Universität Zürich UZH

DIS INTEGRATION
the mass politics of disintegration

# EU and Brexit Corpus:

A corpus of news articles on the EU and Brexit from major newspapers in thirteen European countries between 1992 and 2021

Marco Martini*, Martin Juan José Bucher^, & Stefanie Walter*
**University of Zurich*
*^ETH Zurich*

**Description of the data:**

Our EU and Brexit corpus is a collection of more than 2 million news articles from major newspapers in 13 European countries between 1992 and 2021, which are concerned with reporting on either the European Union, or Brexit, or both. The corpus has been compiled using the LexisNexis news API. The compilation has taken place in the context of the DISINTEGRATION project,[1] headed by Stefanie Walter, at the University of Zurich (https://www.disintegration.ch).

Our country selection is driven by the desire to achieve wide coverage across the membership of the EU. At the same time, our case selection is also dictated by data availability. In particular, the coverage of newspapers from Northern and Eastern European countries as well as smaller countries on LexisNexis is more limited. Our newspaper selection is based on the largest (in terms of circulation) nation-wide appearing dailies and weeklies for all the countries in our sample, that are available via LexisNexis. We generally aimed to achieve a rough balance in term of newspapers' political leaning within countries. In countries, this goal has been easier to achieve in countries with better data availability. In terms of time frame, we opted for coverage ranging back to the signature of the Maastricht Treaty. Due to more limited data availability in earlier years and the fact that 'Brexit' as a topic did not exit prior to 2015, our corpus is richer in the 2010s compared to earlier decades.

The corpus contains not only the text of the articles (separated into heading, subheading, lead, and body), but also a range of meta data (see below for more details). All articles are in original language. The corpus might be of particular interest for researchers seeking to understand media discourse on the EU and/or Brexit over the last decades in one or more of the countries in our data.

**Temporal coverage**:
01.01.1992 - 31.12.2021

[1] This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No 817582 (ERC Consolidator Grant DISINTEGRATION).

**Country coverage (ISO-3 codes):**
AUT, BGR, CHE, DEU, DNK, ESP, FIN, FRA, IRL, ITA, MLT, NLD, POL; NEW (newswires**)

**The corpus also contains reports from a range of large internationally active newswire services, which are treated as a separate category (i.e., NEW). See next section for details.

**News sources (by country):**

| | |
|---|---|
| AUT | Der Standard, Die Presse |
| BGR | 24 Chasa, Dnevnik, Kapital |
| CHE | Blick, NZZ, Sonntags Zeitung, Tages-Anzeiger, Weltwoche, 24 Heures, Le Temps |
| DEU | Die Welt, BILD, Der Spiegel, Die ZEIT |
| DNK | Politiken & Politiken Weekly |
| ESP | El Pais, El Mundo, ABC, El Correo |
| FIN | Kauppalehti |
| FRA | Le Figaro, L'Obs, Le Point, Les Echos, Sud Ouest, Sud Ouest Dimanche, Liberation |
| IRL | The Irish Times, Irish Independent, Irish Daily Mail |
| ITA | Corriere della Sera, ItaliaOggi, La Stampa |
| MLT | Malta Today, The Malta Business Weekly, The Malta Independent |
| NLD | De Volkskrant, Trouw, Algemeen Dagblad, NRC Handelsblad, De Telegraaf |
| POL | Gazeta Wyborcza, Newsweek Polska, Gazeta Prawna, Fakt Polska |
| NEW | Agence France Presse, Deutsche Presse-Agentur, GlobeNewswire, The Associated Press |

**Search terms (used to query the LexisUni API on the sources and dspecified above):**
The LexisNexis query underlying the data is based on the search string “EU|Brexit” for countries with non-latinian languages (e.g., Germany, Ireland) and “UE|Brexit” for countries with latinian languages (e.g. France, Italy). As a consequence, our search selects articles that contain either a reference to the EU or to Brexit (or to both).

We deliberately only used the two short forms “EU” and “UE” rather than the longer versions, because these short from allowed us to query all languages using only these two abbreviations. In contrast, the longer versions would have entailed more complex search strings in each national language, sometime including special characters that might have introduced errors. We pre-tested these decisions prior to compiling the corpus and found that that short forms find essentially all relevant articles on the EU, because the EU is almost always abbreviated in news articles (at least after the first mention).

**Total number of articles (*rows*):**
2’374’982

**Temporal coverage (number of articles per year)**:

| | |
|---|---|
| 1992 | 1757 |
| 1993 | 3063 |
| 1994 | 16943 |
| 1995 | 22947 |
| 1996 | 29287 |
| 1997 | 36801 |
| 1998 | 43956 |
| 1999 | 55870 |
| 2000 | 56270 |
| 2001 | 62106 |

| | |
|---|---|
| 2002 | 68204 |
| 2003 | 73705 |
| 2004 | 93220 |
| 2005 | 89510 |
| 2006 | 91958 |
| 2007 | 98870 |
| 2008 | 102967 |
| 2009 | 93969 |
| 2010 | 109165 |
| 2011 | 110028 |
| 2012 | 116949 |
| 2013 | 104900 |
| 2014 | 112017 |
| 2015 | 109456 |
| 2016 | 129873 |
| 2017 | 110537 |
| 2018 | 113554 |
| 2019 | 121031 |
| 2020 | 96357 |
| 2021 | 98553 |

**Country coverage (number of articles by country)**:

| | |
|---|---|
| AUT | 133206 |
| BGR | 864 |
| CHE | 192947 |
| CHF | 44388 |
| DEU | 125097 |
| DNK | 27770 |
| ESP | 360407 |
| FIN | 11120 |
| FRA | 155084 |
| IRL | 183620 |
| ITA | 194579 |
| MLT | 21665 |
| NLD | 239560 |
| POL | 94954 |
| NEW | 589721 |

**Available Text and Metadata (*columns*):**

| | |
|---|---|
| id | LexisNexis unique document ID |
| title_h1 | Article title |
| title_h2 | Article subtitle |
| author | Author name |
| published | Publication date |
| language | Publication language |
| text_lead | Text of the article lead (‘abstract’) |
| text_body | Main part of the article text |
| publication_name | Name of the news outlet |

| | |
|---|---|
| publication_type | Type of outlet (e.g. daily, weekly, tabloid) |
| publication_edition | Edition of the outlet (e.g. Sunday version) |
| publication_copyright | Copyright information |
| source_section | Section within the news source (e.g. politics, economy, sports) |
| wordcount | Total number of words in the article |
| country | Country code (ISO-3) |
| full_xml | Original XML data file as retrieved from LexisNexis |

**Available data formats:**

The corpus is available in two formats: 1) It can be accessed as a PostgreSQL database on a UZH server. This way it can be effectively queried and subset via SQL commands prior to further analysis. This requires some SQL programming skills, however. 2) alternatively, the corpus can also be immediately accessed as a set of in simple .CSV files organized by country.

**Terms of use:**

We are happy to make this corpus available to researchers, who would like to work with the data. For legal reasons, we cannot however publish the data online or share it with researchers outside of UZH. This is due to the copyright and licensing restrictions enforced by news publishing houses on LexisNexis and their clients. As a consequence, the data need to remain located on UZH servers and interested researchers need to be physically present on UZH premises to work with the raw (text) data.

Researchers interested in working with the data are invited to get in touch with the project PI, Prof. Stefanie Walter, to arrange the data access (walter@ipz.uzh.ch). In the process, it will be necessary to sign a consent form, which specifies policies concerning data access, data handling, and non-disclosure.

**References:**

- Martini, Marco & Stefanie Walter (2023). "Learning from Precedent: How the British Brexit Experience Shapes Nationalist Rhetoric outside the UK." Journal of European Public Policy (*forthcoming*)