pasting without the need for extensive manual revisions. Unlike the original documents, English and French documents have been split into separate document collections and do not alternate in the same document. Researchers with slow internet connections should consider using the 'TXT_TESSERACT' variant, as this still provides a reasonable visual approximation of the original documents, but offers the advantage of drastically reduced file size. A reduced PDF variant of the data set containing only majority opinions is available to assist practitioners.
CSV_TESSERACT	**Quantitative Research (recommended).** A structured representation of the full data set within a single comma-delimited file. Includes the full complement of metadata described in the Codebook. The 'FULL' sub-variant includes the full text of the decisions, whereas the sub-variant 'META' only contains the metadata.
TXT_TESSERACT	**Quantitative Research.** Monolingual TXT files generated with an advanced LSTM neural network machine learning engine from monolingual PDF documents based on the original scans (stored in the collection 'PDF_OriginalSplit'). R users should strongly consider using the package *readtext* to read them into R with the filename metadata intact.
ANALYSIS	**Quantitative Research.** This archive contains almost all of the machine-readable analysis output generated during the data set creation process to facilitate further analysis (CSV for tables, PDF and PNG for plots). Minor analysis results are documented only in the Compilation Report.
TXT_EXTRACTED	**Replication Research and Creation of New Data Sets.** TXT files containing the extracted text layer from the monolingual PDF documents. The quality of the OCR text layer is poor und this variant should not be used for statistical analysis.

Variant	Target Audience and Description
MULT_PDF_ORIGINAL	**Replication Research and Creation of New Data Sets.** The original documents with the original text layer. English, French and sometimes German alternate within the same document. Some very few documents are monolingual. Only recommended for researchers who wish to replicate the machine-readable files or who wish to create a new and improved data set. May be useful in traditional research.
PDF_ORIGINALSPLIT	**Replication Research and Creation of New Data Sets.** The original documents split into monolingual documents. Only recommended for researchers who wish to replicate the machine-readable files or who wish to create a new data set with improved OCR.

5 Variables

5.1 General Remarks

- Missing values are always coded as 'NA'.
- All Strings are encoded in UTF-8.
- All of the metadata contained in the file names was coded manually by the author based on the contents of each document and should exactly reflect the information given in each document. No filename metadata supplied by the Court was retained. Hand-coded data is added automatically at compilation time. Country codes conform to the ISO 3166 Alpha-3 standard and geographical classifications to the M49 standard used by the UN Statistics Division.
- The variable 'fullname' is coded according to case headings as published on the ICJ website and corrected by reviewing the full text of each document. Includes information on the stage of proceedings in parentheses. Introductory phrases such as 'Case concerning...' are omitted.
- The variables 'nchars', 'ntokens', 'ntypes', 'nsentences' and 'year' were calculated automatically based on the content and metadata of each document.
- The variables 'version', 'doi_concept', 'doi_version' and 'license' were added automatically during the data set creation process to document provenance and to comply with FAIR Data Principles F1, F3 and R1.1.

5.2 Structure of TXT File Names

`[court]_[series]_[seriesno]_[shortname]_[applicant]_[respondent]_`
`[date]_[doctype]_[collision]_[stage]_[opinion]_[language]`

5.3 Example TXT File Name

`PCIJ_A_10_Lotus_FRA_TUR_1927-09-07_JUD_01_ME_00_EN.txt`

5.4 Structure of CSV Metadata

```
## Classes 'data.table' and 'data.frame':   259 obs. of  29 variables:
##  $ doc_id              : chr  "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU
    _1923-01-16_APP_01_NA_NA_EN.txt" "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU
    _1923-05-22_APP_01_NA_NA_EN.txt" "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU
    _1923-06-28_JUD_01_IN_00_EN.txt" "PCIJ_A_01_Wimbledon_GBR-FRA-ITA-JPN_DEU
    _1923-08-17_JUD_01_ME_00_EN.txt" ...
##  $ court               : chr  "PCIJ" "PCIJ" "PCIJ" "PCIJ" ...
##  $ series              : chr  "A" "A" "A" "A" ...
##  $ seriesno            : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ caseno              : chr  "A1" "A1" "A1" "A1" ...
##  $ shortname           : chr  "Wimbledon" "Wimbledon" "Wimbledon" "Wimbledon"
    ...
##  $ fullname            : chr  "S.S. Wimbledon" "S.S. Wimbledon" "S.S.
    Wimbledon" "S.S. Wimbledon" ...
##  $ applicant           : chr  "GBR-FRA-ITA-JPN" "GBR-FRA-ITA-JPN" "GBR-FRA-ITA
    -JPN" "GBR-FRA-ITA-JPN" ...
##  $ respondent          : chr  "DEU" "DEU" "DEU" "DEU" ...
##  $ applicant_region    : chr  "Europe|Europe|Europe|Asia" "Europe|Europe|
    Europe|Asia" "Europe|Europe|Europe|Asia" "Europe|Europe|Europe|Asia" ...
##  $ respondent_region   : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ applicant_subregion : chr  "Northern Europe|Western Europe|Southern Europe|
    Eastern Asia" "Northern Europe|Western Europe|Southern Europe|Eastern Asia" "
    Northern Europe|Western Europe|Southern Europe|Eastern Asia" "Northern Europe
    |Western Europe|Southern Europe|Eastern Asia" ...
##  $ respondent_subregion: chr  "Western Europe" "Western Europe" "Western
    Europe" "Western Europe" ...
##  $ date                : IDate, format: "1923-01-16" "1923-05-22" ...
##  $ doctype             : chr  "APP" "APP" "JUD" "JUD" ...
##  $ collision           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stage               : chr  NA NA "IN" "ME" ...
##  $ opinion             : int  NA NA 0 0 1 2 0 1 2 3 ...
##  $ language            : chr  "EN" "EN" "EN" "EN" ...
##  $ year                : int  1923 1923 1923 1923 1923 1923 1924 1924 1924
    1924 ...
##  $ minority            : int  NA NA 0 0 1 1 0 1 1 1 ...
##  $ nchars              : int  4136 1860 5323 37540 10654 17125 72181 37351
    52344 19197 ...
##  $ ntokens             : int  792 351 1026 7189 1953 3143 13075 6920 9936 3529
     ...
##  $ ntypes              : int  326 158 366 1426 617 798 1884 1191 1777 765 ...
##  $ nsentences          : int  26 9 30 179 50 80 325 243 292 103 ...
##  $ version             : chr  "1.0.0" "1.0.0" "1.0.0" "1.0.0" ...
##  $ doi_concept         : chr  "10.5281/zenodo.3840479" "10.5281/zenodo
    .3840479" "10.5281/zenodo.3840479" "10.5281/zenodo.3840479" ...
##  $ doi_version         : chr  "10.5281/zenodo.3840480" "10.5281/zenodo
    .3840480" "10.5281/zenodo.3840480" "10.5281/zenodo.3840480" ...
##  $ license             : chr  "Creative Commons Zero 1.0 Universal" "Creative
    Commons Zero 1.0 Universal" "Creative Commons Zero 1.0 Universal" "Creative
    Commons Zero 1.0 Universal" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

5.5 Detailed Description of Variables

Variable	Type	Details
doc_id	String	(CSV only) The name of the imported TXT file.
text	String	(CSV only) The full content of the imported TXT file.
court	String	The variable only takes the value 'PCIJ', which stands for 'Permanent Court of International Justice'. It is generally only useful if combined with the CD-ICJ or other data sets.
series	String	This variable denotes the PCIJ Series in which the document was published. It takes the values 'A', 'B' or 'AB'.
seriesno	Integer	The number assigned to each collection of documents within a PCIJ Series. Not necessarily unique across series (Series A and B have overlapping numbers).
caseno	String	(CSV only) A combination of the variables 'series' and 'seriesno'. The same case may span multiple case numbers, i.e. preliminary objections decisions often have a different case number than the judgment on the merits. To analyze all stages of a case I recommend a pattern search on the variable 'shortname'. Note: case number A18/19 is coded separately as A18 and A19.
shortname	String	Short name of the case. This was custom-created by the author based on the original title. Short names include well-known components (e.g. 'Lotus') to facilitate quick local searches and try to be as faithful to the full title as possible. Where more than one set of documents exists for a case the stage of proceedings can help differentiate them.
fullname	String	(CSV only) Full name of the case. Coded according to case headings as published on the ICJ website and revised by reviewing the full text of each document. Includes information on the stage of proceedings in parentheses. Introductory phrases such as 'Case concerning...' are omitted.

Variable	Type	Details
applicant	String	The unique identifier of the applicant. In contentious proceedings this is the three-letter (Alpha-3) country code as per the ISO 3166-1 standard. Table 6.1 contains an explanation of all country codes used in the data set. Please note that reserved country codes are in use for historical entities (e.g. Yugoslavia). For advisory proceedings this variable refers to the entity which requested an advisory opinion. In this data set advisory opinions were only ever requested by the Council of the League of Nations, coded as 'LNC'.
respondent	String	The unique identifier of the respondent. In contentious proceedings this is the three-letter (Alpha-3) country code as per the ISO 3166-1 standard. Table 6.1 contains an explanation of all country codes used in the data set. Please note that reserved country codes are in use for historical entities (e.g. the Soviet Union). Advisory proceedings do not have a respondent and therefore always take the value 'NA'.
applicant_region	String	(CSV only) The geographical region of the applicant according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for international organizations.
respondent_region	String	(CSV only) The geographical region of the respondent according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for international organizations.
applicant_subregion	String	(CSV only) The geographical subregion of the applicant according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for international organizations.
respondent_subregion	String	(CSV only) The geographical subregion of the respondent according to the UN M49 standard. Please refer to table 6.1 for details and exceptions. Geographical information is only available for countries, not for international organizations.
date	ISO Date	The date of the document in the format YYYY-MM-DD (extended ISO-8601).

Variable	Type	Details
doctype	String	A three-letter code indicating the type of document. Possible values are 'JUD' (judgments in contentious jurisdiction), 'ADV' (advisory opinions), 'ORD' (orders in all types of jurisidiction), 'REQ' (requests by parties during the proceedings), 'APP' (applications instituting proceedings in both types of jurisdiction), 'DEC' (decision, only used once) or 'ANX' (annexes to documents of the same date, usually a list of documents submitted during the proceedings).
collision	Integer	In some instances several documents with otherwise identical metadata were issued on the same day. This is generally the case for Annexes, but also for a very few substantive documents. Almost all documents take the value '01'. If documents would be assigned identical metadata, the value is incremented.
stage	String	The stage of proceedings, coded based on the title page (primary), or a close reading of the findings (secondary). Possible values are given in table 7. The PCIJ is somewhat more consistent than the ICJ in storing specific stages of proceedings in discrete documents. I am cautiously in favor of performing computational analyses based on this variable, but caution should be exercised nonetheless.
opinion	Integer	A sequential number assigned to each opinion. Majority opinions are always coded '00'. Minority opinions begin with '01' and ascend to the maximum number of minority opinions. For documents of a type other than 'JUD', 'ADV' or 'ORD' this variable takes the value 'NA'.
language	String	The language of the document as a two-letter ISO 639-1 code. This data set mainly contains documents in the languages English ('EN') and French ('FR'), as well as a very few documents in German ('DE').
year	Integer	(CSV only) The year the document was issued. The format is YYYY.
minority	Integer	(CSV only) This variable indicates whether the document is a majority (0) or minority (1) opinion.
nchars	Integer	(CSV only) The number of characters in a given document.

Variable	Type	Details
ntokens	Integer	(CSV only) The number of tokens (an arbitrary character sequence bounded by whitespace) in a given document. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice.
ntypes	Integer	(CSV only) The number of *unique* tokens. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice.
nsentences	Integer	(CSV only) The number of sentences in a given document. The rules for detecting sentence boundaries are very complex and are described in 'Unicode Standard Annex No 29'. This metric can vary significantly depending on tokenizer and parameters used. This count was generated based on plain tokenization with no further pre-processing (e.g. stopword removal, removal of numbers, lowercasing) applied. Analysts should use this number not as an exact figure, but as an estimate of the order of magnitude of a given document's length. If in doubt, perform an independent calculation with the software of your choice.
version	String	(CSV only) The version of the data set in the format MAJOR.MINOR.PATCH, e.g. 1.0.0.

Variable	Type	Details
doi_concept	String	(CSV only) The Digital Object Identifier (DOI) for the *concept* of the data set. Resolving this DOI via www.doi.org allows researchers to always acquire the *latest version* of the data set. The DOI is a persistent identifier suitable for stable long-term citation. Principle F1 of the FAIR Data Principles ('data are assigned globally unique and persistent identifiers') recommends the documentation of each data set with a persistent identifier and Principle F3 its inclusion with the metadata. Even if the CSV data set is transmitted without the accompanying Codebook this allows researchers to establish provenance of the data.
doi_version	String	(CSV only) The Digital Object Identifier (DOI) for the *specific version* of the data set. Resolving this DOI via www.doi.org allows researchers to always acquire this *specific version* of the data set. The DOI is a persistent identifier suitable for stable long-term citation. Principle F1 of the FAIR Data Principles ('data are assigned globally unique and persistent identifiers') recommends the documentation of each data set with a persistent identifier and Principle F3 its inclusion with the metadata. Even if the CSV data set is transmitted without the accompanying Codebook this allows researchers to establish provenance of the data.
license	String	(CSV only) The license of the data set. In this data set the value is always 'Creative Commons Zero 1.0 Universal'. Ensures compliance with FAIR Data principle R1.1 ('clear and accessible data usage license').

6 Applicant and Respondent Codes

6.1 Contentious Jurisdiction: States

Applicants and Respondents in contentious jurisdiction are coded according to the uppercase three-letter (Alpha-3) country codes described in the ISO 3166-1 standard. The codes are taken from the version of the standard which was valid on 4 November 2020. The table below only includes those codes which are used in the data set. The table below only includes those codes which are used in the data set. The regions and subregions assigned to States generally follow the UN Standard Country or Area Codes for Statistics Use, 1999 (Revision 4), also known as the M49 standard.

Please note that where States have ceased to exist (Yugoslavia, Czechoslovakia) their historical three-letter country codes from ISO 3166-1 are used. These are not part of the current ISO 3166-1 standard, but have been transitionally reserved by the ISO 3166 Maintenance Agency to ensure backwards compatibility. The four-letter ISO 3166-3 standard ('Code for formerly used names of countries') is not used in this data set. The regions and subregions for Yugoslavia and Czechoslovakia are taken from M49 revision 2 (1982).

ISO-3	Name	Region	Sub-Region
BEL	Belgium	Europe	Western Europe
BGR	Bulgaria	Europe	Eastern Europe
BRA	Brazil	Americas	Latin America and the Caribbean
CHE	Switzerland	Europe	Western Europe
CHN	China	Asia	Eastern Asia
CSK	Czechoslovakia	Europe	Eastern Europe
DEU	Germany	Europe	Western Europe
DNK	Denmark	Europe	Northern Europe
ESP	Spain	Europe	Southern Europe
EST	Estonia	Europe	Northern Europe
FRA	France	Europe	Western Europe
GBR	United Kingdom	Europe	Northern Europe
GRC	Greece	Europe	Southern Europe
HUN	Hungary	Europe	Eastern Europe
ITA	Italy	Europe	Southern Europe
JPN	Japan	Asia	Eastern Asia
LTU	Lithuania	Europe	Northern Europe
NLD	Netherlands	Europe	Western Europe
NOR	Norway	Europe	Northern Europe
POL	Poland	Europe	Eastern Europe
TUR	Turkey	Asia	Western Asia
YUG	Yugoslavia	Europe	Southern Europe

6.2 Advisory Jurisdiction: Entities

Only a single entity, the Council of the League of Nations, requested advisory opinions from the Permanent Court of International Justice. It is coded as 'LNC'.

7 Stages of Proceedings

This variable encodes a more granular view of the different stages of proceedings that are the subject of PCIJ decisions. The tables is ordered roughly in order of occurrence, although each case only provides documents for a few, select number of stages.

Stage	Doctype	Details
SE	ORD	Settlement
IN	ORD	Request for Intervention
AJ	ORD	Ad Hoc Judges
EV	ORD	Evidence
EX	ORD	Expert Witnesses
JO	ORD	Joinder of Proceedings
IM	ORD	Interim Measures
TL	ORD	Time Limit
DH	ORD	Date of Hearing
PR	ORD	Prorogration
DI	ORD	Discontinuance
PO	ORD, JUD	Preliminary Objections
ME	JUD	Merits

8 Linguistic Metrics

8.1 Explanation of Metrics

To better communicate the scope of the corpus and its constituent documents I provide a number of classic linguistic metrics and visualize their distributions:

Metric	Definition
Characters	Characters roughly correspond to graphemes, the smallest functional unit in a writing system. The word 'judge' is composed of 5 characters, for example.
Tokens	An arbitrary character sequence delimited by whitespace on both sides, e.g. it roughly corresponds to the notion of a 'word'. However, due to its strictly syntactical definition it might also include arbitrary sequences of numbers or special characters.
Types	Unique tokens. If, for example, the token 'human' appeared one hundred times in a given document, it would be counted as only one type.
Sentences	Corresponds approximately to the colloquial definition of a sentence. The exact rules for determining sentence boundaries are very complex and may be reviewed in 'Unicode Standard: Annex No 29'.

8.2 Summary Statistics

8.2.1 English

Metric	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	6,830,889	310	6,326.5	16,267	26,374.09	35,301.5	180,875
ntokens	1,298,030	62	1,294.0	3,107	5,011.70	6,726.0	33,652
ntypes	22,517	46	368.0	701	859.24	1,174.0	3,157
nsentences	38,266	6	38.0	97	147.75	203.5	821

8.2.2 French

Metric	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	6,892,111	345	6,484	15,686	26,406.56	36,374	182,539
ntokens	1,262,293	70	1,217	2,851	4,836.37	6,527	32,752
ntypes	29,204	51	406	774	985.03	1,383	3,741
nsentences	34,176	5	32	88	130.94	179	729

8.3 Explanation of Diagrams

8.3.1 Distributions of Document Length

The diagrams in Section 8.4 are combined violin and box plots. They are especially useful in visualizing distributions of quantitative variables. Their interpretation is fairly straightforward: the greater the area under the curve for a given range, the more frequent the values are in this range. The thick center line of the box indicates the median, the outer lines of the box the first and third quartiles. Whiskers extend outwards to 1.5 times the inter-quartile range (IQR). Outliers beyond 1.5 times IQR are shown as individual points.

Please note that the x-axis is logarithmically scaled, i.e. in powers of 10. It therefore increases in a non-linear fashion. Additional sub-markings are included to assist with interpretation.

8.3.2 Most Frequent Tokens

A token is defined as any character sequence delimited by whitespace on both sides, e.g. it roughly corresponds to the notion of a 'word'. However, due to the strictly syntactical definition tokens might also include arbitrary sequences of numbers or special characters.

The charts in Sections 8.5 and 8.6 show the 50 most frequent tokens for each language, weighted by both term frequency (TF) and term frequency/inverse document frequency (TF-IDF). Sequences of numbers, special symbols and a general list of frequent words for English and French ('stopwords') were removed prior to constructing the list. For details of the calculations, please refer to the Compilation Report and/or the Source Code.

The term frequency tf_{td} is calculated as the raw count of the number of times a term t appears in a document d.

The term frequency/inverse document frequency tf-idf_{td} for a term t in a document d is calculated as follows, with N the total number of documents in a corpus and df_t being the number of documents in the corpus in which the term t appears:

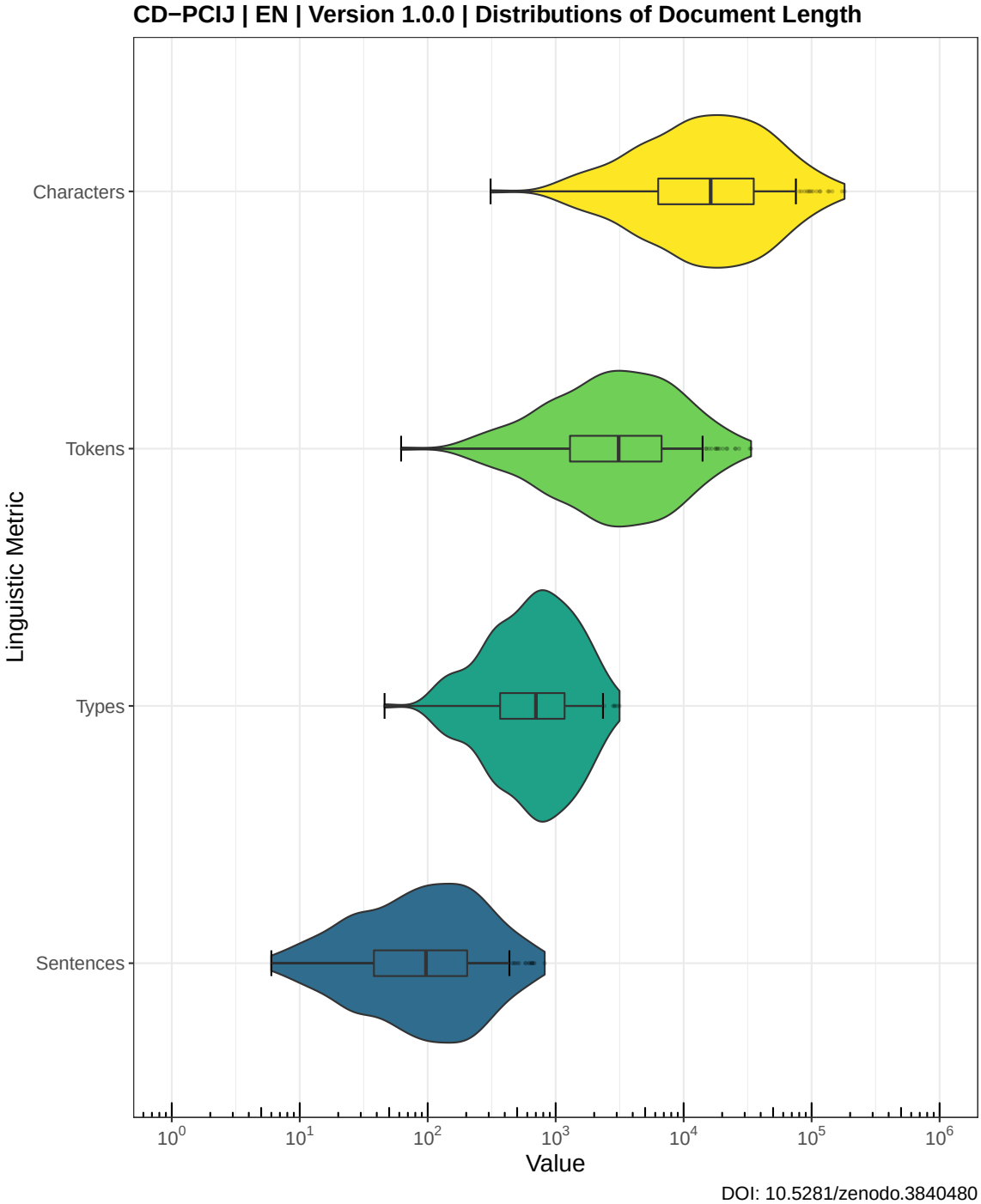
$$\text{tf-idf}_{td} = \mathrm{tf}_{td} \times \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

8.3.3 Tokens over Time

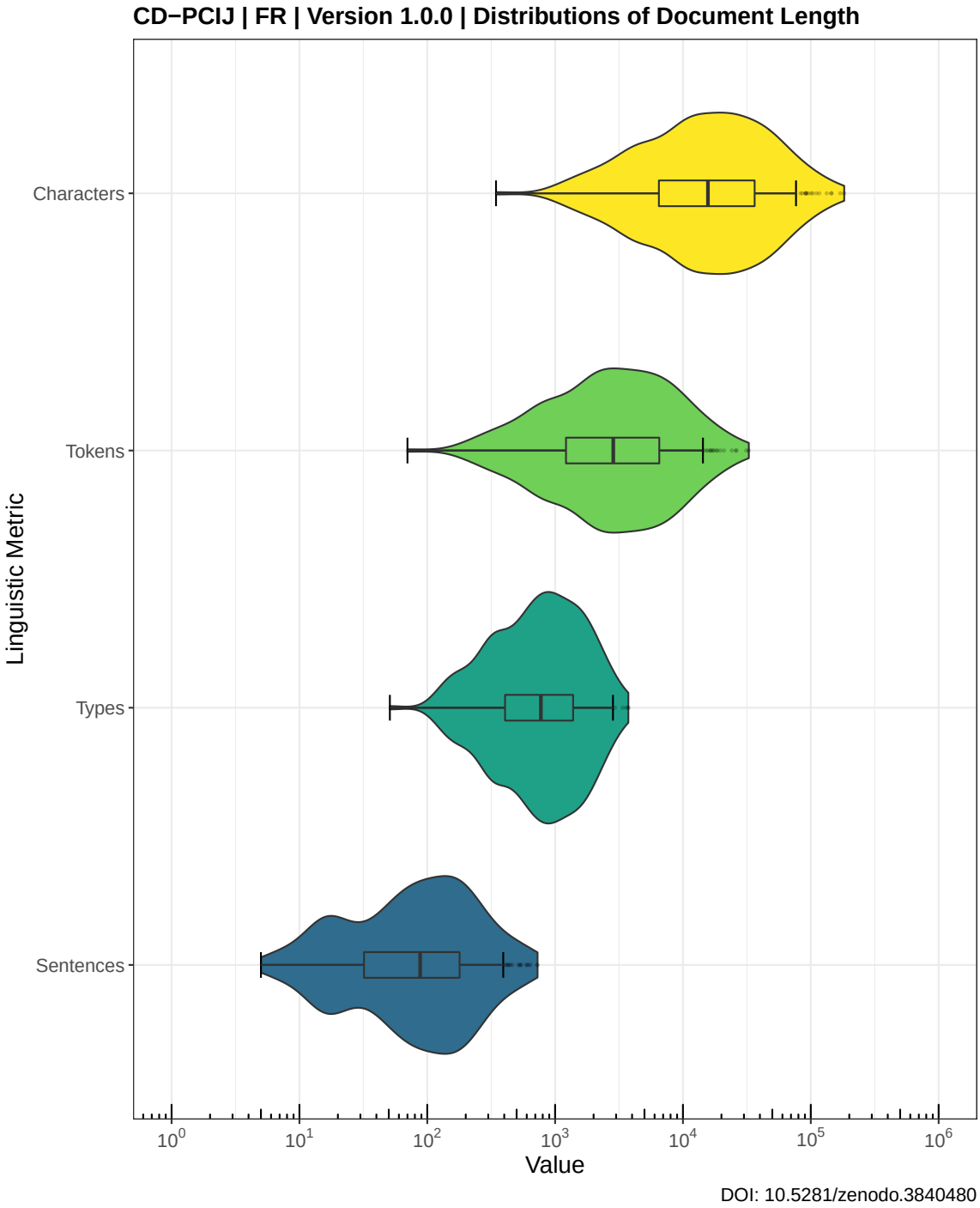
The charts in Section 8.7 show the total published output of the Permanent Court of International Justice for each year as the sum total of the tokens of all published documents (judgments, advisory opinions, orders, appended opinions, appendices, requests). These charts may give a rough estimate of the activity of the Permanent Court of International Justice, although they should be interpreted with caution, as appendices, requests and duplicate documents were not removed for this simple analysis. Please refer to Section 9 for the scope of identical and near-identical documents in the corpus.

8.4 Distributions of Document Length

8.4.1 English

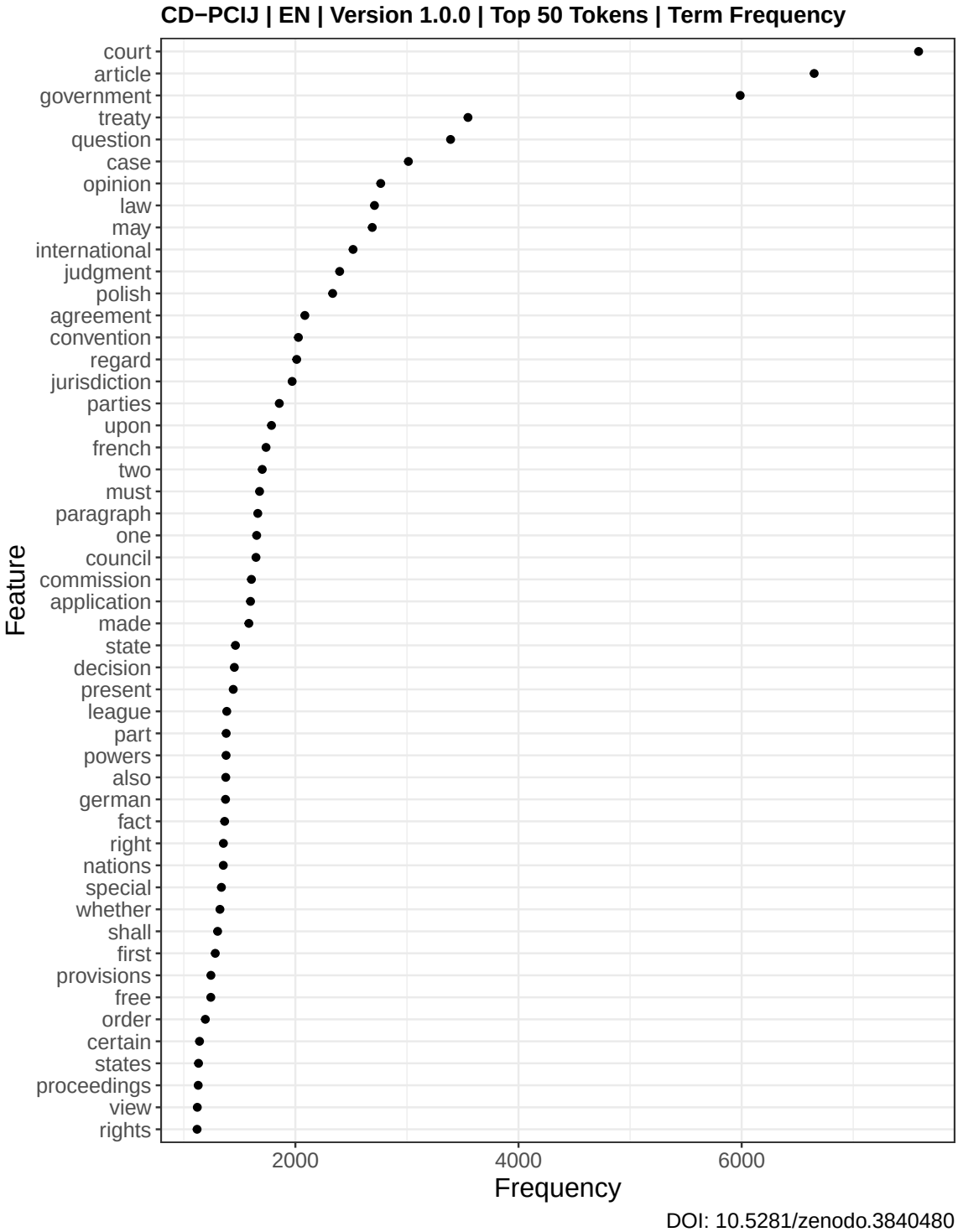


8.4.2 French


CD−PCIJ | FR | Version 1.0.0 | Distributions of Document Length
Characters
Tokens
Types
Sentences
Linguistic Metric
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Value
DOI: 10.5281/zenodo.3840480

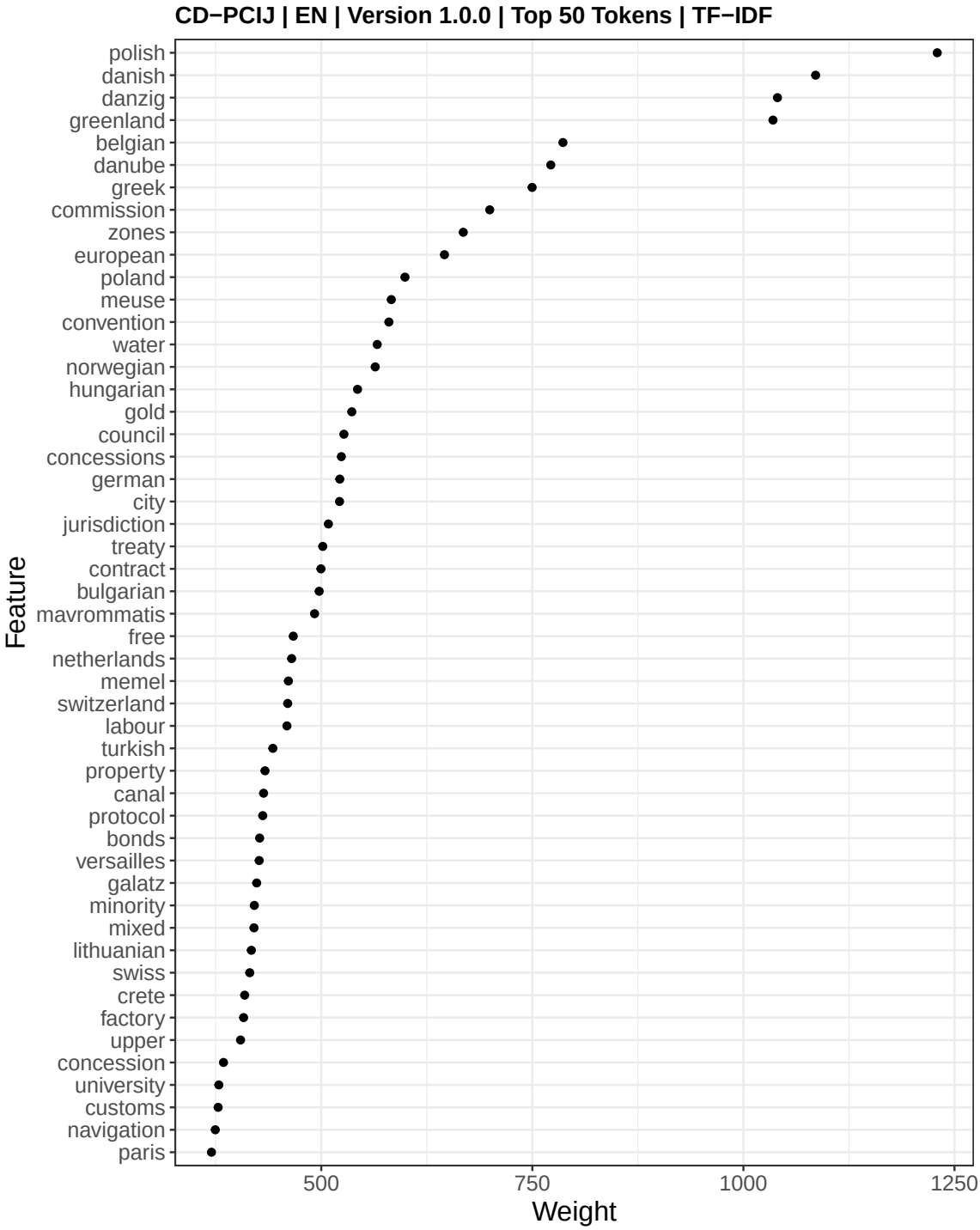
8.5 Most Frequent Tokens (English)

8.5.1 Term Frequency Weighting (TF)



DOI: 10.5281/zenodo.3840480

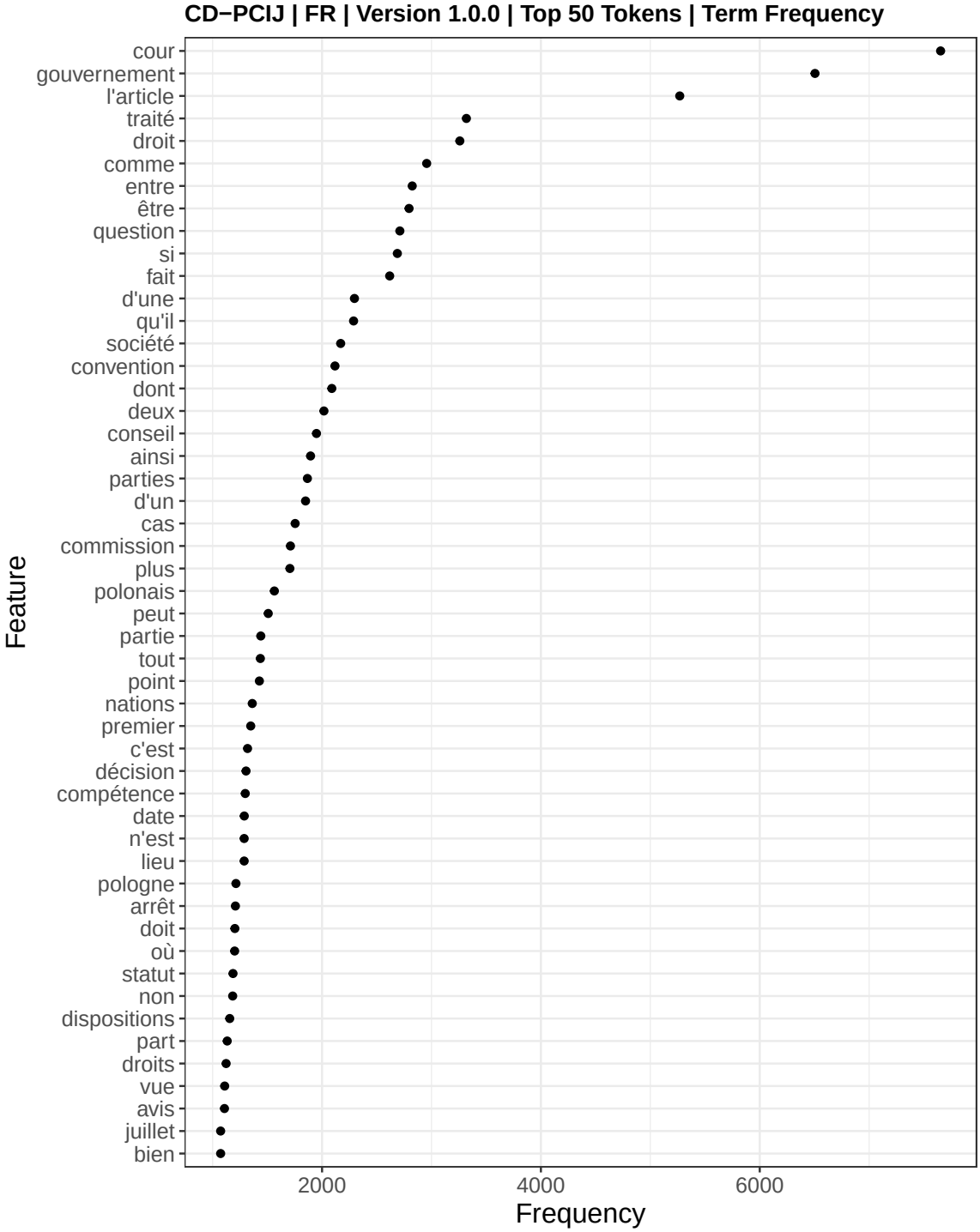
8.5.2 Term Frequency/Inverse Document Frequency Weighting (TF-IDF)



DOI: 10.5281/zenodo.3840480

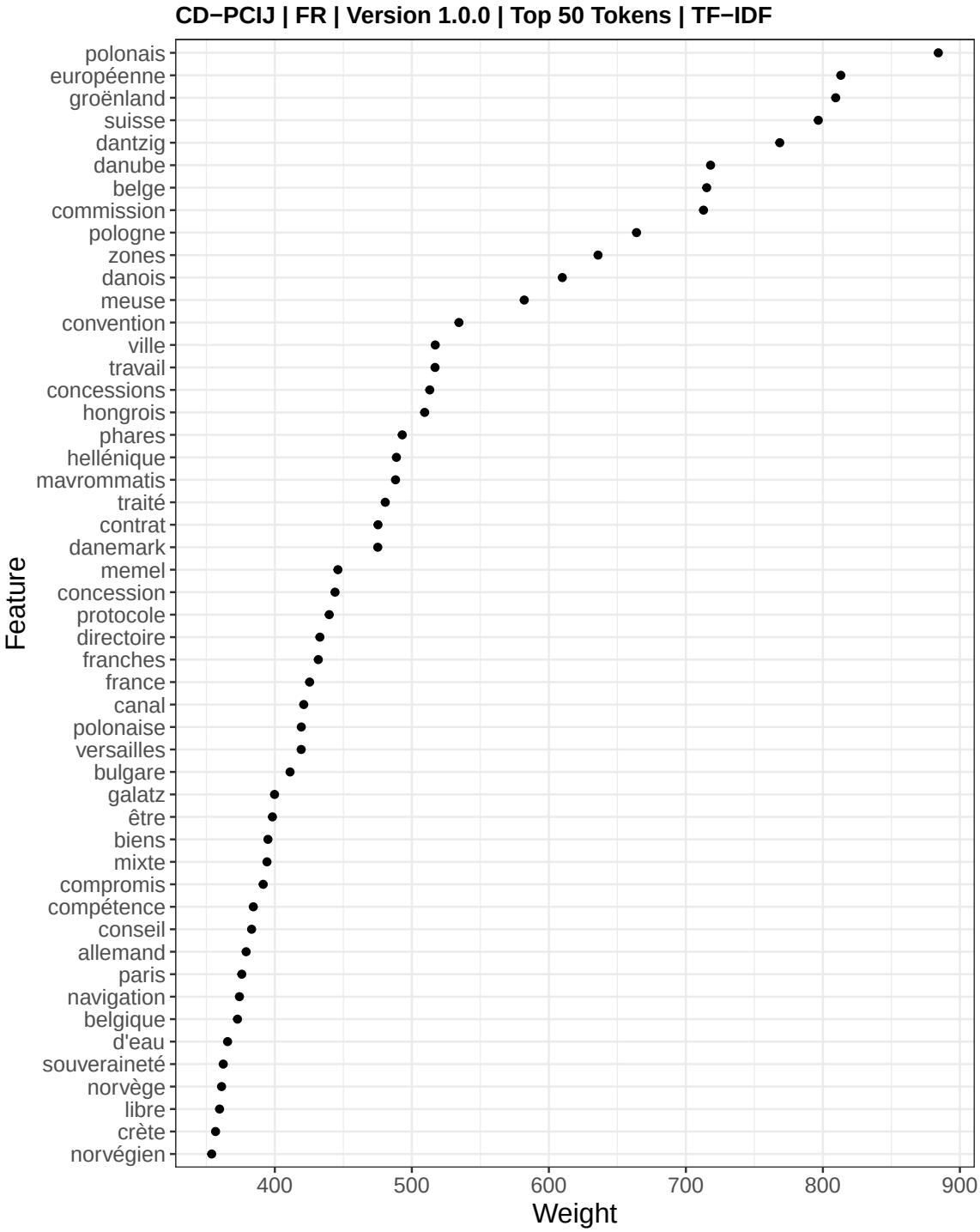
8.6 Most Frequent Tokens (French)

8.6.1 Term Frequency Weighting (TF)



DOI: 10.5281/zenodo.3840480

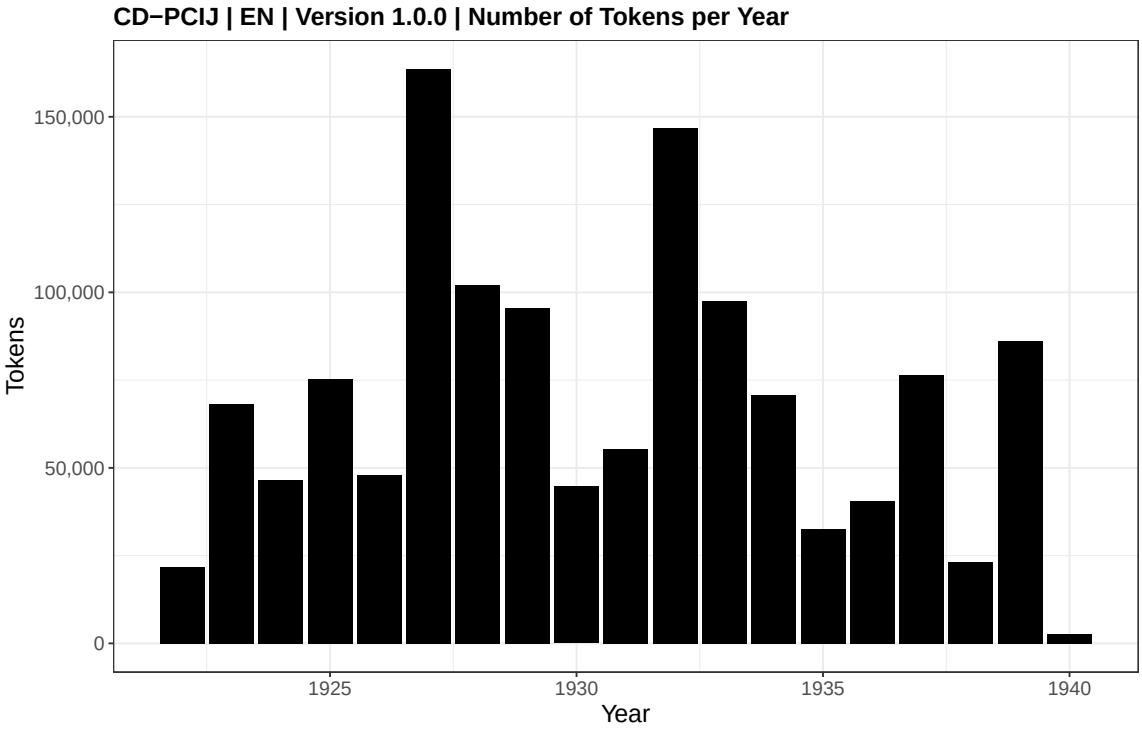
8.6.2 Term Frequency/Inverse Document Frequency Weighting (TF-IDF)



DOI: 10.5281/zenodo.3840480

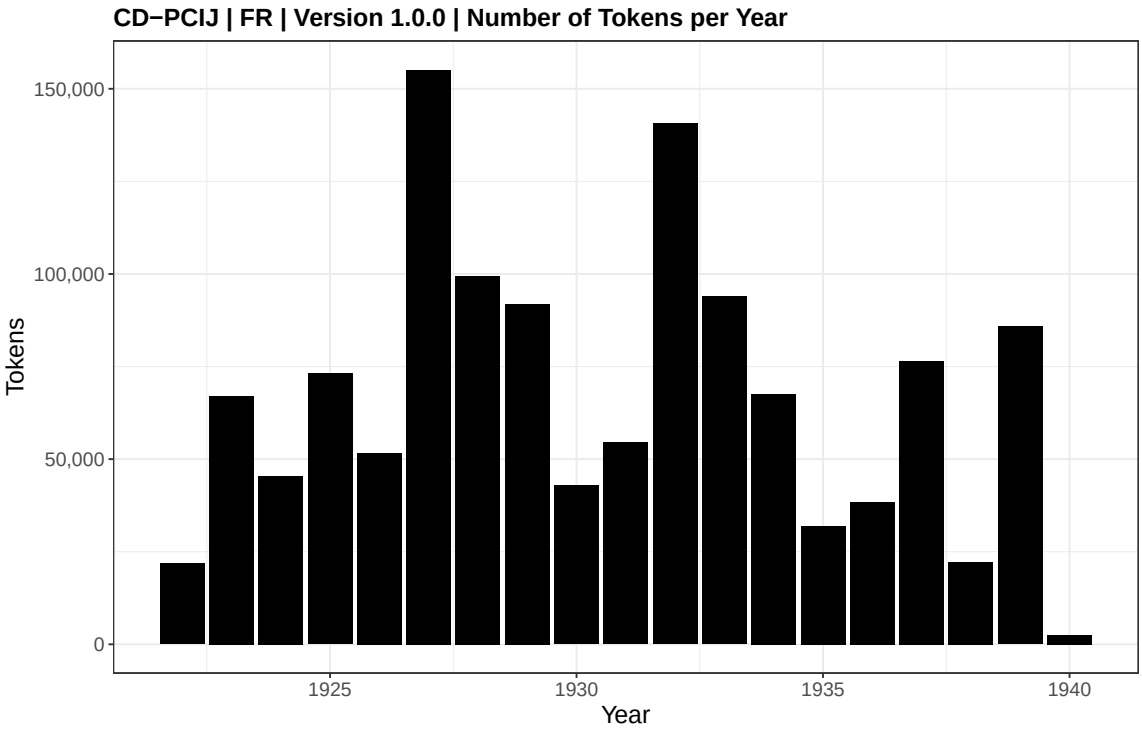
8.7 Tokens over Time

8.7.1 English



DOI: 10.5281/zenodo.3840480

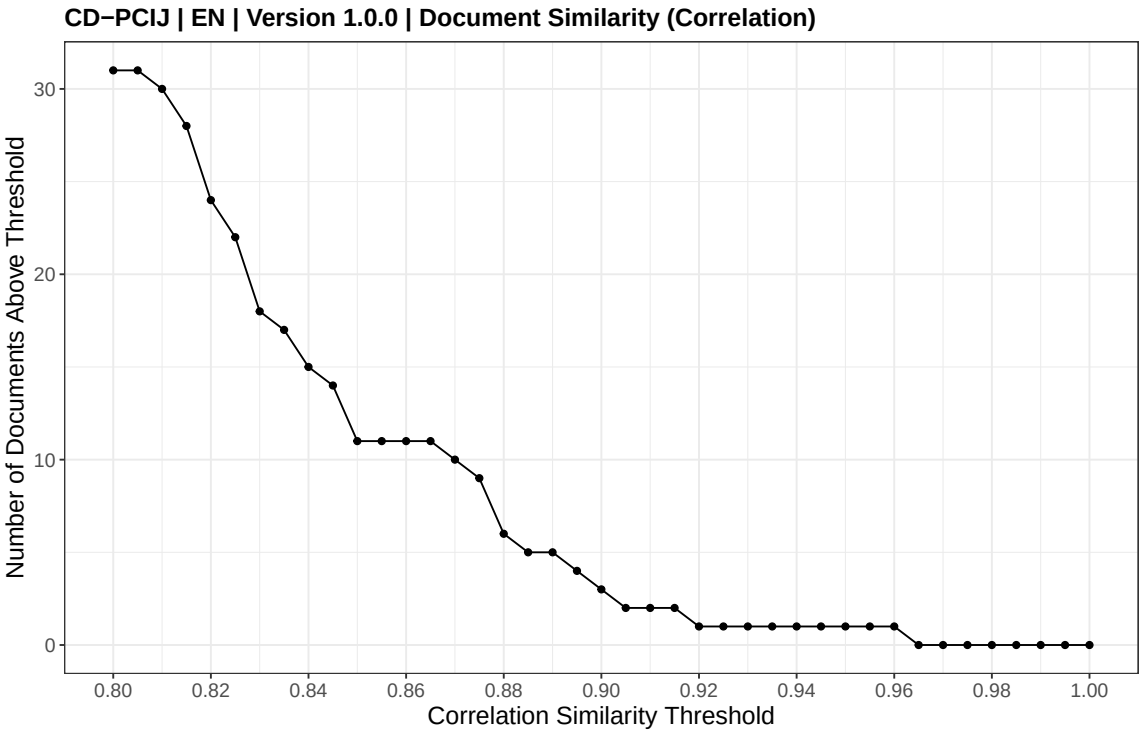
8.7.2 French



DOI: 10.5281/zenodo.3840480

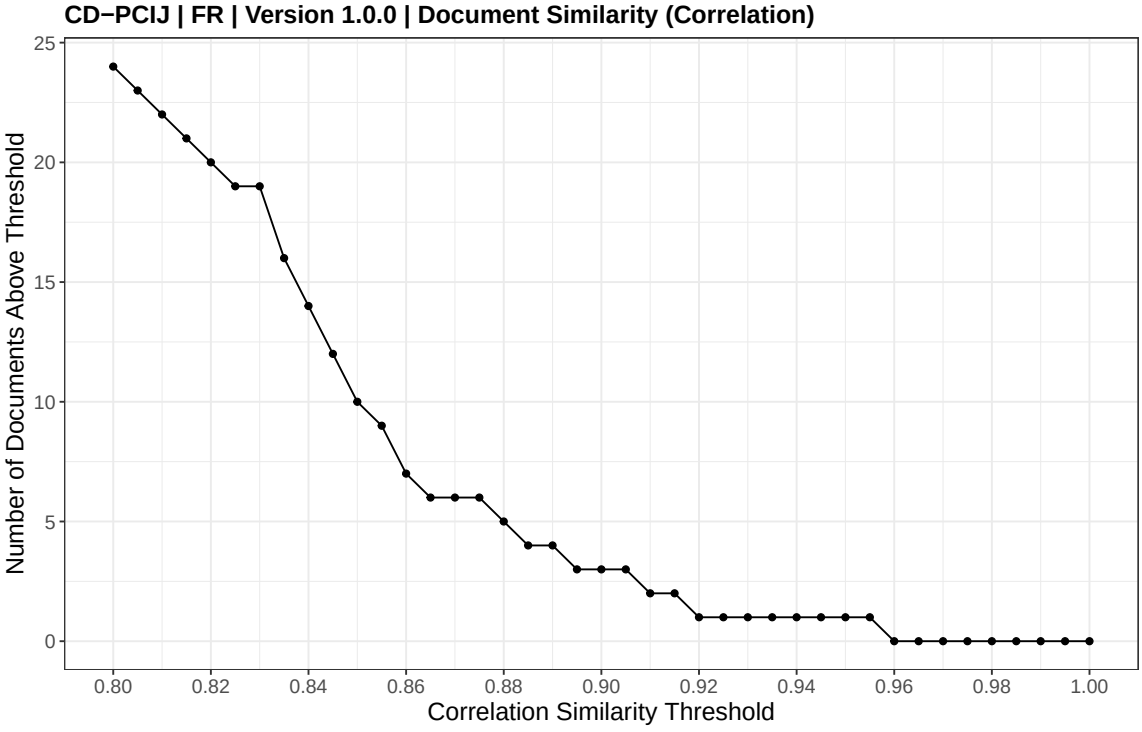
9 Document Similarity

9.1 English



DOI: 10.5281/zenodo.3840480

9.2 French



DOI: 10.5281/zenodo.3840480

9.3 Comment

Analysts generally need not be concerned with deduplicating files in the CD-PCIJ. Only two files (therefore one to drop) are similar enough to be flagged by automatic correlation similarity analysis with a threshold of 0.95. Manual inspection showed that these files differ slightly in content, but are generally identical. Analysts should make their own judgment about whether to exclude one or the other.

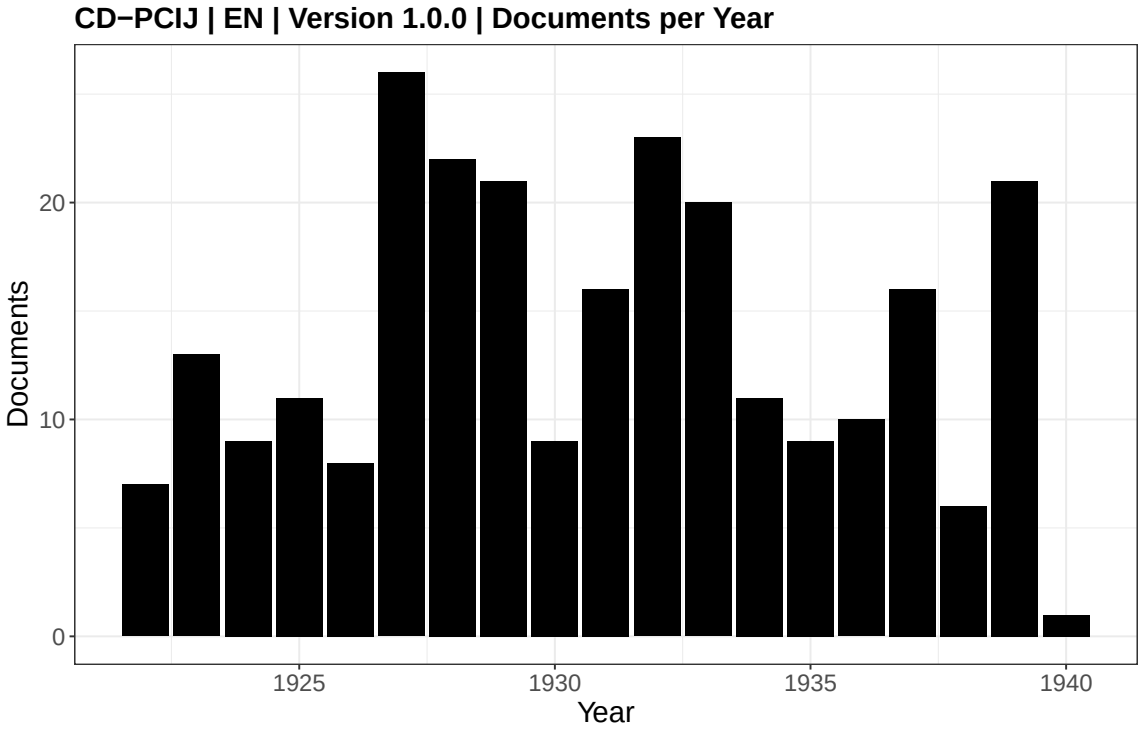
The above figures plot the number of files to be excluded as a function of correlation similarity based on a document-unigram matrix (with the removal of numbers, special symbols and stopwords, as well as lowercasing). Analysts who wish to qualitatively review this computational approach will find the IDs of presumed duplicates, together with the relevant value of correlation similarity, stored as CSV files in the 'ANALYSIS' archive published with the data set (item 17). These document IDs can also easily be read into statistical software and excluded directly from analyses without having to perform one's own similarity analysis. I do, however, recommend double-checking the IDs for false positives. The document pairings and similarity scores are included in a different CSV file (also item 17).

The choice of similarity algorithm, the threshold for marking a document as duplicate and the question of whether duplicate documents should be removed at all should be decided with respect to individual analyses. My goal is to document the Court's output as faithfully as possible and provide analysts with fair warning, as well as the opportunity to make their own choices. Please note that the manner of de-duplication will substantially affect analytical results and should be made after careful consideration of both methodology and data.

10 Metadata Frequency Tables

10.1 By Year

10.1.1 English



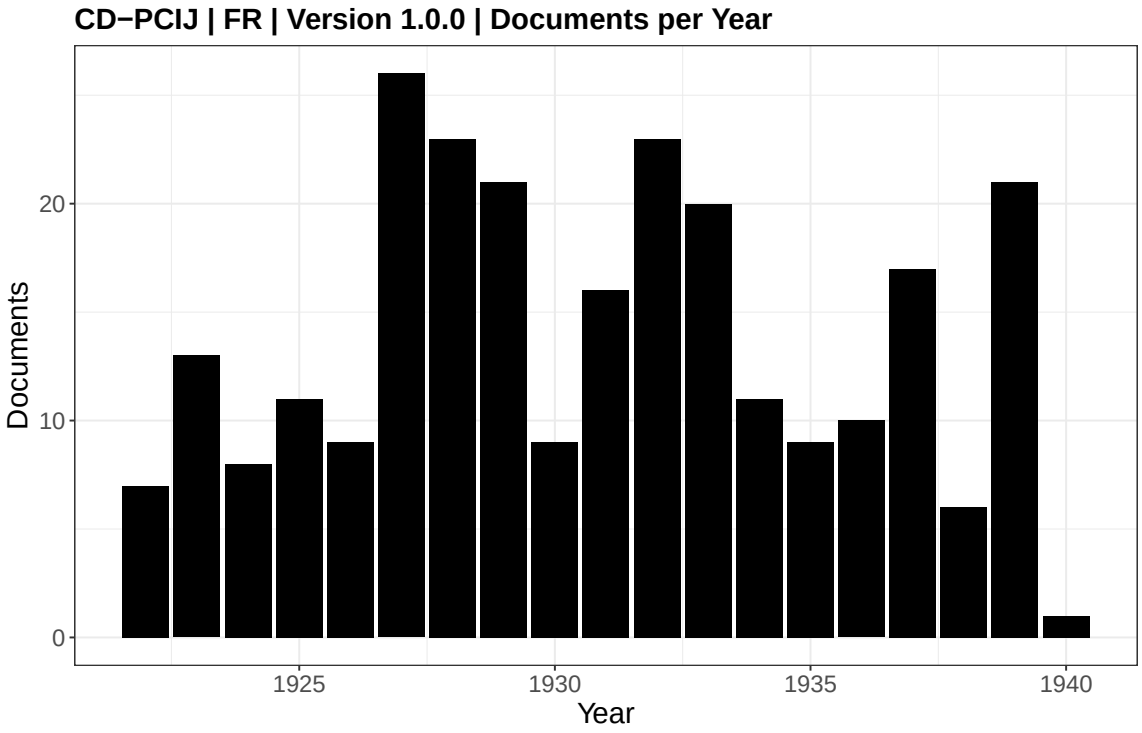
DOI: 10.5281/zenodo.3840480

Year	Documents	% Total	% Cumulative
1922	7	2.70	2.70
1923	13	5.02	7.72
1924	9	3.47	11.20
1925	11	4.25	15.44
1926	8	3.09	18.53
1927	26	10.04	28.57
1928	22	8.49	37.07
1929	21	8.11	45.17
1930	9	3.47	48.65
1931	16	6.18	54.83
1932	23	8.88	63.71
1933	20	7.72	71.43
1934	11	4.25	75.68

(continued)

Year	Documents	% Total	% Cumulative
1935	9	3.47	79.15
1936	10	3.86	83.01
1937	16	6.18	89.19
1938	6	2.32	91.51
1939	21	8.11	99.61
1940	1	0.39	100.00
Total	259	100.00	100.00

10.1.2 French



DOI: 10.5281/zenodo.3840480

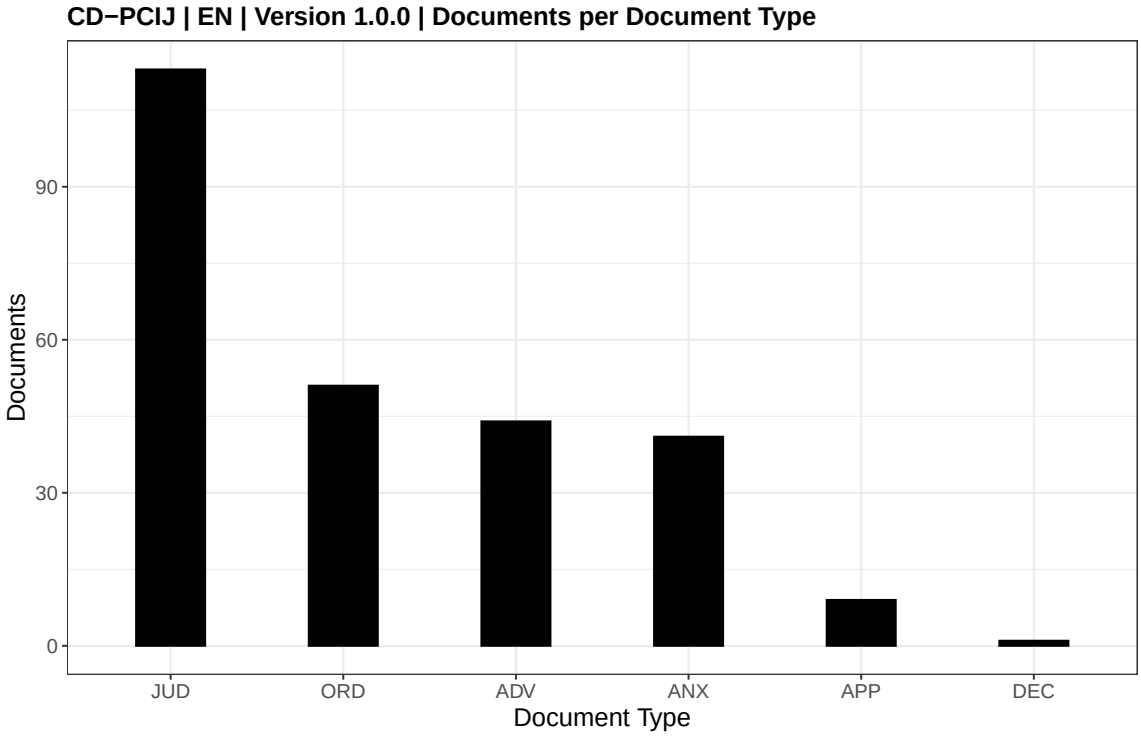
Year	Documents	% Total	% Cumulative
1922	7	2.68	2.68
1923	13	4.98	7.66
1924	8	3.07	10.73
1925	11	4.21	14.94
1926	9	3.45	18.39
1927	26	9.96	28.35
1928	23	8.81	37.16
1929	21	8.05	45.21
1930	9	3.45	48.66
1931	16	6.13	54.79
1932	23	8.81	63.60
1933	20	7.66	71.26
1934	11	4.21	75.48
1935	9	3.45	78.93
1936	10	3.83	82.76
1937	17	6.51	89.27

(continued)

Year	Documents	% Total	% Cumulative
1938	6	2.30	91.57
1939	21	8.05	99.62
1940	1	0.38	100.00
Total	261	100.00	100.00

10.2 By Document Type

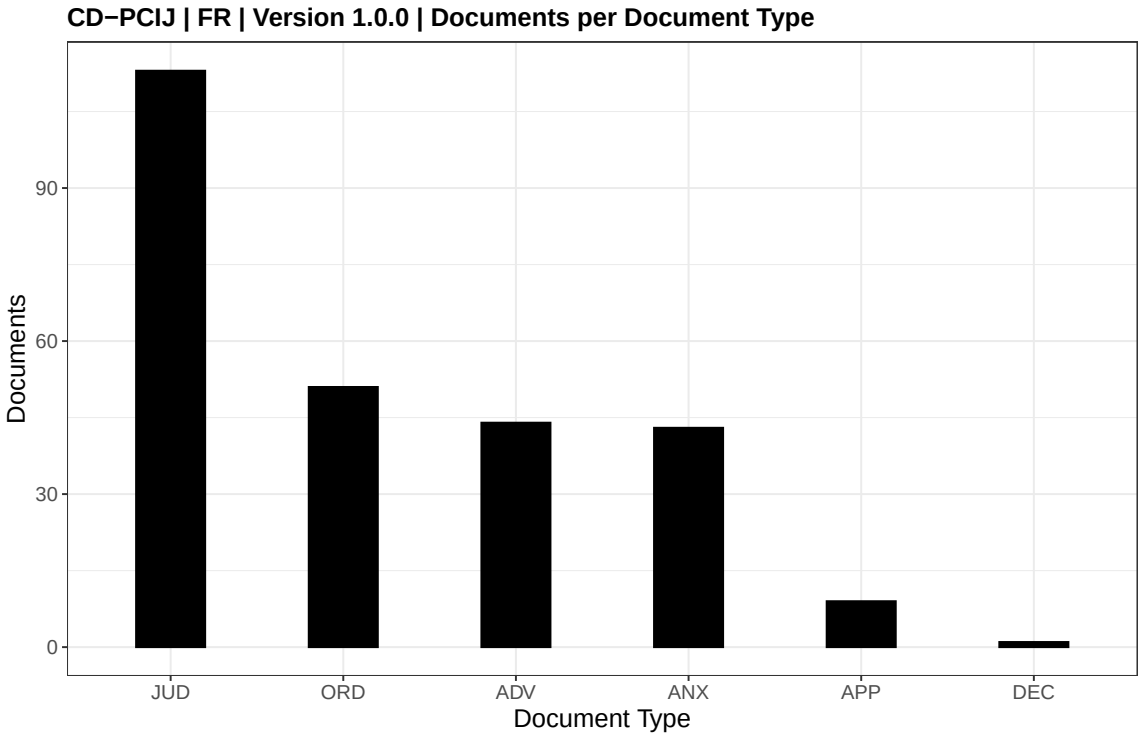
10.2.1 English



DOI: 10.5281/zenodo.3840480

DocType	Documents	% Total	% Cumulative
ADV	44	16.99	16.99
ANX	41	15.83	32.82
APP	9	3.47	36.29
DEC	1	0.39	36.68
JUD	113	43.63	80.31
ORD	51	19.69	100.00
Total	259	100.00	100.00

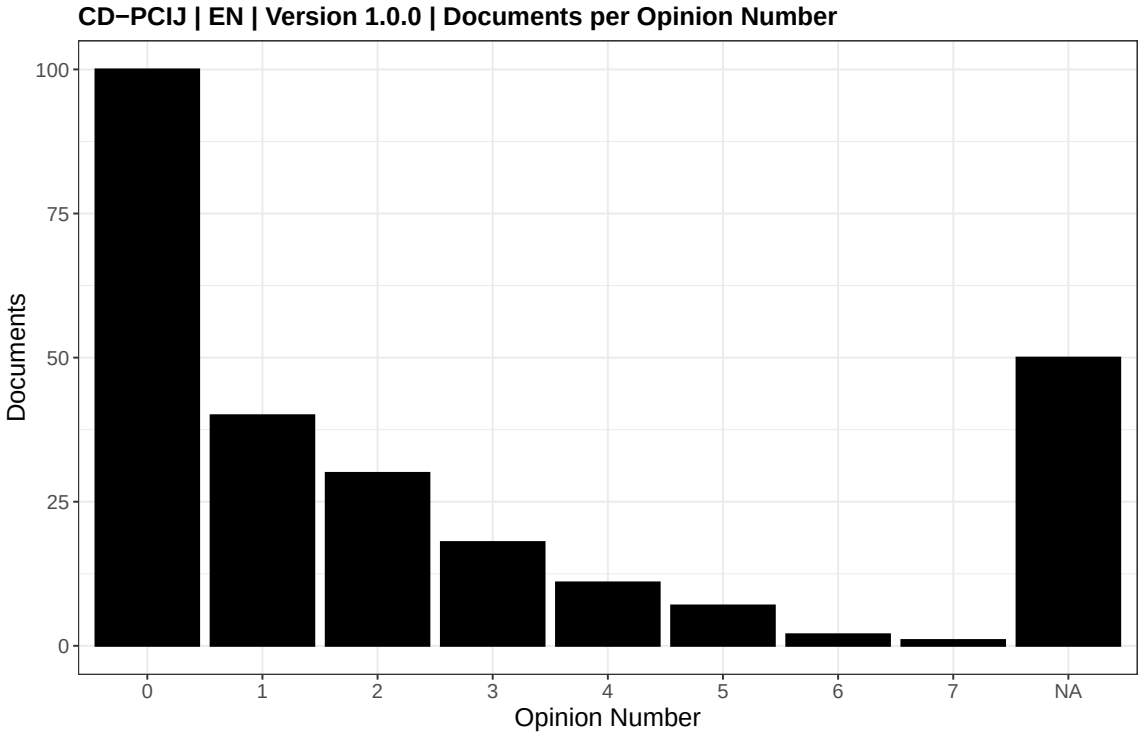
10.2.2 French



DocType	Documents	% Total	% Cumulative
ADV	44	16.86	16.86
ANX	43	16.48	33.33
APP	9	3.45	36.78
DEC	1	0.38	37.16
JUD	113	43.30	80.46
ORD	51	19.54	100.00
Total	261	100.00	100.00

10.3 By Opinion Number

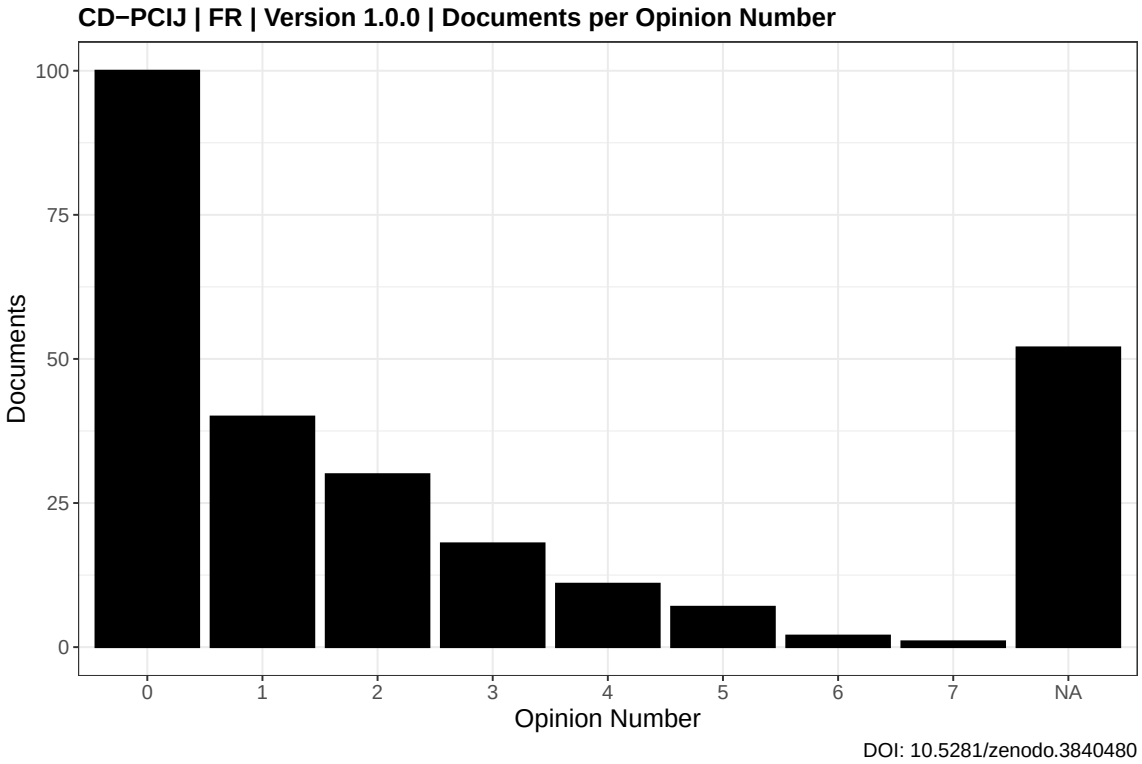
10.3.1 English



DOI: 10.5281/zenodo.3840480

Opinion Number	Documents	% Total	% Cumulative
NA	50	19.31	19.31
0	100	38.61	57.92
1	40	15.44	73.36
2	30	11.58	84.94
3	18	6.95	91.89
4	11	4.25	96.14
5	7	2.70	98.84
6	2	0.77	99.61
7	1	0.39	100.00
Total	259	100.00	100.00

10.3.2 French



Opinion Number	Documents	% Total	% Cumulative
NA	52	19.92	19.92
0	100	38.31	58.24
1	40	15.33	73.56
2	30	11.49	85.06
3	18	6.90	91.95
4	11	4.21	96.17
5	7	2.68	98.85
6	2	0.77	99.62
7	1	0.38	100.00
Total	261	100.00	100.00

10.4 By Applicant

10.4.1 English

Applicant	Documents	% Total	% Cumulative
BEL	27	10.42	10.42
BGR	3	1.16	11.58
CHE	2	0.77	12.36
CSK	4	1.54	13.90
DEU	44	16.99	30.89
DNK	7	2.70	33.59
EST	7	2.70	36.29
FRA	42	16.22	52.51
GBR	7	2.70	55.21
GBR-CSK-DNK-FRA-DEU	4	1.54	56.76
GBR-FRA-ITA-JPN	13	5.02	61.78
GRC	11	4.25	66.02
HUN	8	3.09	69.11
ITA	4	1.54	70.66
LNC	68	26.25	96.91
NLD	7	2.70	99.61
TUR	1	0.39	100.00
Total	259	100.00	100.00

10.4.2 French

Applicant	Documents	% Total	% Cumulative
BEL	27	10.34	10.34
BGR	2	0.77	11.11
CHE	2	0.77	11.88
CSK	4	1.53	13.41
DEU	45	17.24	30.65
DNK	7	2.68	33.33
EST	7	2.68	36.02
FRA	42	16.09	52.11
GBR	7	2.68	54.79
GBR-CSK-DNK-FRA-DEU	4	1.53	56.32
GBR-FRA-ITA-JPN	13	4.98	61.30
GRC	11	4.21	65.52
HUN	8	3.07	68.58
ITA	4	1.53	70.11
LNC	69	26.44	96.55
NLD	8	3.07	99.62
TUR	1	0.38	100.00
Total	261	100.00	100.00

10.5 By Respondent

10.5.1 English

Respondent	Documents	% Total	% Cumulative
NA	68	26.25	26.25
BEL	14	5.41	31.66
BGR	12	4.63	36.29
BRA	4	1.54	37.84
CHE	15	5.79	43.63
CHN	7	2.70	46.33
DEU	6	2.32	48.65
ESP	4	1.54	50.19
FRA	4	1.54	51.74
GBR	11	4.25	55.98
GRC	17	6.56	62.55
HUN	4	1.54	64.09
ITA	1	0.39	64.48
LTU	14	5.41	69.88
NOR	7	2.70	72.59
POL	48	18.53	91.12
TUR	8	3.09	94.21
YUG	15	5.79	100.00
Total	259	100.00	100.00

10.5.2 French

Respondent	Documents	% Total	% Cumulative
NA	69	26.44	26.44
BEL	15	5.75	32.18
BGR	12	4.60	36.78
BRA	4	1.53	38.31
CHE	15	5.75	44.06
CHN	7	2.68	46.74
DEU	6	2.30	49.04
ESP	4	1.53	50.57
FRA	4	1.53	52.11
GBR	11	4.21	56.32
GRC	16	6.13	62.45
HUN	4	1.53	63.98
ITA	1	0.38	64.37
LTU	14	5.36	69.73
NOR	7	2.68	72.41
POL	49	18.77	91.19
TUR	8	3.07	94.25
YUG	15	5.75	100.00
Total	261	100.00	100.00

11 Verification of Cryptographic Signatures

This Codebook automatically verifies the SHA3-512 cryptographic signatures ('hashes') of all ZIP archives during its compilation. SHA3-512 hashes are calculated via system call to the OpenSSL library on Linux systems.

A successful check is indicated by 'Signature verified!'. A failed check will print the line 'ERROR!'

```
# Define Function
sha3test <- function(filename, sig){
    sig.new <- system2("openssl",
                       paste("sha3-512", filename),
                       stdout = TRUE)
    sig.new <- gsub("^.*\\= ", "", sig.new)
    if (sig == sig.new){
        return("Signature verified!")
    }else{
        return("ERROR!")
    }
}

# Import Original Signatures
input <- fread(hashfile)
filename <- input$filename
sha3.512 <- input$sha3.512

# Verify Signatures
sha3.512.result <- mcmapply(sha3test, filename, sha3.512, USE.NAMES = FALSE)

# Print Results
testresult <- data.table(filename, sha3.512.result)

kable(testresult,
      format = "latex",
      align = c("l", "r"),
      booktabs = TRUE,
      col.names = c("File",
                    "Result"))
```

File	Result
CD-PCIJ_1-0-0_EN_CSV_TESSERACT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_EN_CSV_TESSERACT_META.zip	Signature verified!
CD-PCIJ_1-0-0_EN_PDF_ENHANCED_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_EN_PDF_ENHANCED_MajorityOpinions.zip	Signature verified!
CD-PCIJ_1-0-0_EN_PDF_ORIGINALSPLIT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_EN_TXT_EXTRACTED_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_EN_TXT_TESSERACT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_EN-FR_ANALYSIS.zip	Signature verified!
CD-PCIJ_1-0-0_FR_CSV_TESSERACT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_FR_CSV_TESSERACT_META.zip	Signature verified!
CD-PCIJ_1-0-0_FR_PDF_ENHANCED_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_FR_PDF_ENHANCED_MajorityOpinions.zip	Signature verified!
CD-PCIJ_1-0-0_FR_PDF_ORIGINALSPLIT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_FR_TXT_EXTRACTED_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_FR_TXT_TESSERACT_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_MULT_PDF_ORIGINAL_FULL.zip	Signature verified!
CD-PCIJ_1-0-0_Source_Files.zip	Signature verified!

12 Changelog

The Changelog documents changes made to the data set.

Version	Notes
1.0.0	Initial Release

13 Strict Replication Parameters

```
## [1] "OpenSSL 1.1.1l  FIPS 24 Aug 2021"
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 34 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/libflexiblas.so.3.0
##
## locale:
##  [1] LC_CTYPE=en_US.utf8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.utf8        LC_COLLATE=en_US.utf8
##  [5] LC_MONETARY=en_US.utf8    LC_MESSAGES=en_US.utf8
##  [7] LC_PAPER=en_US.utf8       LC_NAME=C
##  [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.utf8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] doParallel_1.0.16         iterators_1.0.13
##  [3] foreach_1.5.1             data.table_1.14.0
##  [5] textcat_1.0-7             quanteda.textplots_0.94
##  [7] quanteda.textstats_0.94.1 quanteda_3.1.0
##  [9] readtext_0.81             RColorBrewer_1.1-2
## [11] viridis_0.6.1             viridisLite_0.4.0
## [13] scales_1.1.1              ggplot2_3.3.5
## [15] rsvg_2.1                  DiagrammeRsvg_0.1
## [17] DiagrammeR_1.0.6.1        magick_2.7.3
## [19] kableExtra_1.3.4          knitr_1.34
## [21] fs_1.5.0                  pdftools_3.0.1
## [23] stringr_1.4.0             mgsub_1.7.3
## [25] rvest_1.0.1               httr_1.4.2
##
## loaded via a namespace (and not attached):
##  [1] jsonlite_1.7.2      RcppParallel_5.1.4  askpass_1.1
##  [4] highr_0.9           selectr_0.4-2       yaml_2.2.1
##  [7] slam_0.1-48         qpdf_1.1            pillar_1.6.2
## [10] lattice_0.20-44     glue_1.4.2          digest_0.6.27
## [13] tau_0.0-24          colorspace_2.0-2    htmltools_0.5.2
## [16] Matrix_1.3-4        pkgconfig_2.0.3     ISOcodes_2021.02.24
## [19] purrr_0.3.4         webshot_0.5.2       svglite_2.0.0
## [22] nsyllable_1.0       tibble_3.1.4        farver_2.1.0
## [25] generics_0.1.0      ellipsis_0.3.2      withr_2.4.2
## [28] magrittr_2.0.1      crayon_1.4.1        evaluate_0.14
## [31] stopwords_2.2       fansi_0.5.0         xml2_1.3.2
## [34] tools_4.0.5         lifecycle_1.0.0     V8_3.4.2
## [37] munsell_0.5.0       compiler_4.0.5      proxyC_0.2.1
## [40] tinytex_0.33        systemfonts_1.0.2   rlang_0.4.11
```

```
## [43] grid_4.0.5           rstudioapi_0.13     htmlwidgets_1.5.4
## [46] visNetwork_2.0.9     labeling_0.4.2      rmarkdown_2.10
## [49] gtable_0.3.0         codetools_0.2-18    curl_4.3.2
## [52] R6_2.5.1             gridExtra_2.3       dplyr_1.0.7
## [55] fastmap_1.1.0        utf8_1.2.2          fastmatch_1.1-3
## [58] stringi_1.7.4        Rcpp_1.0.7          vctrs_0.3.8
## [61] tidyselect_1.1.1     xfun_0.25
```

References

Analytics, Revolution, and Steve Weston. 2020. *Iterators: Provides Iterator Construct.* https://github.com/RevolutionAnalytics/iterators.

Benoit, Kenneth, and Adam Obeng. 2021. *Readtext: Import and Handling for Plain and Formatted Text Files.* https://github.com/quanteda/readtext.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Jiong Wei Lua, and Jouni Kuha. 2021. *Quanteda.textstats: Textual Statistics for the Quantitative Analysis of Textual Data.* https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018a. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

———. 2018b. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

———. 2018c. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774. https://doi.org/10.21105/joss.00774.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, and William Lowe. 2021. *Quanteda: Quantitative Analysis of Textual Data.* https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2021. *Quanteda.textplots: Plots for the Quantitative Analysis of Textual Data.* https://CRAN.R-project.org/package=quanteda.textplots.

Corporation, Microsoft, and Steve Weston. 2020. *DoParallel: Foreach Parallel Adaptor for the Parallel Package.* https://CRAN.R-project.org/package=doParallel.

Dowle, Matt, and Arun Srinivasan. 2021. *Data.table: Extension of 'Data.frame'.* https://CRAN.R-project.org/package=data.table.

Ewing, Mark. 2021. *Mgsub: Safe, Multiple, Simultaneous String Substitution.* https://CRAN.R-project.org/package=mgsub.

Garnier, Simon. 2021a. *Viridis: Colorblind-Friendly Color Maps for R.* https://CRAN.R-project.org/package=viridis.

———. 2021b. *ViridisLite: Colorblind-Friendly Color Maps (Lite Version).* https://CRAN.R-project.org/package=viridisLite.

Hester, Jim, and Hadley Wickham. 2020. *Fs: Cross-Platform File System Operations Based on Libuv.* https://CRAN.R-project.org/package=fs.

Hornik, Kurt, Patrick Mair, Johannes Rauch, Wilhelm Geiger, Christian Buchta, and Ingo Feinerer. 2013. "The textcat Package for n-Gram Based Text Categorization in R." *Journal of Statistical Software* 52 (6): 1–17. https://doi.org/10.18637/jss.v052.i06.

Hornik, Kurt, Johannes Rauch, Christian Buchta, and Ingo Feinerer. 2020. *Textcat: N-Gram Based Text Categorization.* https://CRAN.R-project.org/package=textcat.

Iannone, Richard. 2016. *DiagrammeRsvg: Export Diagrammer Graphviz Graphs as Svg*. https://github.com/rich-iannone/DiagrammeRsvg.

———. 2020. *DiagrammeR: Graph/Network Visualization*. https://github.com/rich-iannone/DiagrammeR.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

Ooms, Jeroen. 2020. *Rsvg: Render Svg Images into Pdf, Png, Postscript, or Bitmap Arrays*. https://github.com/jeroen/rsvg#readme.

———. 2021a. *Magick: Advanced Graphics and Image-Processing in R*. https://CRAN.R-project.org/package=magick.

———. 2021b. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revolution Analytics, and Steve Weston. n.d. *Foreach: Provides Foreach Looping Construct*.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2020. *Httr: Tools for Working with Urls and Http*. https://CRAN.R-project.org/package=httr.

———. 2021. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.