Twitter:
@BerLinguistin

Mastodon:
mas.to/@BerLinguistin

Open Science and Me

Dr Naomi Truan
Assistant Professor in German Sociolinguistics
Leiden University

Workshop goals

- What does open science mean concretely for *you*?
- What steps could you take in the pursuit of open science?
- Why would you take them (or not)?

⇒ focus on **research ethics**
rather than on the technicalities
(going meta!)

Introduction

Open science (OS) unites highly diverse practices:

(a) sharing resources (e.g., code, data, research materials, methods);

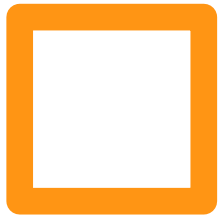
(b) publishing in alternative formats, such as uploading preprints (i.e., manuscripts that have not yet been subjected to a peer-review evaluation) to an open repository or to one's personal website;

(c) sharing research questions and methodologies

(Allen & Mehler, 2019: 2)

Open Data

- "Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike." (Open Data Handbook)
- in linguistics specifically: "overall low rates of sharing materials, data, and protocols" (Bochynska et al. 2022: 1)



Open Access

- "Open access (OA) means free access to information and unrestricted use of electronic resources for everyone." (UNESCO)

A mini biography

My Open Science Journey

- *not* an open science advocate from the very beginning
- a **guided** process
- evolution from indifference to **commitment**
- now something that I do almost automatically, but still **at my own pace**
- just like anything, you have to find **what works for you**, i.e. you have to be able to stand behind it

Not covered today (but important to know/check later!)

Choosing a **license**: https://creativecommons.org/

There is an **excellent independent publishing house in Open Access in linguistics** (with open peer review!): https://langsci-press.org/

Many universities and librairies **cover the costs of Article Processing Charge (APC)**, but we want to learn what is possible without it

Data management plans (DMD): most universities offer workshops / help when applying for grants!

Picture by Lubos Houska on Pixabay

Questions to ask yourself to begin with

1

What could you open (more)?

raw data, annotated data (with metadata), reflexions / blog posts, conference slides, course materials, papers?

2

How could you open it more?

includes the STEPS needed to become (more) open + a reflection on WHERE (on which platforms/repositories)

3

What has prevented you to do so until now?

lack of time, resources, knowledge, confidence, …

4

What would it take for you to become more open?

feeling empowered, knowing that you're allowed to, dedicating an afternoon to the first step, making a plan knowing how it could benefit you, …

5

Which first step could you take today?

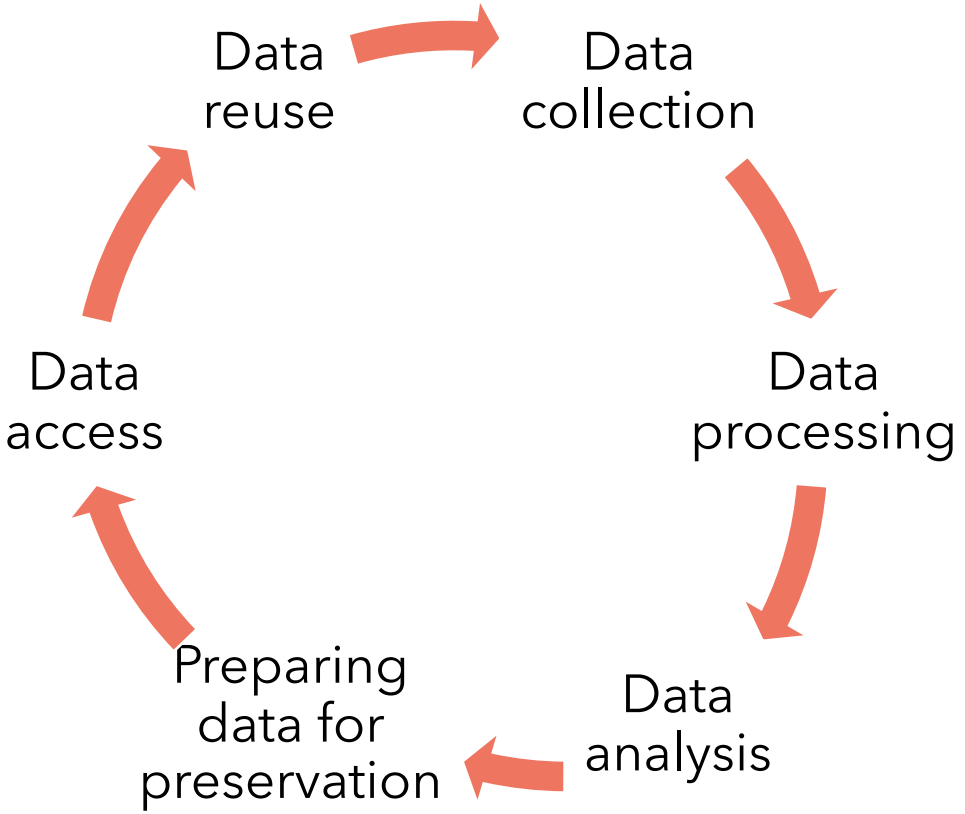
talking to someone, registering for a platform, beginning the description of your metadata, …

Topic one

Open Data

The Research Data Lifecycle Reference Model

UK Data Service



Open Data applied to you

Self reflection

- What **kind of data** are you working with?
- Is your data **'yours'**? Was it produced (collected/annotated/enriched) by you?
- Could you make your data open? Are there any **restrictions** (qualitative interviews, data are part of a project, etc.)?
- What has prevented you until now to make your data open? Time, resources, fears, lack of knowledge… List all your current **limitations**.

What kinds of data?

see: "The range of linguistic data" by Jeff Good (2022)

- "**Raw data** was defined as recorded information in its rawest, digital form, at the level of sampling units (e.g., participants, words, utterances, trials, etc.)." (Bochynska et al. 2022: 10)
- "**Processed data** was defined as a derived form of the data that has undergone changes from its raw state (e.g., extraction of acoustic parameters via Praat, aggregates of responses, etc.)." (Bochynska et al. 2022: 11)
- "**Linguistic annotation** is a kind of linguistic data that 'involves the association of descriptive or analytical notations' with other kinds of language data (Ide 2017:2). Annotation can either be made directly on 'raw' data (i.e., unannotated language data) (see, e.g., Schultze-Berndt 2006:215; Himmelmann 2012:188) or on other annotations." (Good 2022: 31)
- "The term **metadata** is generally applied to data about other data when the new data are not seen as directly supporting further analysis." (Good 2022: 41)

Open Data: WHY?

Why can/should we make our data openly available? (1/2)

- not only because we have to (e.g. for grants)!
- research ethics:
  - the preparation of data for analysis is, in itself, an **integral part of the research process**: "Because transcription is an act of interpretation and representation, it is also an act of power" (Bucholtz 2000: 1463)
  - the scientific community should be able to **check the data (in its context)**: In a PhD thesis or a paper, constrained by the lack of space, we often present only extracts that are truncated for the needs of the demonstration (= which say what we want them to say)

Why can/should we make our data openly available? (2/2)

- Direct benefits for you at the PhD level:
  - preparing your data in an open way encourages you to **systematize and record your choices precisely**—which will greatly facilitates not only the writing (in progress) of the thesis itself, but also the methodology
  - you will **gain place in papers** -> you can point to the persistent identifier created by national archiving platforms instead of having to describe the corpus each time
  - you may **become visible through your open data** even before your first publications

to be continued…


Naomi Truan
@BerLinguistin
I just discovered randomly while looking for something else that my manually annotated corpus of parliamentary debates has been used in a PhD thesis!
If you can, share your data, not only your analysis — you never know whose work you may make easier ☺️ #OpenScience
3:33 PM · Jan 26, 2022
View Tweet analytics
3 Retweets
25 Likes

Open Data: HOW?

The FAIR principles

see: article in *Nature* in 2016 & FAIR principles

1

Findability

2

Accessibility

3

Interoperability

4

Reuse of digital assets

The FAIR principles (1/4)

- FINDABLE:
  -> globally unique and persistent **identifier**: an example is your ORCID number, the link should remain active (thus, repository!)
  -> data **+** metadata: never the data alone, but always with its **extensive and generous documentation**
  -> identifier of the data + metadata: if separate files, the **association between a metadata file and the dataset** should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata
  -> registered or indexed in a **searchable resource**: not (only) on a personal blog, a Padlet, or Academia.edu / Research Gate (which are not open access anyways), make your data set **discoverable** (also by indexing machines)

The FAIR principles (2/4)

Tromsø recommendations for citation of research data in linguistics

- ACCESSIBLE:

  -> (meta)data are **retrievable** by their identifier: most users of the internet retrieve data by 'clicking on a link' (http(s))

  -> the protocol is open, free, and universally implementable: to maximise data reuse, the protocol should be **free (no-cost)** and **open (-sourced)** -> should impact your **choice of repository**, for instance, Skype is proprietary (so, not what we want)

The FAIR principles (2/4)

- ACCESSIBLE:

  -> the protocol allows for an authentication and authorisation procedure, where necessary: IMPORTANT! The 'A' in FAIR does not necessarily mean 'open' or 'free'. Rather, it implies that one should provide the **exact conditions under which the data are accessible**. Hence, even heavily protected and private data can be FAIR.

  -> metadata are accessible, even when the data are no longer available: **metadata are valuable in and of themselves**, when planning research, especially replication studies. Even if the original data are missing, tracking down people, institutions or publications associated with the original research can be extremely useful.

The FAIR principles (3/4)

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form." (https://tei-c.org/)

- INTEROPERABLE:
  - -> (meta)data use a **formal, accessible, shared, and broadly applicable language** for knowledge representation: data that should be **readable for machines**
  - -> (meta)data use vocabularies that follow FAIR principles
  - -> (meta)data include qualified **references to other (meta)data**: you should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described

The FAIR principles (4/4)

- REUSABLE:
  -> (meta)data are **richly described** with a plurality of accurate and relevant attributes: see next slides
  -> (meta)data are released with a clear and accessible data usage **license**: what **usage rights** do you attach to your data?
  -> (meta)data are associated with detailed provenance: include a **description of the workflow** that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else that you may have transformed or completed?
  -> (meta)data meet domain-relevant **community standards**

Open Data applied to you

Discuss

- In the (open) **corpora you are using**, how does the metadata look like?
- Which **metadata** is (should be) associated with your data?
- How could **your metadata** look like? What would you have to explain to make your data intelligible to someone who does not know anything about your research?

Preparing data in order for it to be open

Extracts from "Research data management", UK Data Service

- "Documentation deposited alongside data files should enable users, **with no prior knowledge of the research project and data collected**, to understand exactly how the research was carried out and what the data mean, in order to **(re)use the data correctly in their respective projects and for their respective purposes**."
- "This requires **clear and detailed data description and annotation**. Besides the information that is needed to reuse the data, data also need to be accompanied by information for citing and discovering the data."
- "To prepare data for secondary research, researchers should document data appropriately. They should also **explain the procedures and fieldwork methods, the objectives and methodology of the research, and explicitly describe the meanings of variables and codes used**."
- "Additionally, they should **describe any derivation, transformations, de-identification (pseudonymization/anonymization) or data cleaning** carried out."

“REUSABLE” from the FAIR principles

Examples from go-fair.org

- Some points to take into consideration (non-exhaustive list):
  - Describe the **scope of your data**: for **what purpose** was it generated/collected?
  - Mention any **particularities or limitations** about the data that other users should be aware of.
  - Specify the **date of generation/collection** of the data, the lab **conditions**, **who** prepared the data, the **parameter settings**, the name and version of the **software** used.
  - Is it **raw or processed** data?
  - Ensure that **all variable names** are explained or self-explanatory (i.e., defined in the research field’s controlled vocabulary).
  - Clearly specify and document **the version** of the archived and/or reused data.

Preparing data for it to be open (1/3)

- be open about the **research question** and the issues underlying the **composition and annotation of the corpus**
- in my case: each of the online corpora can be studied independently of the others, but the corpora also work together (FR, DE, UK)
- the contrastive dimension had significant consequences on the **annotation choices** (e.g. the coding of unauthorized interventions in parliament)

- document the **process** (not only the end product!)
- release a first version **as soon as possible (*not* only once you have published on it!)**
- be ready to make some **changes** / update the data and/or metadata

Preparing data for it to be open (2/3)

- clearly state what one would expect to find, **what is 'missing'**, why it is not there
- for example, the **contrastive aspect** of my corpus has had several **implications in terms of (non) coding**, but I addressed this aspect only in a later data paper (Truan & Romary 2022)

Important notice:

The French corpus has been added after the German and British corpora had been collected and annotated. I followed the same protocol to look for the national plenary debates in France, but many debates are missing. Nevertheless, the corresponding European Councils are still in the table (see document "Description of the French Corpus"), so that the comparison with the German and British corpora remains possible.

Descriptor: speech type
(debate, interruption, vote explanation, etc.)
"From a linguistic point of view, this descriptor, which is not included in the data model of the corpus provided by (Truan, 2017), is particularly important when it comes to differentiate effects of register variation ranging from highly formulaic to less formal speech (as in the case of e.g. interruptions)." (Diwersy, Frontini & Luxardo 2018)

Preparing data for it to be open (3/3)

- VERY IMPORTANT: **choose a license*** (ask your librarian/Open Science officer)
- tell people **how to quote** the corpus

Screenshot of the bibliographic information for the corpus of German parliamentary debates online on ORTOLANG https://www.ortolang.fr/market/corpora/de-parl

Tromsø recommendations for citation of research data in linguistics

* like I'm doing for these slides!

Topic two

Open Access

Debunking a few myths

Discuss

- Have you already published in open access? If so, what kind of publication was it? If not, why didn't you?
- What *counts* as a publication for you? Why or why not?

Common Misconceptions about Open Access (1/4)

- it is **expensive**
- it is **time-consuming**
- it is **low quality**
- it consists in putting your research papers on **Academia.edu** or **Research Gate**
- it is only a **reaction external pressure** ("You need to be visible", "You agreed on publishing everything in open access to get this grant")
- it is **worthless** / doesn't bring anything

Common Misconceptions about Open Access (2/4)

- it is **expensive**
  - only gold Open Access is (when it includes an article publishing charge or APC)
  - OA also (and maybe foremost) means making your research available on open repositories, e.g. the preprints (without formatting from the journals)

- it is **time-consuming**
  - yes, it can definitely be *at the beginning*
  - once you have chosen an open repositor, uploading your preprint, postprint, or open access paper is a matter of minutes! It's like writing your academic CV for the first time and then only having to update it

Common Misconceptions about Open Access (3/4)

- it is **low quality**
  - "Many people associate OA with low quality, basically because they attach OA to relatively new journals [...] This view is misleading because it reduces OA to journals…
  - … while you can publish in OA afterwards, after being published in a non OA journal."

https://icietla.hypotheses.org/872

Common Misconceptions about Open Access (4/4)

- it consists in putting your research papers on **Academia.edu** or **Research Gate**
  - ➢ these are NOT open access platforms, but commercial ones!!

https://icietla.hypotheses.org/114

- it is **worthless**
  - ➢ in fact, publishing in OA increases visibility + reception (including a higher number of citations) [see "Effective Strategies for Increasing Citation Frequency". I didn't have time to look for sources specific to linguistics for this, but this has been shown in a vast variety of disciplines, for instance for PhD theses]


Thread
Naomi Truan
@BerLinguistin
Little reminder: As an author, you're *ALWAYS* allowed to upload the author's version of your text to an open repository, before, during, or after peer-review, NO MATTER what the journal policy says! Your text is yours & remains yours, go share it with the world! 🙏 #OpenScience
Naomi Truan @BerLinguistin · Apr 29, 2021
You've written a paper, especially in the humanities and social sciences, and you're not sure where and when you're allowed to publish it in #OpenAccess? This mini thread on #OpenScience is for you, #AcademicTwitter! 🤓🔓 twitter.com/BerLinguistin/...
4:21 PM · Jan 11, 2022
View Tweet analytics
9 Retweets
1 Quote Tweet
48 Likes

Open Access applied to you

Self reflection

- What is this one paper you could make available right now but have not until now? Why is that?
- Will you publish link to your paper on your personal website / university page? If so, what steps are necessary *before*?

A mini guide (1/2)

See: https://twitter.com/BerLinguistin/status/1297967280920440833

- PRE-PRINT: "The pre-print is the **author's manuscript version** of the publication that has been submitted to a journal for consideration for publication."
  - ➢ so, **not accepted (yet)**
- POST-PRINT: "The post-print is the **author's final manuscript** [...], contains **all revisions** made during the peer-review process. It does not, however, reflect any layout or copy editing done by the publisher [...]"
  - ➢ the content *with* revisions but *no* layout
- PUBLISHED VERSION: "the **final version of the article produced by the publisher**": "the printed version found in books, proceedings and journals" or "a PDF"
  - ➢ usually the version with a 12-month embargo in the humanities (i.e. may be made available only after 1 year)

A mini guide (2/2)

See: https://twitter.com/BerLinguistin/status/1297967280920440833

- The policy may depend on the publisher (check on https://v2.sherpa.ac.uk/romeo/) but what is sure is:
- the PRE-PRINT is **always** yours (**basically it's your text**) and you can publish it whenever/wherever you want (and get feedback *before* submitting to a journal, for instance)
- the POST-PRINT, the final paper, accepted for publication, with revisions after all the (nice and useful) comments from reviewer 2, but not the final pagination and maybe 1-2 typos **may be uploaded as well** (because tbh the publisher hasn't put too much effort in it!)
- the PUBLISHED VERSION is not entirely yours anymore, because the publisher has formatted it according to the journal's policy & it looks nice!

Open Science for Everyone?

Finding *your* way in neoliberal, classist, patriarchal, ableist (etc.) academia

Dealing With Being Exposed: Setting Boundaries While Being Open

- Examples of **how vulnerability manifests** include, but are not limited to:
  - working on **sensitive (often critical) topics** that mat be difficult to bring in mainstream (highly reputed) journals, thus having many preprints "out there" that remain unpublished in a traditional sense;
  - increasing, through Open Access publications, **the chance of being read and thus potentially criticized**, especially for projects at an early stage / before the PhD defense;
  - not being supported by one's supervisor(s) in one's Open Science journey, which may be seen as a waste of time or not prestigious enough, in worst cases as narcissism, and thus **damage the person's reputation**;

https://icietla.hypotheses.org/2771

Dealing With Being Exposed: Setting Boundaries While Being Open

- Examples of **how vulnerability manifests** include, but are not limited to:
  - being afraid of professional and academic consequences **if research findings or conclusions are not as solid or unequivocal as previously thought** (this includes the fact that the author themself may change their mind);
  - being subject to **pressures** from peers and/or supervisors and not being able to weigh in when it comes to publication strategies (basically, not having a word to say);
  - being **misunderstood when advocating for Open Science**, for instance by scholars who mistakenly think that Open Science is restricted to publishing in an Open Access journal or paying the Open Access fees, etc.

https://icietla.hypotheses.org/2771

Going Open Access is Like Transitioning Into a Vegetarian Lifestyle

- It doesn't have to be all or nothing and you'll be fine.
- Nobody forces you to make anything available online that is just not ready (yet). By putting things in open access, you **take the control back on your publications**. You decide what to publish, when, and where (i.e. on which platform).
- You ensure the dissemination of your work **on your own terms**, following **your agenda** and **your priorities**.

https://icietla.hypotheses.org/872

Bonus

To delve into the topic later

Open Repositories

Examples of specialized corpora in OA:
- political discourses (Barbaresi 2018)
- parliamentary corpora (CLARIN)

Open Data (OD)

- Ranking Web of Repositories
- Directory of Open Access Repositories: "OpenDOAR is the quality-assured, global Directory of Open Access Repositories. You can search and browse through thousands of registered repositories based on a range of features, such as location, software or type of material held."
- ORTOLANG: "Platform of linguistic tools and resources for an optimized treatment of the French language"
- Tromsø Repository of Language and Linguistics (TROLLing)

Open Access (OA)

- MetaArXiv preprints: "An interdisciplinary archive of articles focused on improving research transparency and reproducibility"
- OSF: "OSF is a free, open platform to support your research and enable collaboration."
- SocArXiv: "Open archive of the social sciences SocArXiv papers are moderated before appearing."
- HAL-SHS: "The open archive HAL SHS is devoted to archiving and dissemination of scientific literature, published or unpublished, from universities or research institutions in all disciplines of human and social sciences."

➢ The one I am using, very friendly and responsive team

Resources & Links

Farran, Emily K., Dr Priya Silverstein, Aminath A. Ameen, Iliana Misheva & Camilla Gilmore. 2020. Open Research: Examples of good practice, and resources across disciplines. OSF Preprints. https://doi.org/10.31219/osf.io/3r8hb.

Open Data (OD)

- Linguistic Linked Open Data
- The Open Handbook of Linguistic Data Management
- Reproducible research practices and transparency across linguistics
- Dealing With Being Exposed: Setting Boundaries While Being Open
- How sharing my data has changed how I write
- Paper on building and annotating a corpus of parliamentary debates (methodology)
- Step-by-step guide

Open Access (OA)

- Open Access Network
- Online library and publication platform: "OAPEN promotes and supports the transition to open access for academic books by providing open infrastructure services to stakeholders in scholarly communication."
- Directory of Open Access Books
- Directory of Open Access Journals and Articles
- Open Edition Books (mostly in French)
- Think Check Submit: "Identify trusted publishers for your research"
- How to make the most of your publications in the humanities?
- Open Scientist Handbook

Resources & Links

Open Educational Resources (OER)

- OER Commons: https://www.oercommons.org/
- Blog series on teaching Open Science in Linguistics: https://icietla.hypotheses.org/open-science-in-education
- Peer-reviewed papers by myself on the topic:
  - https://www.herausforderung-lehrerinnenbildung.de/index.php/hlz/article/view/4343/4598 (seminar concept, in German)
  - https://www.irrodl.org/index.php/irrodl/article/view/6201 (interpretation of the students' attitudes towards OS)
  - https://riojournal.com/article/86663/ (autoethnographic insights from teachers in higher education in Germany)
- Examples of OER in Linguistics:
  - YouTube: https://www.youtube.com/@AlexanderLasch
  - Padlet: https://padlet.com/berlinguistin/fbrhor1lir39mtaz / link on OER Commons

Inspiration


Picture by Joe on Pixabay

- Open Science Radio (podcasts)
- Val.Es.Co (most important Spanish research group on informal language with a huge online corpus) is currently tweeting a lot about methodological issues related to informal language collection, annotation, analysis*
- Platform on Spanish linguistics and language teaching, which also serves some OA purposes, but not only*
- Database to mainly Spanish-speaking journals and theses and, whenever also OA, it gives a direct link*


Picture by Tran Mau Tri Tam on Pixabay

* Thank you to Barbara De Cock for the tips!

Thank you!

Dr Naomi Truan

Assistant Professor in German Sociolinguistics, Leiden University

Fellow of *Freies Wissen* in 2020-2021

Winner of the Open Science Research Data Award of the French Ministry of Higher Education, Research and Innovation in the category “reuse of data” in 2022

n.a.l.truan@hum.leidenuniv.nl

Website: https://naomitruan.wordpress.com/

Blog: https://icietla.hypotheses.org/

Publications in Open Access: https://cv.hal.science/naomi-truan