Corpus of Decisions

International Court of Justice

(CD-ICJ)

Codebook

Version 2023-10-22

DOI: 10.5281/zenodo.10030647

Title	Corpus of Decisions: International Court of Justice
Abbreviation	CD-ICJ
Author	Seán Fobbe
Version	2023-10-22
Download	https://doi.org/10.5281/zenodo.10030647
License	CC0 1.0 Universal

Citation

Seán Fobbe (2023). Corpus of Decisions: International Court of Justice (CD-ICJ). Version 2023-10-22. Zenodo. DOI: 10.5281/zenodo.10030647.

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 2023-10-22. 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 **International Court of Justice (ICJ)** is the primary judicial organ of the United Nations and one of the most consequential courts in international law.

Called the 'World Court' by many, it is the only international court with general thematic jurisdiction. While critics occasionally note the lack of compulsory jurisdiction and sharply limited access to the Court,[1] its opinions continue to have an outsize influence on the modern interpretation, codification and wider development of international law. Every international legal textbook covers the workings and decisions of the Court *in extenso* and participation in international moot courts, such as the Philip C. Jessup Moot Court, without regular reference to and citation of the International Court of Justice's decisions, is unthinkable.

The **Corpus of Decisions: International Court of Justice (CD-ICJ)** collects and presents for the first time in human- and machine-readable form all published decisions of the International Court of Justice. Among these are judgments, advisory opinions and orders, as well as their respective appended minority opinions (declarations, separate opinions and dissenting opinions).

This data set is designed to be complementary to and fully compatible with the *Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ)*, which is also available open access.[2]

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.[3]

[1] Only States may be party to proceedings in contentious jurisdiction and only certain bodies of international organizations may request advisory opinions.

[2] Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ). <https://doi.org/10.5281/zenodo.3840480>.

[3] 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[4] 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)
icj.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)
icj.en <- readtext("EN_TXT_BEST_FULL/*.txt",
                   docvarsfrom = "filenames",
                   docvarnames = c("court",
                                   "caseno",
                                   "shortname",
                                   "applicant",
                                   "respondent",
                                   "date",
                                   "doctype",
                                   "collision",
                                   "opinion",
                                   "language"),
                   dvsep = "_",
                   encoding = "UTF-8")
```

[4] 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: International Court of Justice (CD-ICJ)** collects and structures in human- and machine-readable form all published decisions of the International Court of Justice. Among these are judgments, advisory opinions and orders, as well as their respective appended minority opinions (declarations, separate opinions and dissenting opinions).

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: Permanent Court of International Justice (CD-PCIJ)*, which is also available open access.[5]

3.3 Table of Sources

Data Source	Citation
Primary Data Source	https://www.icj-cij.org
Source Code	https://doi.org/10.5281/zenodo.10030648
Country Codes	https://doi.org/10.5281/zenodo.10030648
Entity Codes	https://doi.org/10.5281/zenodo.10030648
Cases Names and Parties	https://doi.org/10.5281/zenodo.10030648

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 International Court of Justice that were published on the official website of the International Court of Justice on 2023-10-22.

[5] Corpus of Decisions: Permanent Court of International Justice (CD-PCIJ). <https://doi.org/10.5281/zenodo.3840480>.

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 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 ICJ 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. Automatic language detection is not foolproof and some bilingual documents marked as monolingual may have gone undetected (*language mismatch*).
4. Lengthy quotations in languages other than the language indicated in the metadata may further 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.[6]

[6] https://creativecommons.org/publicdomain/zero/1.0/legalcode

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.

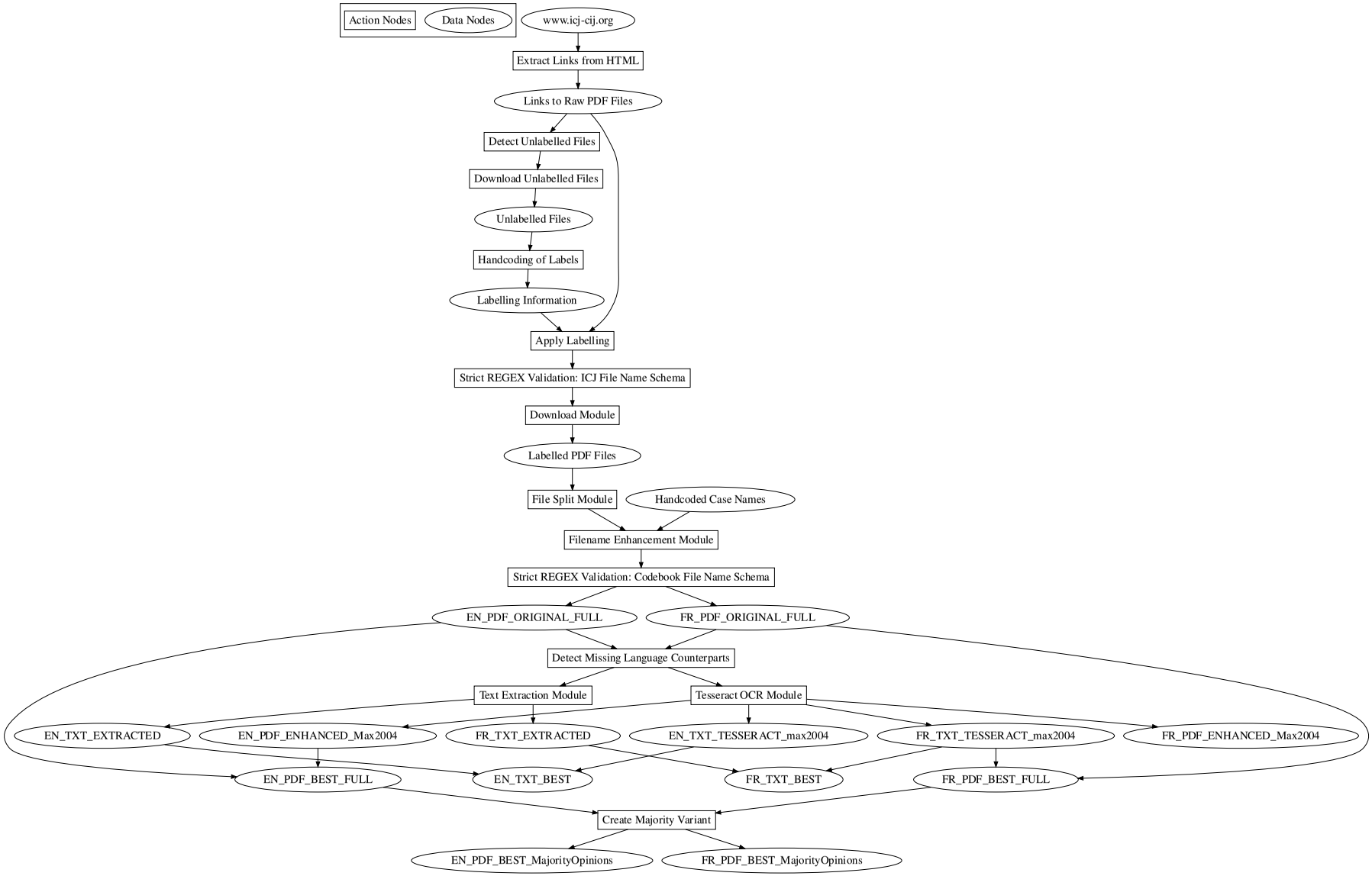


Figure 1: Workflow Schematic Part 1: Download, Labelling, Conversion and Sorting of Documents

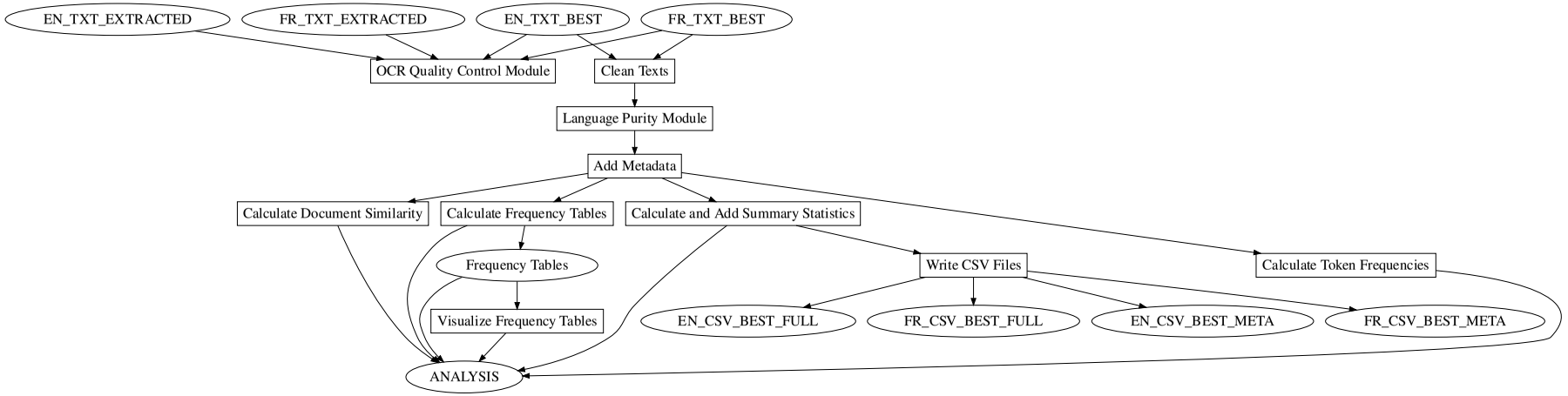


Figure 2: Workflow Schematic Part 2: Ingestion, Pre-Processing, Analysis and Creation of CSV Files

4 Variants and Primary Target Audiences

The data set is provided in two 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_BEST	**Traditional Legal Research (recommended).** A synthesis of all born-digital documents issued by the ICJ combined with older scanned documents (prior to 2005) which were given a new and enhanced text layer created with an advanced 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. Researchers with slow internet connections should consider using the 'TXT_BEST' 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_BEST	**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_BEST	**Quantitative Research.** A synthesis of TXT files created by combining the extracted text of all born-digital documents issued by the ICJ (2005 and later) and the OCR texts from older scanned documents (prior to 2005) generated with an advanced LSTM neural network machine learning engine. 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.

Variant	Target Audience and Description
TXT_EXTRACTED	**Replication Research and Creation of New Data Sets.** TXT files containing the extracted text layer from all original documents as published by the ICJ. The quality of the original OCR text for older documents is poor and this variant should not be used for statistical analysis. Documents dated 2005 or later were born-digital and can be used for all purposes.
TXT_TESSERACT	**Replication Research and Creation of New Data Sets.** TXT files containing the OCR text generated with an advanced LSTM neural network machine learning engine for documents predating 2005. Fully included in the BEST variant, but provided separately for reasons of transparency.
PDF_ORIGINAL	**Replication Research and Creation of New Data Sets.** The original documents with the original text layer. Only recommended for researchers who wish to replicate the machine-readable files or who wish to create a new and improved data set. Not recommended for traditional research, as the quality of the original OCR text layer is quite poor.
PDF_ENHANCED	**Replication Research and Creation of New Data Sets.** Scanned documents of opinions rendered before 2005 which were given a new and enhanced text layer generated with an advanced LSTM neural network machine learning engine. Fully included in the BEST variant, but provided separately for reasons of transparency.

5 Variables

5.1 General Remarks

- Missing values are always coded as 'NA'.
- All Strings are encoded in UTF-8.
- A significant part of the metadata was included with the files downloaded from the Court's website.
- The variables 'shortname', 'applicant', 'respondent', 'stage', 'applicant_region', 'applicant_subregion', 'respondent_region' and 'respondent_subregion' were coded manually by the author of the data set and 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. Includes the full names of the parties 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]_[caseno]_[shortname]_[applicant]_[respondent]_[date]_[doctype]_[collision]_[stage]_[opinion]_[language]`

5.3 Example TXT File Name

`ICJ_001_CorfuChannel_GBR_ALB_1949-04-09_JUD_01_ME_05_EN.txt`

5.4 Structure of CSV Metadata

```
## Classes 'data.table' and 'data.frame':   2289 obs. of  27 variables:
##  $ doc_id              : chr  "ICJ_001_CorfuChannel_GBR_ALB_1947-07-31_ORD_01_
    NA_00_EN.txt" "ICJ_001_CorfuChannel_GBR_ALB_1947-12-10_ORD_01_NA_00_EN.txt" "
    ICJ_001_CorfuChannel_GBR_ALB_1948-03-25_JUD_01_PO_00_EN.txt" "ICJ_001_
    CorfuChannel_GBR_ALB_1948-03-25_JUD_01_PO_01_EN.txt" ...
##  $ court               : chr  "ICJ" "ICJ" "ICJ" "ICJ" ...
##  $ caseno              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ shortname           : chr  "CorfuChannel" "CorfuChannel" "CorfuChannel" "
    CorfuChannel" ...
##  $ fullname            : chr  "Corfu Channel (United Kingdom of Great Britain
    and Northern Ireland v. Albania)" "Corfu Channel (United Kingdom of Great
    Britain and Northern Ireland v. Albania)" "Corfu Channel (United Kingdom of
    Great Britain and Northern Ireland v. Albania)" "Corfu Channel (United
    Kingdom of Great Britain and Northern Ireland v. Albania)" ...
##  $ applicant           : chr  "GBR" "GBR" "GBR" "GBR" ...
##  $ respondent          : chr  "ALB" "ALB" "ALB" "ALB" ...
##  $ applicant_region    : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ respondent_region   : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ applicant_subregion : chr  "Northern Europe" "Northern Europe" "Northern
    Europe" "Northern Europe" ...
##  $ respondent_subregion: chr  "Southern Europe" "Southern Europe" "Southern
    Europe" "Southern Europe" ...
##  $ date                : IDate, format: "1947-07-31" "1947-12-10" ...
##  $ doctype             : chr  "ORD" "ORD" "JUD" "JUD" ...
##  $ collision           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stage               : chr  NA NA "PO" "PO" ...
##  $ opinion             : int  0 0 0 1 2 0 0 0 1 2 ...
##  $ language            : chr  "EN" "EN" "EN" "EN" ...
##  $ year                : int  1947 1947 1948 1948 1948 1948 1948 1949 1949
    1949 ...
##  $ minority            : int  0 0 0 1 1 0 0 0 1 1 ...
##  $ nchars              : int  5782 2435 48544 3883 34300 6148 7846 172995
    25052 23145 ...
##  $ ntokens             : int  1141 464 9046 744 6633 1169 1602 33912 4750 4310
     ...
##  $ ntypes              : int  368 175 1353 284 1037 373 516 3674 1037 1026 ...
##  $ nsentences          : int  45 12 272 28 209 32 53 1254 161 118 ...
##  $ version             : IDate, format: "2023-10-22" "2023-10-22" ...
##  $ doi_concept         : chr  "10.5281/zenodo.3826444" "10.5281/zenodo
    .3826444" "10.5281/zenodo.3826444" "10.5281/zenodo.3826444" ...
##  $ doi_version         : chr  "10.5281/zenodo.10030647" "10.5281/zenodo
    .10030647" "10.5281/zenodo.10030647" "10.5281/zenodo.10030647" ...
##  $ 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 'ICJ', which stands for 'International Court of Justice'. It is generally only useful if combined with the CD-PCIJ or other data sets.
caseno	Integer	The case number assigned by the ICJ. The same case may span multiple case numbers, i.e. the Interpretation or Revision stages have different case numbers than the original judgment. To analyze all stages of a case I recommend a pattern search on the variable 'shortname'. Note: case number 2 is unassigned and there are no documents for case number 2 available on the ICJ website.
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. 'Nicaragua') to facilitate quick local searches and try to be as faithful to the full title as possible. For requests concerning interpretation or revision of a judgment the shortname is followed by 'Interpretation' or 'Revision'.
fullname	String	(CSV only) Full name of the case as published on the ICJ website. Includes the full names of the Parties. Introductory phrases such as 'Case concerning...' are omitted.
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. the Soviet Union). For advisory proceedings this variable refers to the entity which requested an advisory opinion. Table 6.2 explains the detailed advisory coding decisions.

Variable	Type	Details
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 UN bodies or 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 UN bodies or 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 UN bodies or 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 UN bodies or international organizations.
date	ISO Date	The date of the document in the format YYYY-MM-DD (ISO-8601).
doctype	String	A three-letter code indicating the type of document. Possible values are 'JUD' (judgments in contentious jurisdiction), 'ADV' (advisory opinions) and 'ORD' (orders in all types of jurisidiction).
collision	Integer	In rare instances the International Court of Justice issued several decisions of the same type in the same proceedings on the same day. Most documents take the value '01'. If documents with otherwise identical metadata would be issued, the value is incremented.

Variable	Type	Details
stage	String	The stage of proceedings in contentious jurisdiction, coded based on the title page (primary), or a close reading of the findings (secondary). Possible values are 'PO' (preliminary objections), 'ME' (merits), 'IN' (intervention) and 'CO' (compensation). Please note that the ICJ is very inconsistent in how it classifies admissibility; it can occur in the same document either together with a decision on jurisdiction or a decision on the merits. I have chosen to code pure admissibility decisions as 'ME' (e.g. Second Phase of Nottebohm). In general all of the above types of decisions can occur in the same document. I therefore do not recommend this variable for computational analysis unless great care is taken to understand its limitations. Currently only judgments are coded, orders will be added in the future.
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.
language	String	The language of the document as a two-letter ISO 639-1 code. This data set contains documents in the languages English ('EN') and French ('FR').
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.
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.

Variable	Type	Details
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	ISO Date	(CSV only) The version of the data set as a date in long form as per ISO-8601. The version represents the date on which the data set creation process was begun and the data was acquired from the website of the Court.
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.

Variable	Type	Details
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 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 (Soviet Union, Yugoslavia, Serbia and Montenegro, 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). The Soviet Union is coded as 'Europe/Eastern Europe' (the M49 standard considers the SUN its own region). Serbia and Montenegro was never included in the M49 standard and has been assigned the same region and subregion as Yugoslavia.

ISO-3	Name	Region	Sub-Region
ALB	Albania	Europe	Southern Europe
ARE	United Arab Emirates	Asia	Western Asia
ARG	Argentina	Americas	Latin America and the Caribbean
ARM	Armenia	Asia	Western Asia
AUS	Australia	Oceania	Australia and New Zealand
AZE	Azerbaijan	Asia	Western Asia
BDI	Burundi	Africa	Sub-Saharan Africa
BEL	Belgium	Europe	Western Europe
BEN	Benin	Africa	Sub-Saharan Africa
BFA	Burkina Faso	Africa	Sub-Saharan Africa
BGR	Bulgaria	Europe	Eastern Europe
BHR	Bahrain	Asia	Western Asia
BIH	Bosnia and Herzegovina	Europe	Southern Europe
BLZ	Belize	Americas	Latin America and the Caribbean
BOL	Bolivia	Americas	Latin America and the Caribbean
BRA	Brazil	Americas	Latin America and the Caribbean
BWA	Botswana	Africa	Sub-Saharan Africa
CAN	Canada	Americas	Northern America
CHE	Switzerland	Europe	Western Europe
CHL	Chile	Americas	Latin America and the Caribbean
CMR	Cameroon	Africa	Sub-Saharan Africa

(continued)

ISO-3	Name	Region	Sub-Region
COD	Democratic Republic of the Congo	Africa	Sub-Saharan Africa
COL	Colombia	Americas	Latin America and the Caribbean
CRI	Costa Rica	Americas	Latin America and the Caribbean
CSK	Czechoslovakia	Europe	Eastern Europe
DEU	Germany	Europe	Western Europe
DJI	Djibouti	Africa	Sub-Saharan Africa
DMA	Dominica	Americas	Latin America and the Caribbean
DNK	Denmark	Europe	Northern Europe
ECU	Ecuador	Americas	Latin America and the Caribbean
EGY	Egypt	Africa	Northern Africa
ESP	Spain	Europe	Southern Europe
ETH	Ethiopia	Africa	Sub-Saharan Africa
FIN	Finland	Europe	Northern Europe
FRA	France	Europe	Western Europe
GAB	Gabon	Africa	Sub-Saharan Africa
GBR	United Kingdom	Europe	Northern Europe
GEO	Georgia	Asia	Western Asia
GIN	Guinea	Africa	Sub-Saharan Africa
GMB	Gambia	Africa	Sub-Saharan Africa
GNB	Guinea-Bissau	Africa	Sub-Saharan Africa
GNQ	Equatorial Guinea	Africa	Sub-Saharan Africa
GRC	Greece	Europe	Southern Europe
GTM	Guatemala	Americas	Latin America and the Caribbean
GUY	Guyana	Americas	Latin America and the Caribbean
HND	Honduras	Americas	Latin America and the Caribbean
HRV	Croatia	Europe	Southern Europe
HUN	Hungary	Europe	Eastern Europe
IDN	Indonesia	Asia	South-eastern Asia
IND	India	Asia	Southern Asia
IRN	Iran	Asia	Southern Asia
ISL	Iceland	Europe	Northern Europe
ISR	Israel	Asia	Western Asia
ITA	Italy	Europe	Southern Europe
JPN	Japan	Asia	Eastern Asia
KEN	Kenia	Africa	Sub-Saharan Africa

(continued)

ISO-3	Name	Region	Sub-Region
KHM	Cambodia	Asia	South-eastern Asia
LBN	Lebanon	Asia	Western Asia
LBR	Liberia	Africa	Sub-Saharan Africa
LBY	Libya	Africa	Northern Africa
LIE	Liechtenstein	Europe	Western Europe
MEX	Mexico	Americas	Latin America and the Caribbean
MHL	Marshall Islands	Oceania	Micronesia
MKD	North Macedonia	Europe	Southern Europe
MLI	Mali	Africa	Sub-Saharan Africa
MLT	Malta	Europe	Southern Europe
MMR	Myanmar	Asia	South-eastern Asia
MYS	Malaysia	Asia	South-eastern Asia
NAM	Namibia	Africa	Sub-Saharan Africa
NER	Niger	Africa	Sub-Saharan Africa
NGA	Nigeria	Africa	Sub-Saharan Africa
NIC	Nicaragua	Americas	Latin America and the Caribbean
NLD	Netherlands	Europe	Western Europe
NOR	Norway	Europe	Northern Europe
NRU	Nauru	Oceania	Micronesia
NZL	New Zealand	Oceania	Australia and New Zealand
PAK	Pakistan	Asia	Southern Asia
PER	Peru	Americas	Latin America and the Caribbean
PRT	Portugal	Europe	Southern Europe
PRY	Paraguay	Americas	Latin America and the Caribbean
PSE	Palestine	Asia	Western Asia
QAT	Qatar	Asia	Western Asia
ROU	Romania	Europe	Eastern Europe
RUS	Russia	Europe	Eastern Europe
RWA	Rwanda	Africa	Sub-Saharan Africa
SAU	Saudi-Arabia	Asia	Western Asia
SCG	Serbia and Montenegro	Europe	Southern Europe
SEN	Senegal	Africa	Sub-Saharan Africa
SGP	Singapore	Asia	South-eastern Asia
SLV	El Salvador	Americas	Latin America and the Caribbean
SOM	Somalia	Africa	Sub-Saharan Africa
SRB	Serbia	Europe	Southern Europe

(continued)

ISO-3	Name	Region	Sub-Region
SUN	Soviet Union	Europe	Eastern Europe
SVK	Slovakia	Europe	Eastern Europe
SWE	Sweden	Europe	Northern Europe
SYR	Syrian Arab Republic	Asia	Western Asia
TCD	Chad	Africa	Sub-Saharan Africa
THA	Thailand	Asia	South-eastern Asia
TLS	Timor Leste	Asia	South-eastern Asia
TUN	Tunisia	Africa	Northern Africa
TUR	Turkey	Asia	Western Asia
UGA	Uganda	Africa	Sub-Saharan Africa
UKR	Ukraine	Europe	Eastern Europe
URY	Uruguay	Americas	Latin America and the Caribbean
USA	United States of America	Americas	Northern America
VEN	Venezuela	Americas	Latin America and the Caribbean
YUG	Yugoslavia	Europe	Southern Europe
ZAF	South Africa	Africa	Sub-Saharan Africa

6.2 Advisory Jurisdiction: Entities

Entities who requested an advisory opinion from the International Court of Justice are not Applicants in the strict sense, but have been coded under this variable to reduce clutter. I have tried to choose widely used codes for each entity.

Note that the *International Maritime Organization (IMO)* was known as the 'Inter-Governmental Maritime Consultative Organization' at the time it requested the advisory opinon. I have coded it with the modern 'IMO', as the organization only underwent a change of name and its legal continuity is not in doubt.

I was unable to discover a well-known acronym for the *Committee on Applications for Review of Administrative Tribunal Judgements* and custom-coded it as 'CARAT'.

Code	Entity
CARAT	Committee on Applications for Review of Administrative Tribunal Judgements
ECOSOC	UN Economic and Social Council
IFAD	International Fund for Agricultural Development
IMO	Inter-Governmental Maritime Consultative Organization
UNECO	United Nations Educational, Scientific and Cultural Organization
UNGA	UN General Assembly
UNSC	United Nations Security Council
WHO	World Health Organization

7 Linguistic Metrics

7.1 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'.

7.2 Summary Statistics

7.2.1 English

Metric	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	89,402,566	373	4,532	16,114	39,057.48	44,391	744,487
ntokens	15,767,521	71	763	2,836	6,888.39	8,043	142,528
ntypes	89,506	53	290	705	1,039.21	1,392	9,950
nsentences	538,741	1	20	94	235.36	265	5,645

7.2.2 French

Metric	Total	Min	Quart1	Median	Mean	Quart3	Max
nchars	95,209,156	396	4,639.25	16,983.0	41,831.79	47,728.00	817,862
ntokens	16,239,787	69	788.75	2,933.5	7,135.23	8,395.75	148,455
ntypes	110,464	55	318.00	825.5	1,225.57	1,665.00	11,867
nsentences	535,058	1	23.00	95.0	235.09	263.25	5,577

7.3 Explanation of Diagrams

7.3.1 Distributions of Document Length

The diagrams in Section 7.4 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.

7.3.2 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 7.5 and 7.6 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:

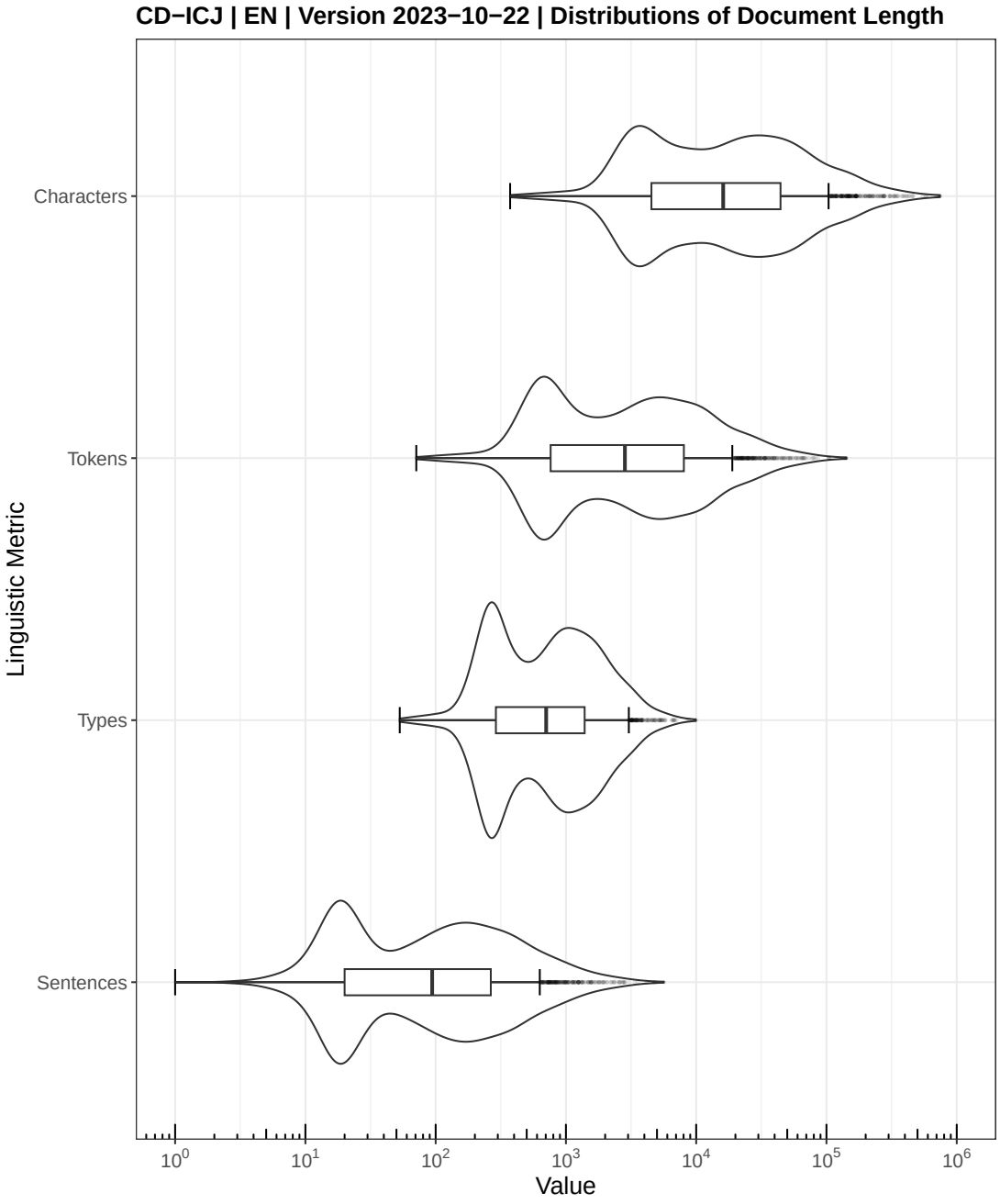
$$\text{tf-idf}_{td} = \mathrm{tf}_{td} \times \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

7.3.3 Tokens over Time

The charts in Section 7.7 show the total output of the International Court of Justice for each year as the sum total of the tokens of all published decisions (judgments, advisory opinions, orders, appended opinions). These charts may give a rough estimate of the activity of the International Court of Justice, although they should be interpreted with caution, as duplicate and highly similar opinions were not removed for this simple analysis. Please refer to Section 8 for the scope of identical and near-identical documents in the corpus.

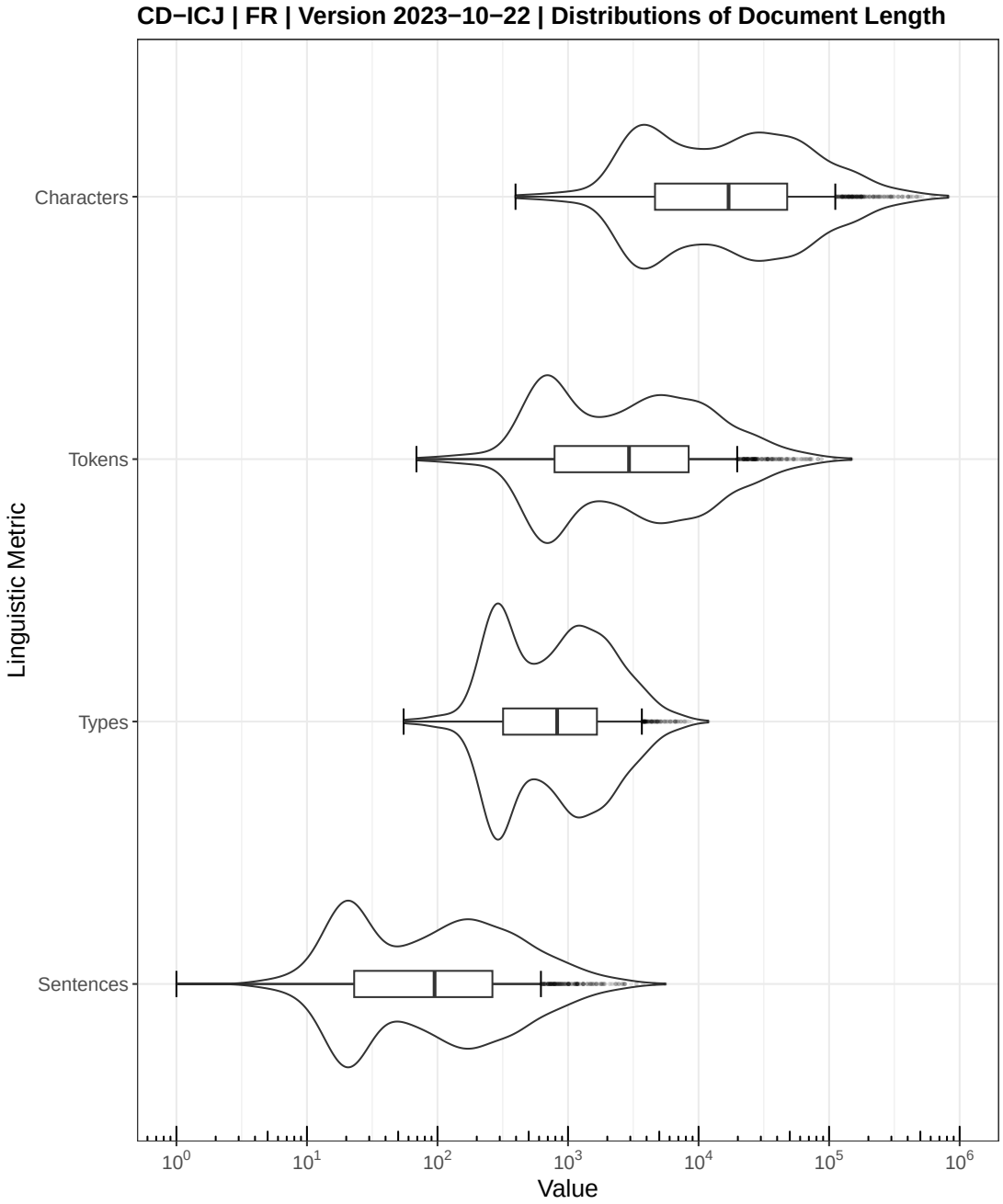
7.4 Distributions of Document Length

7.4.1 English



DOI: 10.5281/zenodo.10030647

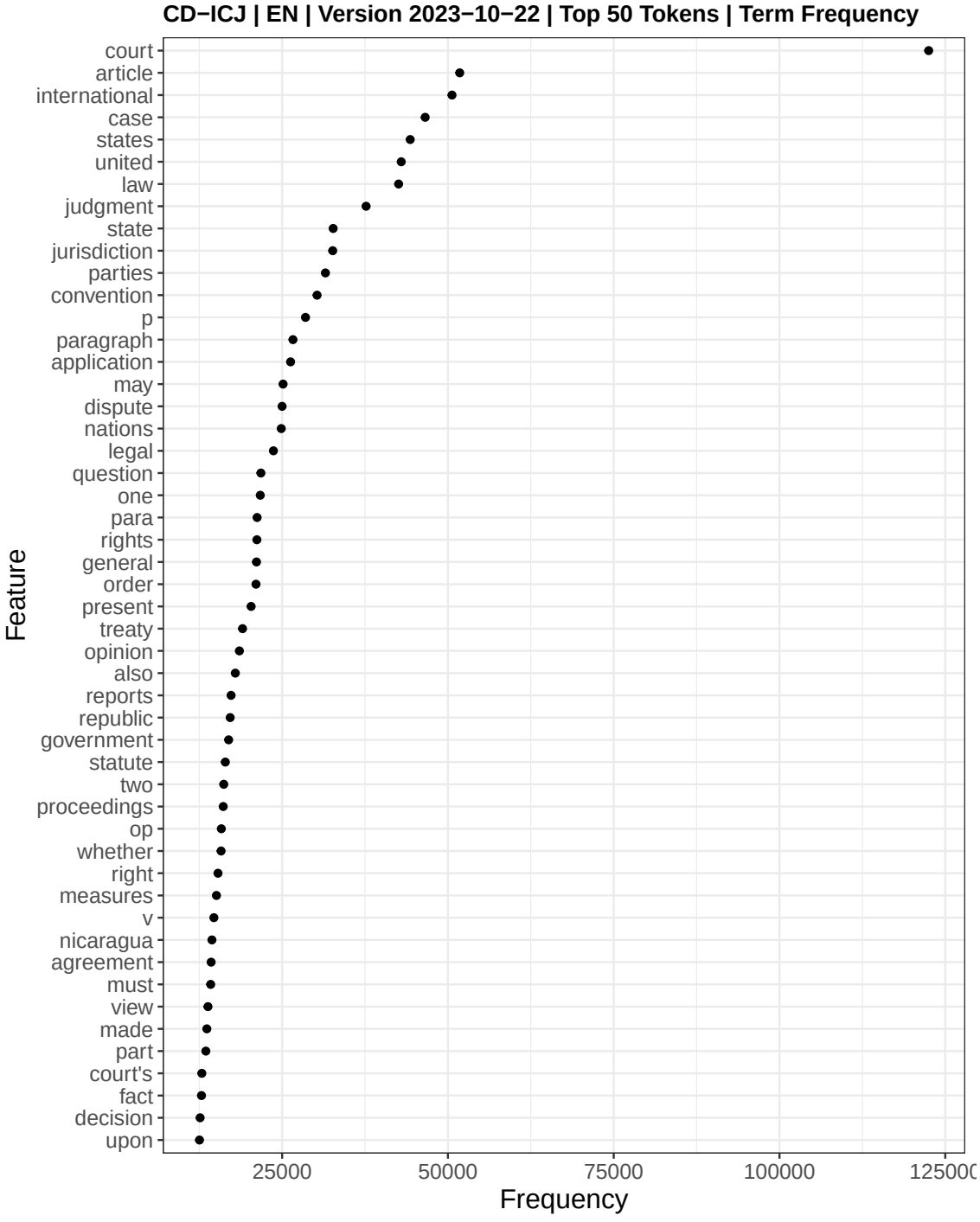
7.4.2 French



DOI: 10.5281/zenodo.10030647

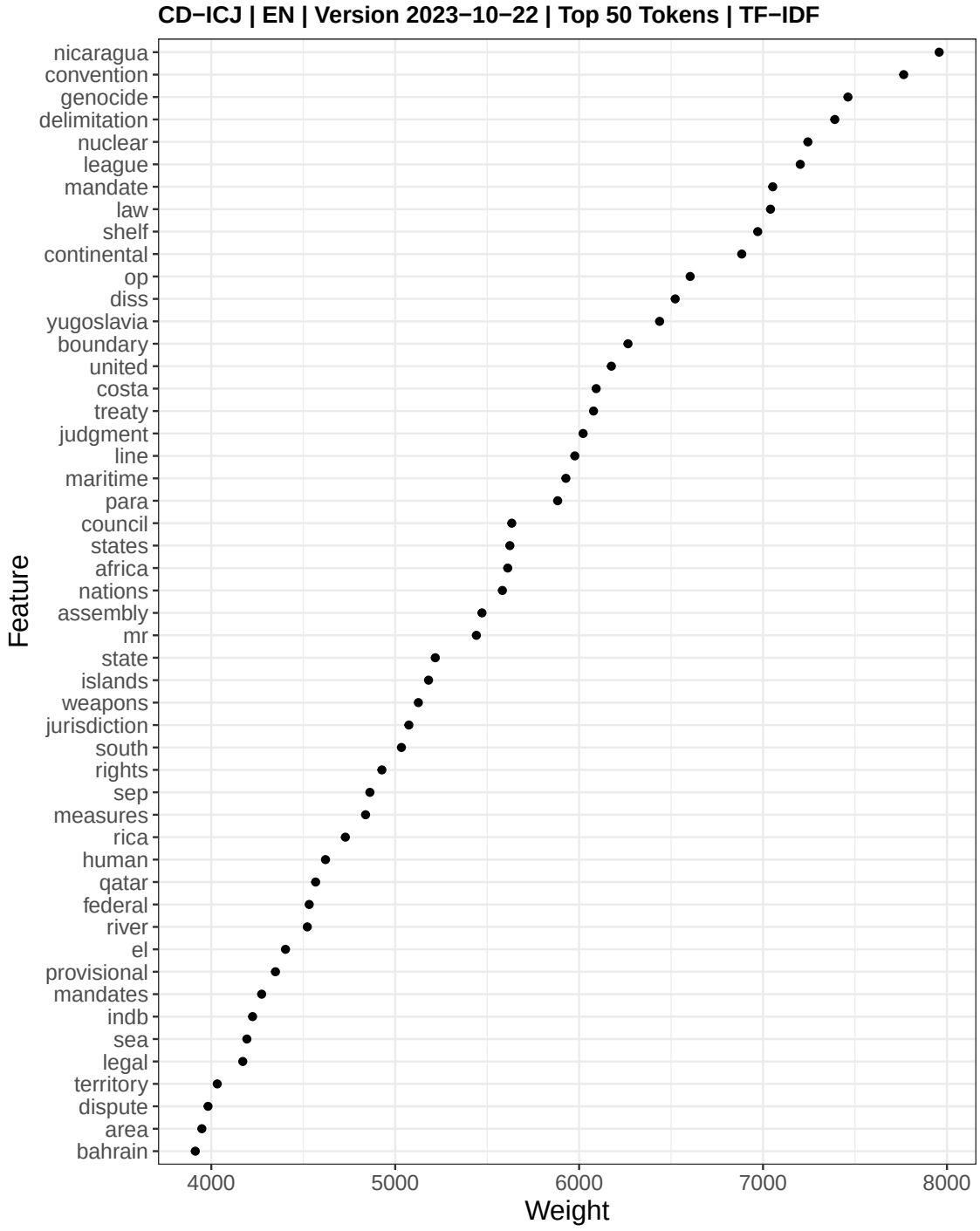
7.5 Most Frequent Tokens (English)

7.5.1 Term Frequency Weighting (TF)



DOI: 10.5281/zenodo.10030647

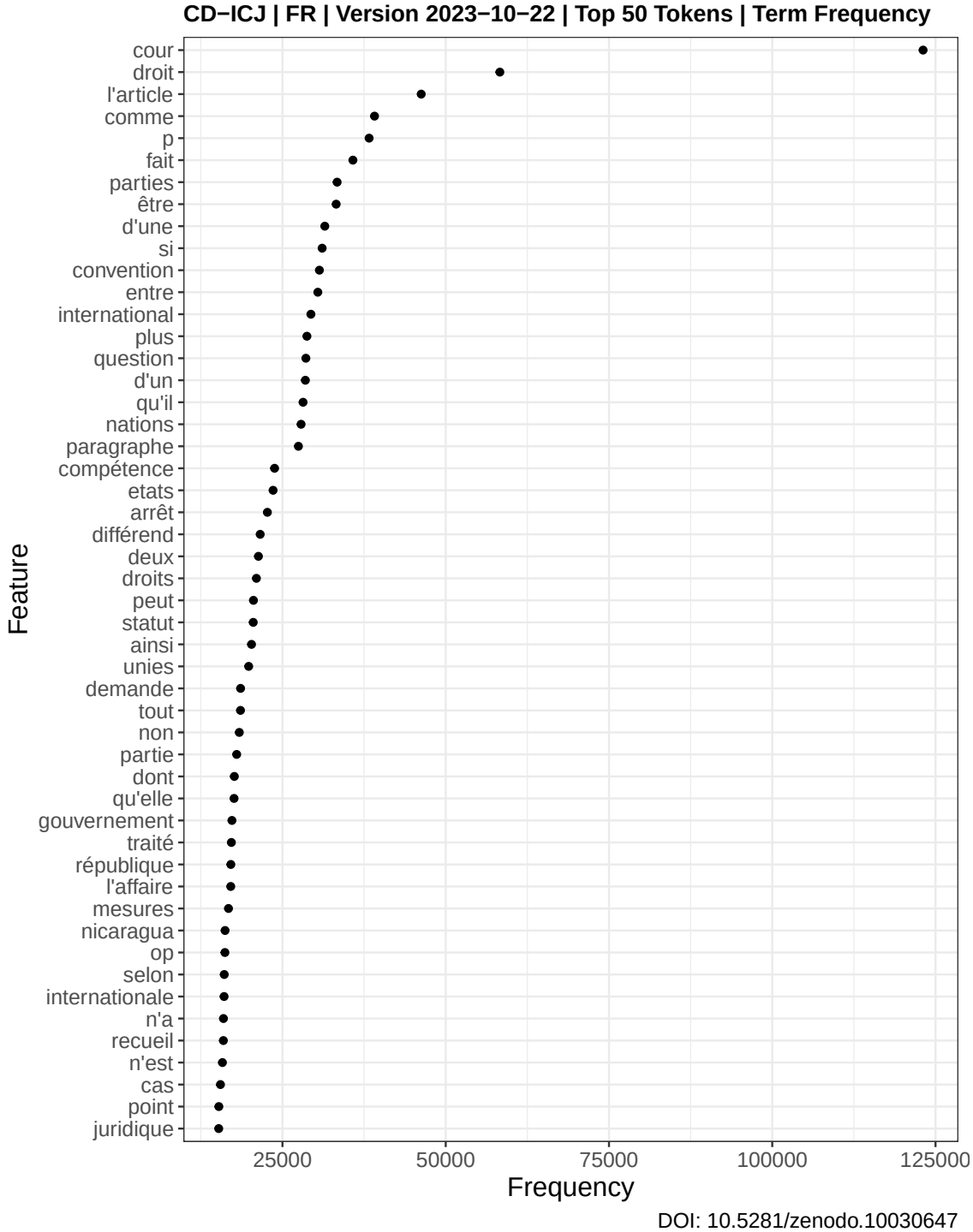
7.5.2 Term Frequency/Inverse Document Frequency Weighting (TF-IDF)



DOI: 10.5281/zenodo.10030647

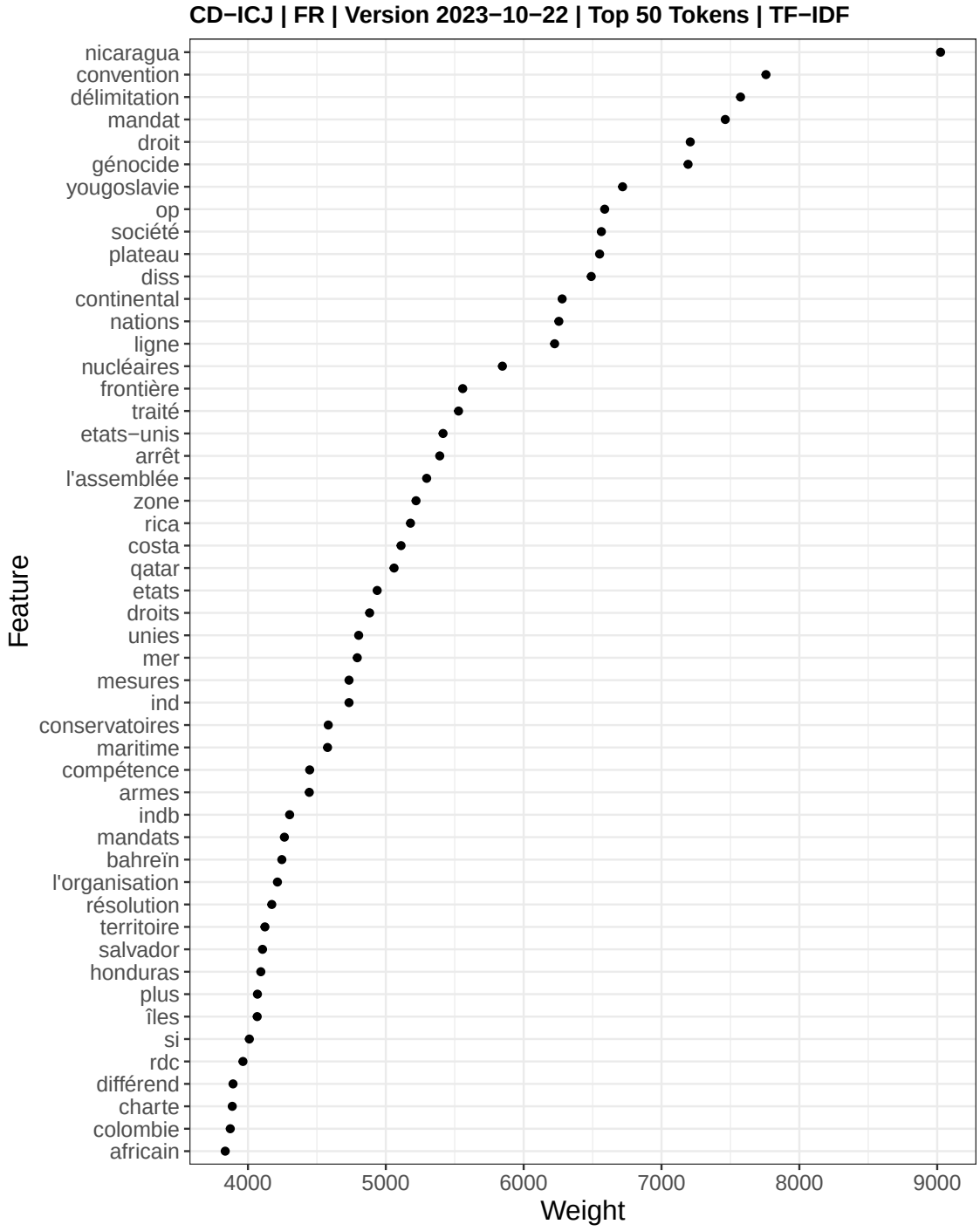
7.6 Most Frequent Tokens (French)

7.6.1 Term Frequency Weighting (TF)



DOI: 10.5281/zenodo.10030647

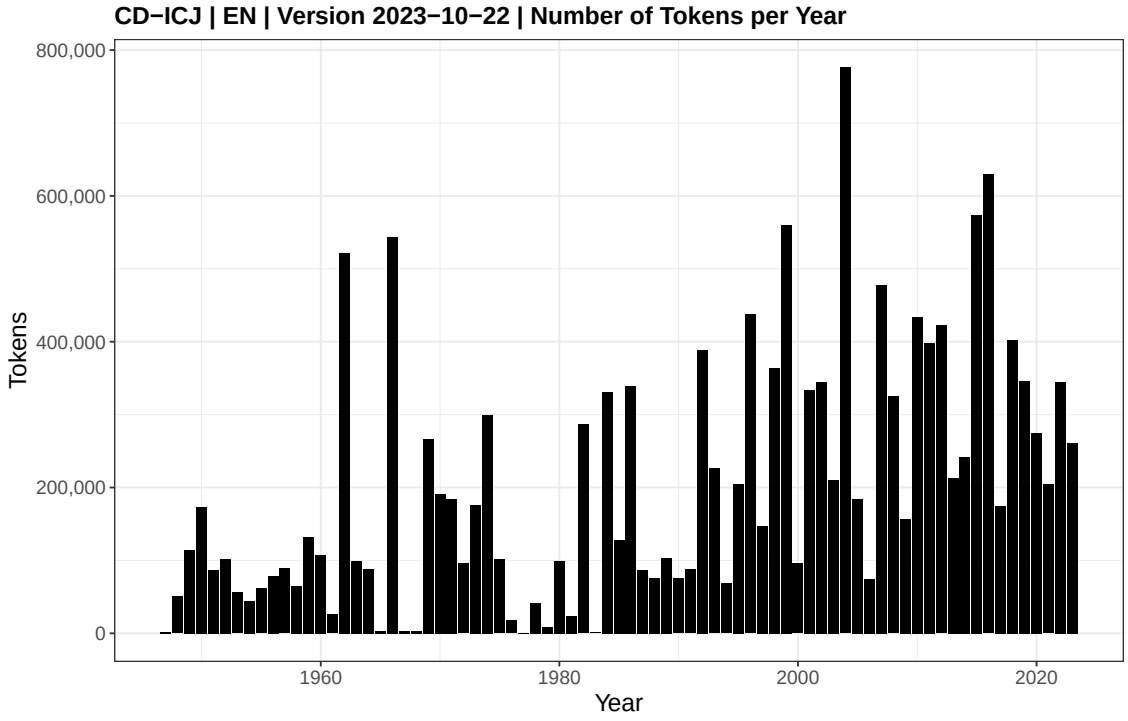
7.6.2 Term Frequency/Inverse Document Frequency Weighting (TF-IDF)



DOI: 10.5281/zenodo.10030647

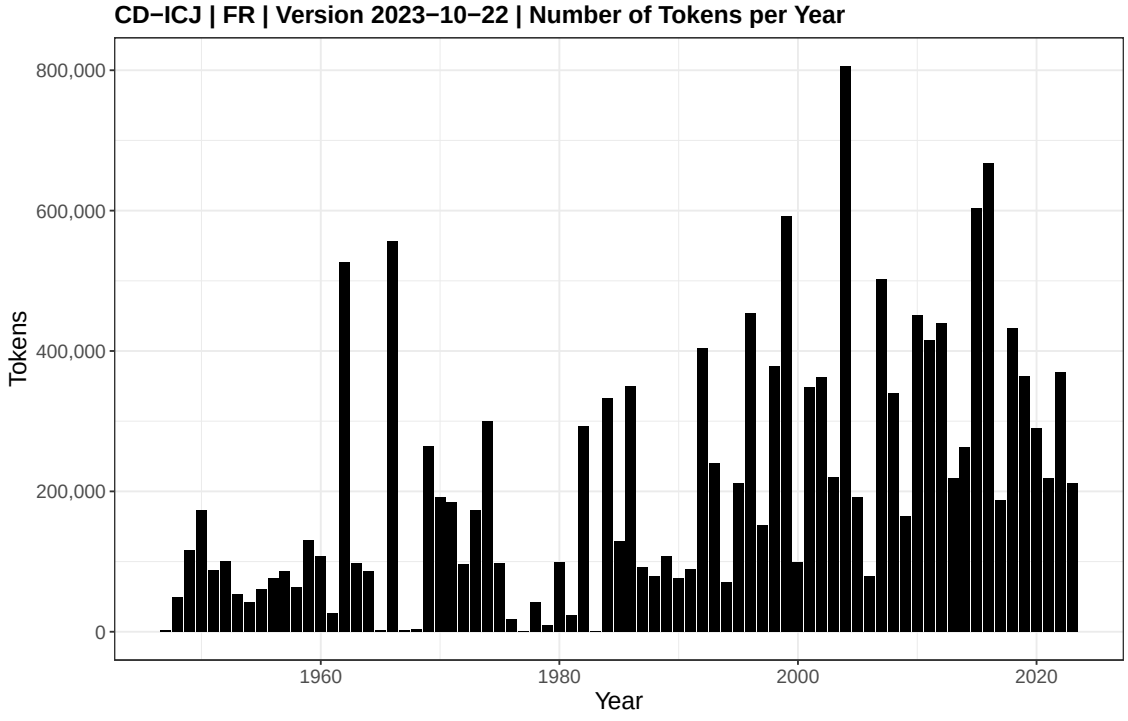
7.7 Tokens over Time

7.7.1 English



DOI: 10.5281/zenodo.10030647

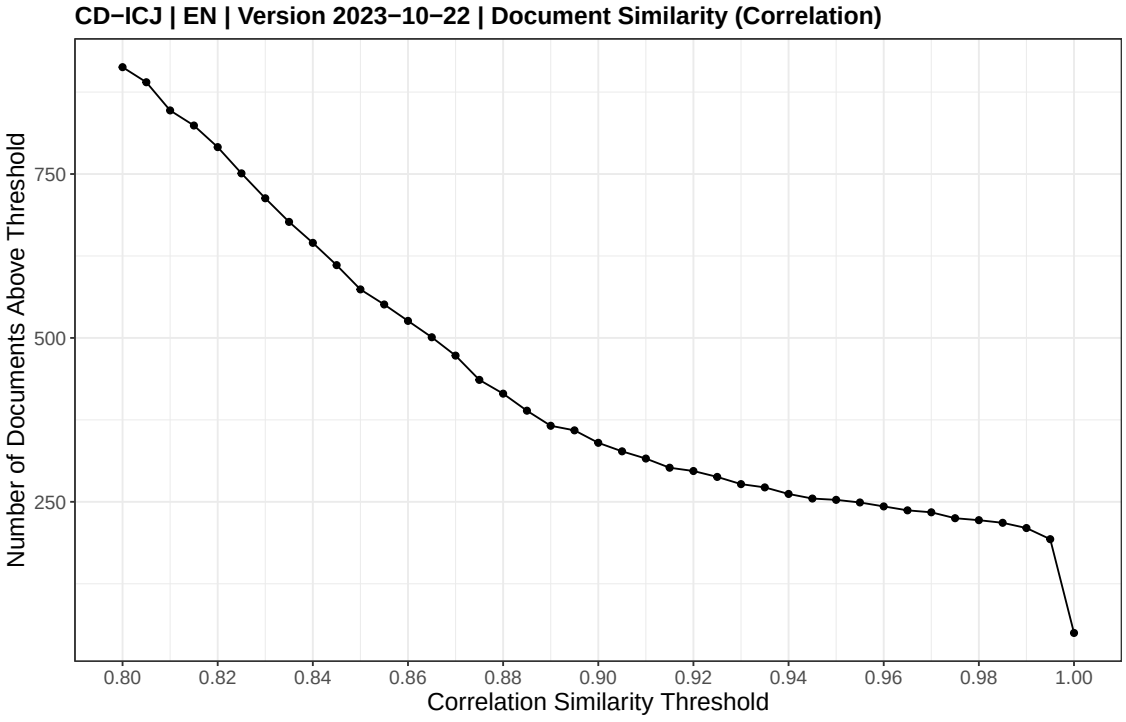
7.7.2 French



DOI: 10.5281/zenodo.10030647

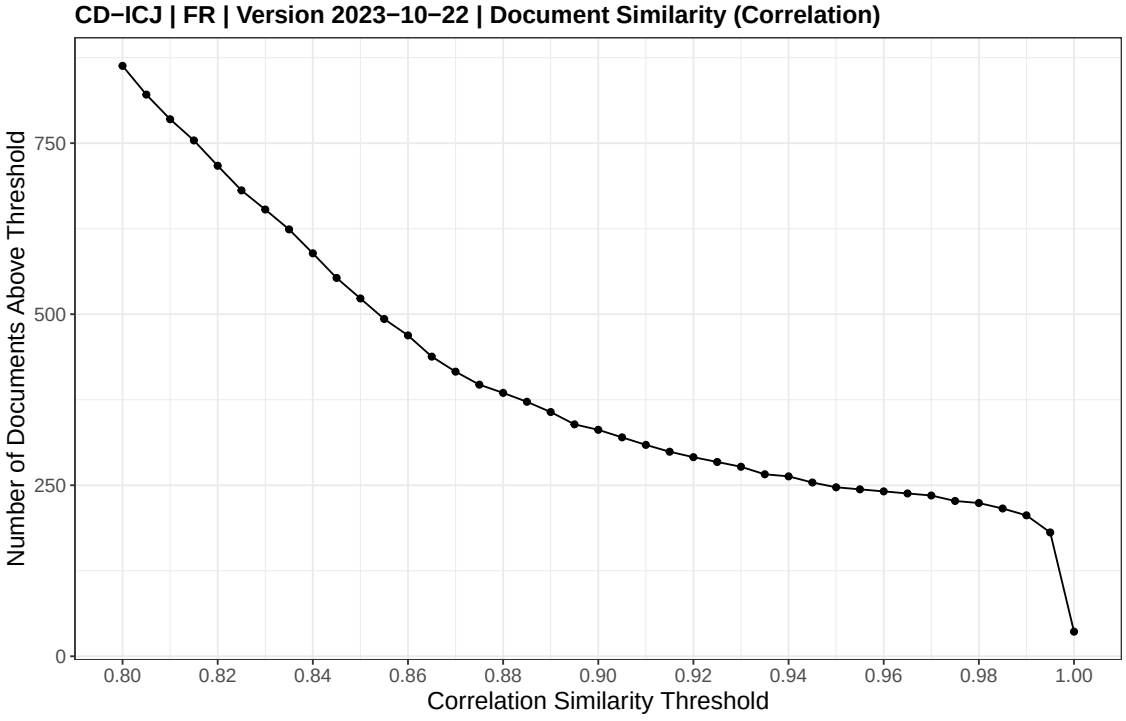
8 Document Similarity

8.1 English



DOI: 10.5281/zenodo.10030647

8.2 French



DOI: 10.5281/zenodo.10030647

8.3 Comment

Analysts are advised that the CD-ICJ contains a non-negligible number of highly similar to near-identical documents. This is due to the Court's long-standing practice of issuing formally different decisions for each Applicant-Respondent pair in the course of the same proceedings. A prime example of such proceedings are the *Use of Force* cases, for which the judgments are identical in content, but differ only in the names of the Parties across more than half a dozen different judgments.

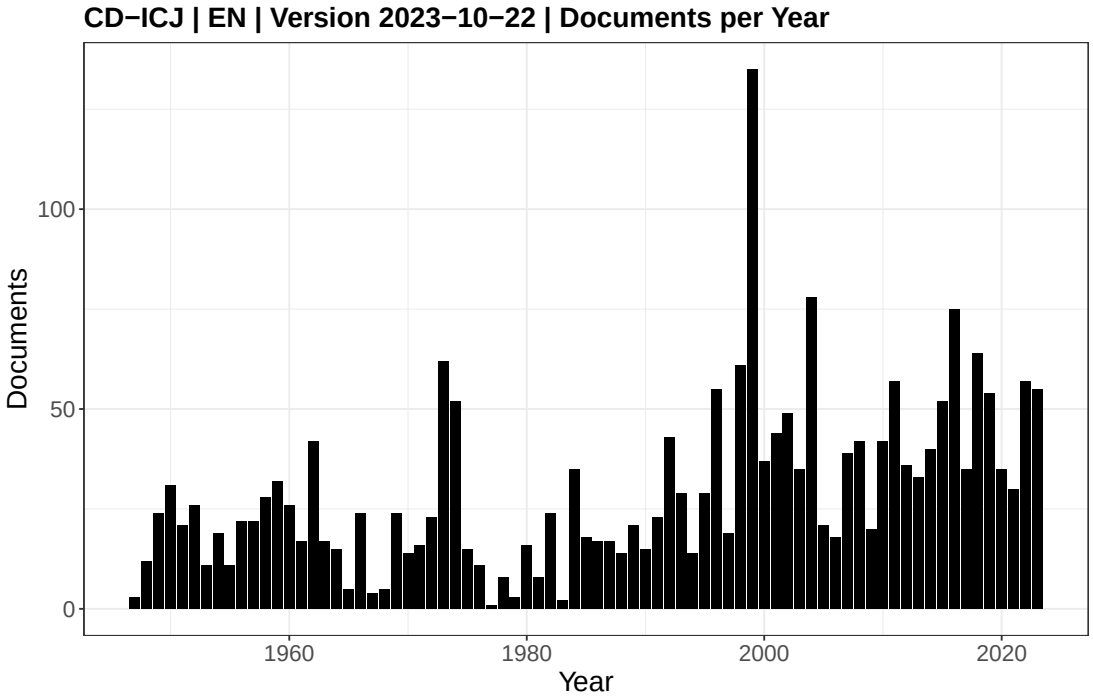
The above figures plot the number of files to be excluded as a function of correlation similarity based on a document-unigram matrix (with the removal of numbers, special symbols and stopwords, as well as lowercasing). Analysts who wish to qualitatively review this computational approach will find the 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 the data.

9 Metadata Frequency Tables

9.1 By Year

9.1.1 English



DOI: 10.5281/zenodo.10030647

Year	Documents	% Total	% Cumulative
1947	3	0.13	0.13
1948	12	0.52	0.66
1949	24	1.05	1.70
1950	31	1.35	3.06
1951	21	0.92	3.98
1952	26	1.14	5.11
1953	11	0.48	5.59
1954	19	0.83	6.42
1955	11	0.48	6.90
1956	22	0.96	7.86
1957	22	0.96	8.82
1958	28	1.22	10.05
1959	32	1.40	11.45

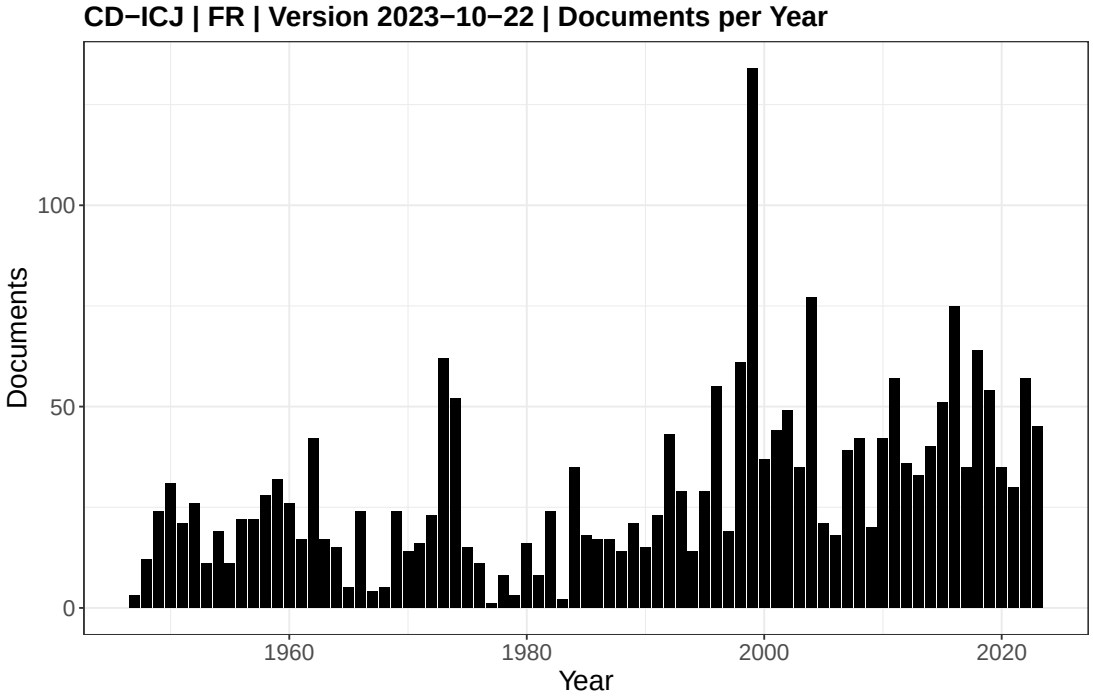
(continued)

Year	Documents	% Total	% Cumulative
1960	26	1.14	12.58
1961	17	0.74	13.32
1962	42	1.83	15.16
1963	17	0.74	15.90
1964	15	0.66	16.56
1965	5	0.22	16.78
1966	24	1.05	17.82
1967	4	0.17	18.00
1968	5	0.22	18.22
1969	24	1.05	19.27
1970	14	0.61	19.88
1971	16	0.70	20.58
1972	23	1.00	21.58
1973	62	2.71	24.29
1974	52	2.27	26.56
1975	15	0.66	27.22
1976	11	0.48	27.70
1977	1	0.04	27.74
1978	8	0.35	28.09
1979	3	0.13	28.22
1980	16	0.70	28.92
1981	8	0.35	29.27
1982	24	1.05	30.32
1983	2	0.09	30.41
1984	35	1.53	31.94
1985	18	0.79	32.72
1986	17	0.74	33.46
1987	17	0.74	34.21
1988	14	0.61	34.82
1989	21	0.92	35.74
1990	15	0.66	36.39
1991	23	1.00	37.40
1992	43	1.88	39.27

(continued)

Year	Documents	% Total	% Cumulative
1993	29	1.27	40.54
1994	14	0.61	41.15
1995	29	1.27	42.42
1996	55	2.40	44.82
1997	19	0.83	45.65
1998	61	2.66	48.32
1999	135	5.90	54.22
2000	37	1.62	55.83
2001	44	1.92	57.75
2002	49	2.14	59.90
2003	35	1.53	61.42
2004	78	3.41	64.83
2005	21	0.92	65.75
2006	18	0.79	66.54
2007	39	1.70	68.24
2008	42	1.83	70.07
2009	20	0.87	70.95
2010	42	1.83	72.78
2011	57	2.49	75.27
2012	36	1.57	76.85
2013	33	1.44	78.29
2014	40	1.75	80.03
2015	52	2.27	82.31
2016	75	3.28	85.58
2017	35	1.53	87.11
2018	64	2.80	89.91
2019	54	2.36	92.27
2020	35	1.53	93.80
2021	30	1.31	95.11
2022	57	2.49	97.60
2023	55	2.40	100.00
Total	2289	100.00	100.00

9.1.2 French



DOI: 10.5281/zenodo.10030647

Year	Documents	% Total	% Cumulative
1947	3	0.13	0.13
1948	12	0.53	0.66
1949	24	1.05	1.71
1950	31	1.36	3.08
1951	21	0.92	4.00
1952	26	1.14	5.14
1953	11	0.48	5.62
1954	19	0.83	6.46
1955	11	0.48	6.94
1956	22	0.97	7.91
1957	22	0.97	8.88
1958	28	1.23	10.11
1959	32	1.41	11.51
1960	26	1.14	12.65
1961	17	0.75	13.40
1962	42	1.85	15.25

(continued)

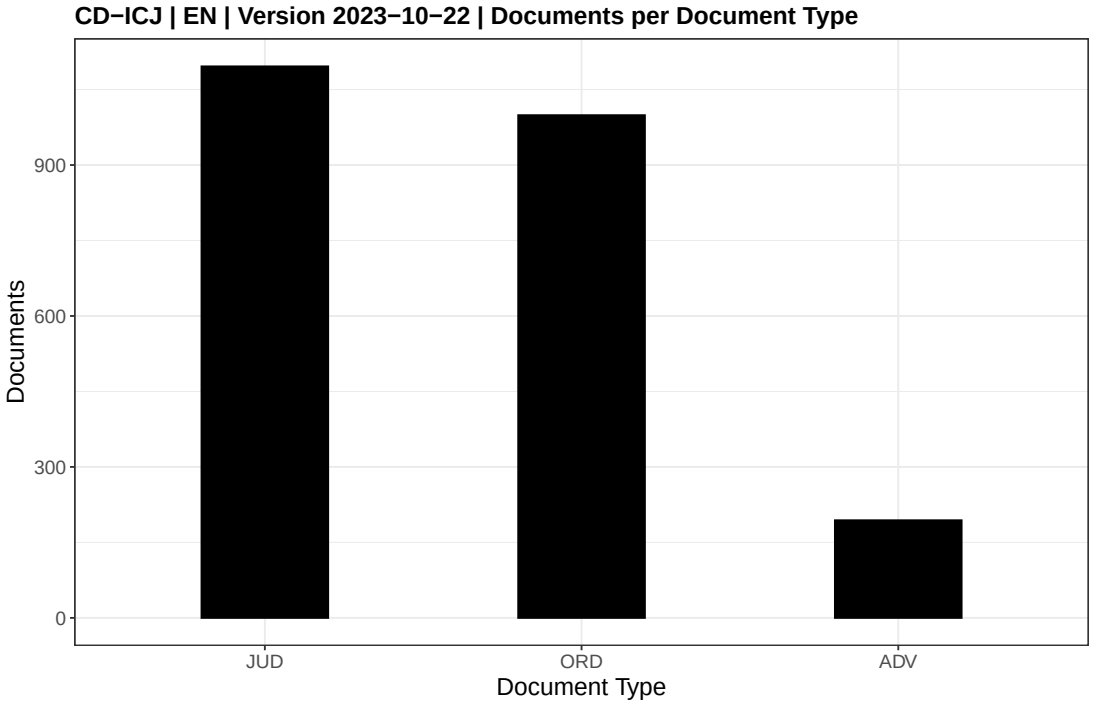
Year	Documents	% Total	% Cumulative
1963	17	0.75	15.99
1964	15	0.66	16.65
1965	5	0.22	16.87
1966	24	1.05	17.93
1967	4	0.18	18.10
1968	5	0.22	18.32
1969	24	1.05	19.38
1970	14	0.62	19.99
1971	16	0.70	20.69
1972	23	1.01	21.70
1973	62	2.72	24.43
1974	52	2.28	26.71
1975	15	0.66	27.37
1976	11	0.48	27.86
1977	1	0.04	27.90
1978	8	0.35	28.25
1979	3	0.13	28.38
1980	16	0.70	29.09
1981	8	0.35	29.44
1982	24	1.05	30.49
1983	2	0.09	30.58
1984	35	1.54	32.12
1985	18	0.79	32.91
1986	17	0.75	33.66
1987	17	0.75	34.40
1988	14	0.62	35.02
1989	21	0.92	35.94
1990	15	0.66	36.60
1991	23	1.01	37.61
1992	43	1.89	39.50
1993	29	1.27	40.77
1994	14	0.62	41.39
1995	29	1.27	42.66
1996	55	2.42	45.08

(continued)

Year	Documents	% Total	% Cumulative
1997	19	0.83	45.91
1998	61	2.68	48.59
1999	134	5.89	54.48
2000	37	1.63	56.11
2001	44	1.93	58.04
2002	49	2.15	60.19
2003	35	1.54	61.73
2004	77	3.38	65.11
2005	21	0.92	66.04
2006	18	0.79	66.83
2007	39	1.71	68.54
2008	42	1.85	70.39
2009	20	0.88	71.27
2010	42	1.85	73.11
2011	57	2.50	75.62
2012	36	1.58	77.20
2013	33	1.45	78.65
2014	40	1.76	80.40
2015	51	2.24	82.64
2016	75	3.30	85.94
2017	35	1.54	87.48
2018	64	2.81	90.29
2019	54	2.37	92.66
2020	35	1.54	94.20
2021	30	1.32	95.52
2022	57	2.50	98.02
2023	45	1.98	100.00
Total	2276	100.00	100.00

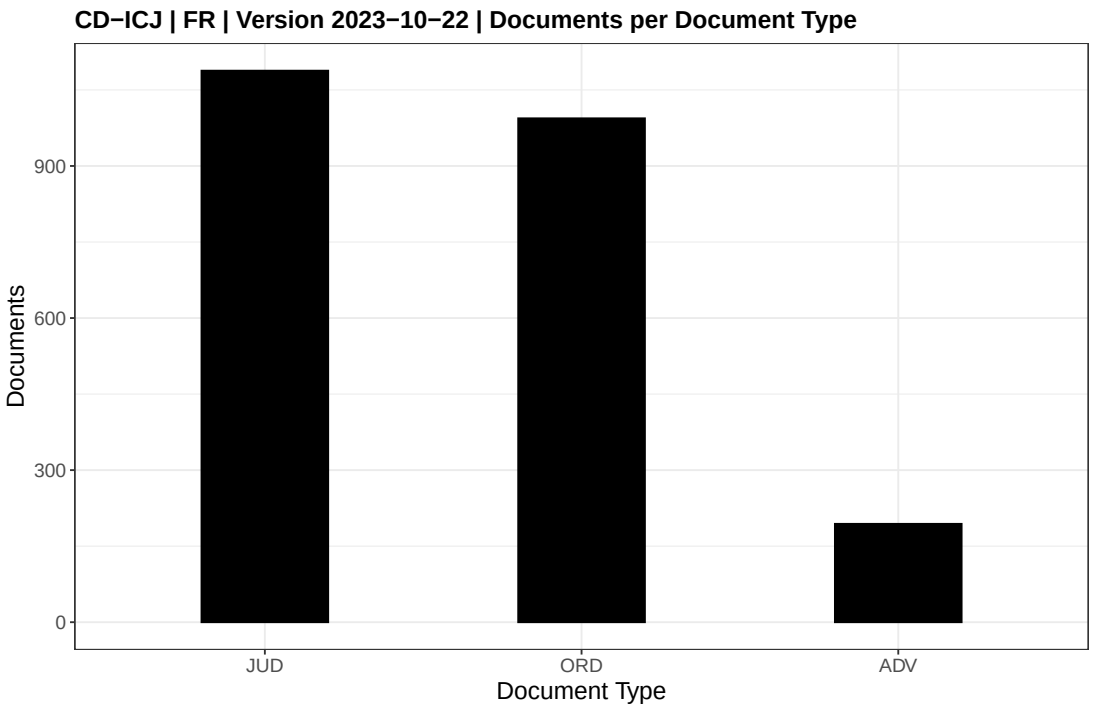
9.2 By Document Type

9.2.1 English



DocType	Documents	% Total	% Cumulative
ADV	194	8.48	8.48
JUD	1096	47.88	56.36
ORD	999	43.64	100.00
Total	2289	100.00	100.00

9.2.2 French

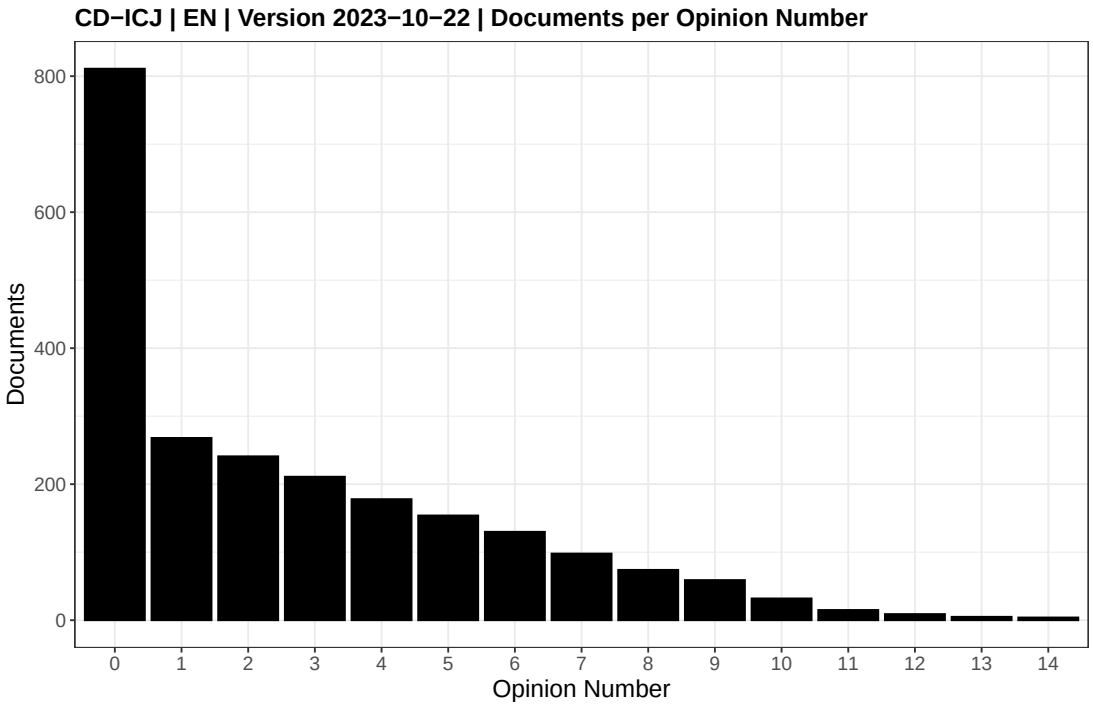


DOI: 10.5281/zenodo.10030647

DocType	Documents	% Total	% Cumulative
ADV	194	8.52	8.52
JUD	1088	47.80	56.33
ORD	994	43.67	100.00
Total	2276	100.00	100.00

9.3 By Opinion Number

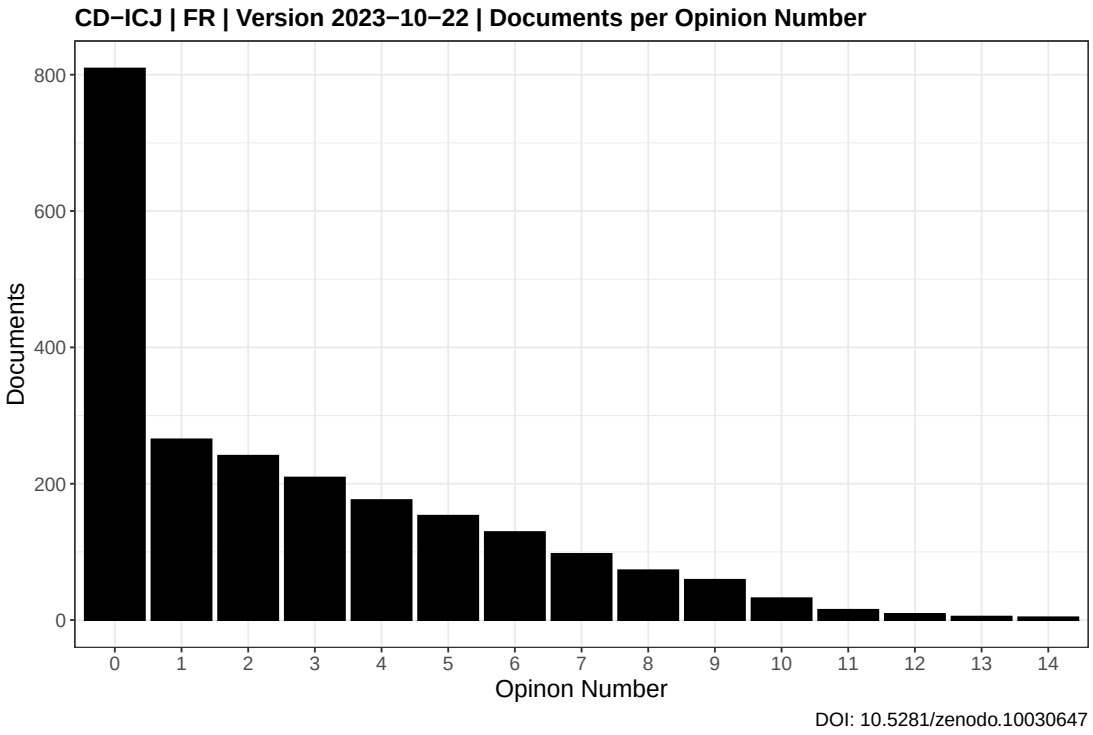
9.3.1 English



DOI: 10.5281/zenodo.10030647

Opinion Number	Documents	% Total	% Cumulative
0	811	35.43	35.43
1	268	11.71	47.14
2	241	10.53	57.67
3	211	9.22	66.89
4	178	7.78	74.66
5	154	6.73	81.39
6	130	5.68	87.07
7	98	4.28	91.35
8	74	3.23	94.58
9	59	2.58	97.16
10	32	1.40	98.56
11	15	0.66	99.21
12	9	0.39	99.61
13	5	0.22	99.83
14	4	0.17	100.00
Total	2289	100.00	100.00

9.3.2 French



Opinion Number	Documents	% Total	% Cumulative
0	809	35.54	35.54
1	265	11.64	47.19
2	241	10.59	57.78
3	209	9.18	66.96
4	176	7.73	74.69
5	153	6.72	81.41
6	129	5.67	87.08
7	97	4.26	91.34
8	73	3.21	94.55
9	59	2.59	97.14
10	32	1.41	98.55
11	15	0.66	99.21
12	9	0.40	99.60
13	5	0.22	99.82
14	4	0.18	100.00
Total	2276	100.00	100.00

9.4 By Applicant

9.4.1 English

Applicant	Documents	% Total	% Cumulative
ARG	20	0.87	0.87
ARM	16	0.70	1.57
AUS	48	2.10	3.67
AZE	8	0.35	4.02
BEL	64	2.80	6.82
BEN	7	0.31	7.12
BFA	16	0.70	7.82
BHR-EGY-ARE	6	0.26	8.08
BHR-EGY-SAU-ARE	6	0.26	8.34
BIH	49	2.14	10.48
BLZ	1	0.04	10.53
BOL	14	0.61	11.14
BWA	12	0.52	11.66
CAN	12	0.52	12.19
CAN-SWE-UKR-GBR	1	0.04	12.23
CARAT	32	1.40	13.63
CHE	16	0.70	14.33
CHL	8	0.35	14.68
CMR	56	2.45	17.13
COD	90	3.93	21.06
COL	12	0.52	21.58
CRI	69	3.01	24.60
DEU	84	3.67	28.27
DJI	11	0.48	28.75
DMA	1	0.04	28.79
DNK	13	0.57	29.36
ECOSOC	11	0.48	29.84
ECU	4	0.17	30.01
ESP	13	0.57	30.58
ETH	30	1.31	31.89

(continued)

Applicant	Documents	% Total	% Cumulative
FIN	7	0.31	32.20
FRA	32	1.40	33.60
GAB	2	0.09	33.68
GBR	77	3.36	37.05
GEO	16	0.70	37.75
GIN	24	1.05	38.79
GMB	13	0.57	39.36
GNB	16	0.70	40.06
GNQ	28	1.22	41.28
GRC	36	1.57	42.86
GTM	3	0.13	42.99
GUY	17	0.74	43.73
HND	7	0.31	44.04
HRV	31	1.35	45.39
HUN	16	0.70	46.09
IDN	14	0.61	46.70
IFAD	5	0.22	46.92
IMO	4	0.17	47.09
IND	25	1.09	48.19
IRN	82	3.58	51.77
ISR	8	0.35	52.12
ITA	7	0.31	52.42
KHM	30	1.31	53.74
LBR	30	1.31	55.05
LBY	84	3.67	58.72
LIE	19	0.83	59.55
MEX	19	0.83	60.38
MHL	51	2.23	62.60
MKD	8	0.35	62.95
MYS	11	0.48	63.43
NIC	156	6.82	70.25
NLD	11	0.48	70.73
NRU	11	0.48	71.21

(continued)

Applicant	Documents	% Total	% Cumulative
NZL	37	1.62	72.83
PAK	12	0.52	73.35
PER	13	0.57	73.92
PRT	35	1.53	75.45
PRY	7	0.31	75.75
PSE	1	0.04	75.80
QAT	51	2.23	78.03
ROU	4	0.17	78.20
SCG	163	7.12	85.32
SLV	24	1.05	86.37
SOM	17	0.74	87.11
TLS	11	0.48	87.59
TUN	19	0.83	88.42
UKR	39	1.70	90.13
UNESCO	9	0.39	90.52
UNGA	144	6.29	96.81
UNSC	16	0.70	97.51
USA	21	0.92	98.43
WHO	20	0.87	99.30
YUG	16	0.70	100.00
Total	2289	100.00	100.00

9.4.2 French

Applicant	Documents	% Total	% Cumulative
ARG	20	0.88	0.88
ARM	16	0.70	1.58
AUS	48	2.11	3.69
AZE	8	0.35	4.04
BEL	64	2.81	6.85
BEN	6	0.26	7.12
BFA	16	0.70	7.82
BHR-EGY-ARE	6	0.26	8.08
BHR-EGY-SAU-ARE	6	0.26	8.35
BIH	49	2.15	10.50
BLZ	1	0.04	10.54
BOL	14	0.62	11.16
BWA	12	0.53	11.69
CAN	12	0.53	12.21
CAN-SWE-UKR-GBR	1	0.04	12.26
CARAT	32	1.41	13.66
CHE	16	0.70	14.37
CHL	8	0.35	14.72
CMR	56	2.46	17.18
COD	90	3.95	21.13
COL	12	0.53	21.66
CRI	69	3.03	24.69
DEU	84	3.69	28.38
DJI	11	0.48	28.87
DMA	1	0.04	28.91
DNK	13	0.57	29.48
ECOSOC	11	0.48	29.96
ECU	4	0.18	30.14
ESP	13	0.57	30.71
ETH	30	1.32	32.03
FIN	7	0.31	32.34

(continued)

Applicant	Documents	% Total	% Cumulative
FRA	32	1.41	33.74
GAB	2	0.09	33.83
GBR	77	3.38	37.21
GEO	16	0.70	37.92
GIN	24	1.05	38.97
GMB	13	0.57	39.54
GNB	16	0.70	40.25
GNQ	28	1.23	41.48
GRC	36	1.58	43.06
GTM	3	0.13	43.19
GUY	17	0.75	43.94
HND	7	0.31	44.24
HRV	31	1.36	45.61
HUN	16	0.70	46.31
IDN	14	0.62	46.92
IFAD	5	0.22	47.14
IMO	4	0.18	47.32
IND	25	1.10	48.42
IRN	82	3.60	52.02
ISR	8	0.35	52.37
ITA	7	0.31	52.68
KHM	30	1.32	54.00
LBR	30	1.32	55.32
LBY	83	3.65	58.96
LIE	19	0.83	59.80
MEX	19	0.83	60.63
MHL	51	2.24	62.87
MKD	8	0.35	63.22
MYS	11	0.48	63.71
NIC	148	6.50	70.21
NLD	11	0.48	70.69
NRU	11	0.48	71.18
NZL	37	1.63	72.80
PAK	12	0.53	73.33

(continued)

Applicant	Documents	% Total	% Cumulative
PER	13	0.57	73.90
PRT	35	1.54	75.44
PRY	7	0.31	75.75
PSE	1	0.04	75.79
QAT	51	2.24	78.03
ROU	4	0.18	78.21
SCG	163	7.16	85.37
SLV	24	1.05	86.42
SOM	17	0.75	87.17
TLS	10	0.44	87.61
TUN	19	0.83	88.44
UKR	37	1.63	90.07
UNESCO	9	0.40	90.47
UNGA	144	6.33	96.79
UNSC	16	0.70	97.50
USA	21	0.92	98.42
WHO	20	0.88	99.30
YUG	16	0.70	100.00
Total	2276	100.00	100.00

9.5 By Respondent

9.5.1 English

Respondent	Documents	% Total	% Cumulative
NA	241	10.53	10.53
ALB	19	0.83	11.36
ARE	22	0.96	12.32
ARG	1	0.04	12.36
ARM	8	0.35	12.71
AUS	32	1.40	14.11
AZE	16	0.70	14.81
BDI	3	0.13	14.94
BEL	43	1.88	16.82
BGR	21	0.92	17.74
BHR	29	1.27	19.00
BLZ	3	0.13	19.13
BOL	8	0.35	19.48
BRA	1	0.04	19.53
CAN	35	1.53	21.06
CHE	4	0.17	21.23
CHL	28	1.22	22.46
COD	24	1.05	23.50
COL	90	3.93	27.44
CRI	19	0.83	28.27
CSK	1	0.04	28.31
DEU	27	1.18	29.49
DNK	22	0.96	30.45
EGY	1	0.04	30.49
ESP	52	2.27	32.77
FRA	137	5.99	38.75
FRA-GBR-USA	7	0.31	39.06
GBR	107	4.67	43.73
GNQ	2	0.09	43.82
GRC	8	0.35	44.17
GTM	11	0.48	44.65
HND	44	1.92	46.57

(continued)

Respondent	Documents	% Total	% Cumulative
HUN	1	0.04	46.61
IND	54	2.36	48.97
IRN	20	0.87	49.85
ISL	49	2.14	51.99
ITA	44	1.92	53.91
JPN	17	0.74	54.65
KEN	17	0.74	55.40
LBN	7	0.31	55.70
LBY	19	0.83	56.53
MLI	8	0.35	56.88
MLT	22	0.96	57.84
MMR	13	0.57	58.41
MYS	14	0.61	59.02
NAM	12	0.52	59.55
NER	15	0.66	60.20
NGA	38	1.66	61.86
NIC	75	3.28	65.14
NLD	44	1.92	67.06
NOR	33	1.44	68.50
PAK	42	1.83	70.34
PER	12	0.52	70.86
PRT	21	0.92	71.78
QAT	12	0.52	72.30
RUS	55	2.40	74.71
RWA	19	0.83	75.54
SCG	43	1.88	77.41
SEN	33	1.44	78.86
SGP	11	0.48	79.34
SRB	31	1.35	80.69
SUN	4	0.17	80.87
SVK	16	0.70	81.56
SWE	11	0.48	82.04
TCD	8	0.35	82.39

(continued)

Respondent	Documents	% Total	% Cumulative
THA	30	1.31	83.70
TUR	20	0.87	84.58
UGA	35	1.53	86.11
UKR	4	0.17	86.28
URY	20	0.87	87.16
USA	211	9.22	96.37
VEN	17	0.74	97.12
YUG	6	0.26	97.38
ZAF	60	2.62	100.00
Total	2289	100.00	100.00

9.5.2 French

Respondent	Documents	% Total	% Cumulative
NA	241	10.59	10.59
ALB	19	0.83	11.42
ARE	22	0.97	12.39
ARG	1	0.04	12.43
ARM	8	0.35	12.79
AUS	31	1.36	14.15
AZE	16	0.70	14.85
BDI	3	0.13	14.98
BEL	43	1.89	16.87
BGR	21	0.92	17.79
BHR	29	1.27	19.07
BLZ	3	0.13	19.20
BOL	8	0.35	19.55
BRA	1	0.04	19.60
CAN	35	1.54	21.13
CHE	4	0.18	21.31
CHL	28	1.23	22.54
COD	24	1.05	23.59
COL	82	3.60	27.20
CRI	19	0.83	28.03
CSK	1	0.04	28.08
DEU	27	1.19	29.26
DNK	22	0.97	30.23
EGY	1	0.04	30.27
ESP	52	2.28	32.56
FRA	137	6.02	38.58
FRA-GBR-USA	7	0.31	38.88
GBR	107	4.70	43.59
GNQ	2	0.09	43.67
GRC	8	0.35	44.02
GTM	11	0.48	44.51
HND	44	1.93	46.44
HUN	1	0.04	46.49

(continued)

Respondent	Documents	% Total	% Cumulative
IND	54	2.37	48.86
IRN	20	0.88	49.74
ISL	49	2.15	51.89
ITA	44	1.93	53.82
JPN	17	0.75	54.57
KEN	17	0.75	55.32
LBN	7	0.31	55.62
LBY	19	0.83	56.46
MLI	8	0.35	56.81
MLT	22	0.97	57.78
MMR	13	0.57	58.35
MYS	14	0.62	58.96
NAM	12	0.53	59.49
NER	14	0.62	60.11
NGA	38	1.67	61.78
NIC	75	3.30	65.07
NLD	44	1.93	67.00
NOR	33	1.45	68.45
PAK	42	1.85	70.30
PER	12	0.53	70.83
PRT	21	0.92	71.75
QAT	12	0.53	72.28
RUS	53	2.33	74.60
RWA	19	0.83	75.44
SCG	43	1.89	77.33
SEN	33	1.45	78.78
SGP	11	0.48	79.26
SRB	31	1.36	80.62
SUN	4	0.18	80.80
SVK	16	0.70	81.50
SWE	11	0.48	81.99
TCD	8	0.35	82.34
THA	30	1.32	83.66

(continued)

Respondent	Documents	% Total	% Cumulative
TUR	20	0.88	84.53
UGA	35	1.54	86.07
UKR	4	0.18	86.25
URY	20	0.88	87.13
USA	210	9.23	96.35
VEN	17	0.75	97.10
YUG	6	0.26	97.36
ZAF	60	2.64	100.00
Total	2276	100.00	100.00

10 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!'

```
# Function: Test SHA3-Hashes
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-ICJ_2023-10-22_EN_CSV_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_EN_CSV_BEST_META.zip	Signature verified!
CD-ICJ_2023-10-22_EN_PDF_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_EN_PDF_BEST_MajorityOpinions.zip	Signature verified!
CD-ICJ_2023-10-22_EN_PDF_ENHANCED_max2004.zip	Signature verified!
CD-ICJ_2023-10-22_EN_PDF_ORIGINAL_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_EN_TXT_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_EN_TXT_EXTRACTED_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_EN_TXT_TESSERACT_max2004.zip	Signature verified!
CD-ICJ_2023-10-22_EN-FR_ANALYSIS.zip	Signature verified!
CD-ICJ_2023-10-22_FR_CSV_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_FR_CSV_BEST_META.zip	Signature verified!
CD-ICJ_2023-10-22_FR_PDF_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_FR_PDF_BEST_MajorityOpinions.zip	Signature verified!
CD-ICJ_2023-10-22_FR_PDF_ENHANCED_max2004.zip	Signature verified!
CD-ICJ_2023-10-22_FR_PDF_ORIGINAL_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_FR_TXT_BEST_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_FR_TXT_EXTRACTED_FULL.zip	Signature verified!
CD-ICJ_2023-10-22_FR_TXT_TESSERACT_max2004.zip	Signature verified!
CD-ICJ_2023-10-22_Source_Files.zip	Signature verified!
CD-ICJ_2023-10-22_UnlabelledFiles.zip	Signature verified!

11 Changelog

The Changelog documents changes made to the data set. Versions are named according to the day on which the data creation process began.

11.1 Version 2023-10-22

- Full recompilation of data set
- Scope extended up to case number 190: *Aerial Incident of 8 January 2020* (Canada, Sweden, Ukraine and United Kingdom v. Islamic Republic of Iran)
- Add fix for lowercase components in URL basenames
- Updated Python toolchain
- Align docker config with Debian as host system

11.2 Version 2023-05-07

- Full recompilation of data set
- Entire computational environment now version-controlled with Docker
- Scope extended up to case number 187: *Obligations of States in respect of climate change* (Advisory Opinion)
- Upgrade Tesseract OCR to version 5.3.1
- Upgrade OCR training data to "tesseract_best"
- Simplified config file
- Simplified function loading
- Ensure that debug mode only processes cases once
- Fix download manifest
- Update download function
- Contents of source ZIP file linked to Git manifest

11.3 Version 2022-09-07

- Full recompilation of data set
- Scope extended up to case number 183: *Jurisdictional Immunities* (Germany v Italy)
- Upgraded OCR to Tesseract 5.0.1
- CHANGELOG and README converted to external markdown files
- The ZIP archive of source files includes the TEX files
- Config file converted to TOML format
- All R packages are version-controlled with {renv}
- Data set creation process cleans up all files from previous runs before a new data set is created
- Removed redundant color from violin plots
- Added custom split instructions for the 2021-07-21 Order in the Amity Treaty case

11.4 Version 2021-11-23

- Initial Release

12 Strict Replication Parameters

```
## [1] "OpenSSL 3.0.2 15 Mar 2022 (Library: OpenSSL 3.0.2 15 Mar 2022)"
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.2 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] doParallel_1.0.17         iterators_1.0.14
##  [3] foreach_1.5.2             data.table_1.14.8
##  [5] textcat_1.0-8             quanteda.textplots_0.94.2
##  [7] quanteda.textstats_0.96.1 quanteda_3.2.4
##  [9] readtext_0.81             RColorBrewer_1.1-3
## [11] viridis_0.6.2             viridisLite_0.4.1
## [13] scales_1.2.1              ggplot2_3.4.1
## [15] rsvg_2.4.0                DiagrammeRsvg_0.1
## [17] DiagrammeR_1.0.9          magick_2.7.4
## [19] kableExtra_1.3.4          knitr_1.42
## [21] fs_1.6.1                  pdftools_3.3.3
## [23] stringr_1.5.0             mgsub_1.7.3
## [25] rvest_1.0.3               httr_1.4.5
## [27] RcppTOML_0.2.2            rmarkdown_2.20
##
## loaded via a namespace (and not attached):
##  [1] jsonlite_1.8.4      RcppParallel_5.1.7  askpass_1.1
##  [4] selectr_0.4-2       yaml_2.3.7          slam_0.1-50
##  [7] qpdf_1.3.0          pillar_1.8.1        lattice_0.20-45
## [10] glue_1.6.2          digest_0.6.31       tau_0.0-24
## [13] colorspace_2.1-0    htmltools_0.5.4     Matrix_1.5-1
## [16] pkgconfig_2.0.3     ISOcodes_2022.09.29 webshot_0.5.4
## [19] svglite_2.1.1       nsyllable_1.0.1     tibble_3.2.0
## [22] farver_2.1.1        generics_0.1.3      withr_2.5.0
## [25] cli_3.6.0           magrittr_2.0.3      evaluate_0.20
## [28] stopwords_2.3       fansi_1.0.4         xml2_1.3.3
## [31] tools_4.2.2         lifecycle_1.0.3     V8_4.2.2
## [34] munsell_0.5.0       compiler_4.2.2      proxyC_0.3.3
```

```
## [37] tinytex_0.44         systemfonts_1.0.4   rlang_1.0.6
## [40] grid_4.2.2           rstudioapi_0.14     htmlwidgets_1.6.1
## [43] visNetwork_2.1.2     labeling_0.4.2      gtable_0.3.1
## [46] codetools_0.2-18     curl_5.0.0          R6_2.5.1
## [49] gridExtra_2.3        dplyr_1.1.0         fastmap_1.1.1
## [52] utf8_1.2.3           fastmatch_1.1-3     stringi_1.7.12
## [55] Rcpp_1.0.10          vctrs_0.5.2         tidyselect_1.2.0
## [58] xfun_0.37
```

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2023. *Rmarkdown: Dynamic Documents for R.*

Analytics, Revolution, and Steve Weston. 2022. *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. 2023. *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. 2022. *Quanteda: Quantitative Analysis of Textual Data.* https://quanteda.io.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2022. *Quanteda.textplots: Plots for the Quantitative Analysis of Textual Data.*

Corporation, Microsoft, and Steve Weston. 2022. *DoParallel: Foreach Parallel Adaptor for the Parallel Package.* https://github.com/RevolutionAnalytics/doparallel.

Dowle, Matt, and Arun Srinivasan. 2023. *Data.table: Extension of 'Data.frame'.*

Eddelbuettel, Dirk. 2023. *RcppTOML: Rcpp Bindings to Parser for "Tom's Obvious Markup Language".* http://dirk.eddelbuettel.com/code/rcpp.toml.html.

Ewing, Mark. 2021. *Mgsub: Safe, Multiple, Simultaneous String Substitution.*

Garnier, Simon. 2021. *Viridis: Colorblind-Friendly Color Maps for R.*

———. 2022. *ViridisLite: Colorblind-Friendly Color Maps (Lite Version).*

Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2023. *Fs: Cross-Platform File System Operations Based on Libuv.*

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. 2023. *Textcat: N-Gram Based Text Categorization.*

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

———. 2022. *DiagrammeR: Graph/Network Visualization*. https://github.com/rich-iannone/DiagrammeR.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*.

Ooms, Jeroen. 2022. *Rsvg: Render Svg Images into Pdf, Png, (Encapsulated) Postscript, or Bitmap Arrays*.

———. 2023a. *Magick: Advanced Graphics and Image-Processing in R*.

———. 2023b. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*.

R Core Team. 2022. *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.

———. 2022a. *Rvest: Easily Harvest (Scrape) Web Pages*.

———. 2022b. *Stringr: Simple, Consistent Wrappers for Common String Operations*.

———. 2023. *Httr: Tools for Working with Urls and Http*.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*.

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.

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

———. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with Kable and Pipe Syntax*.