Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility

Naomi Truan, Sophie Granger, Jo Lychnara

Research Data Management Workflow

Pseudonymizing Interviews in the Humanities & Social Sciences

Naomi Truan, Sophie Granger, Jo Lychnara

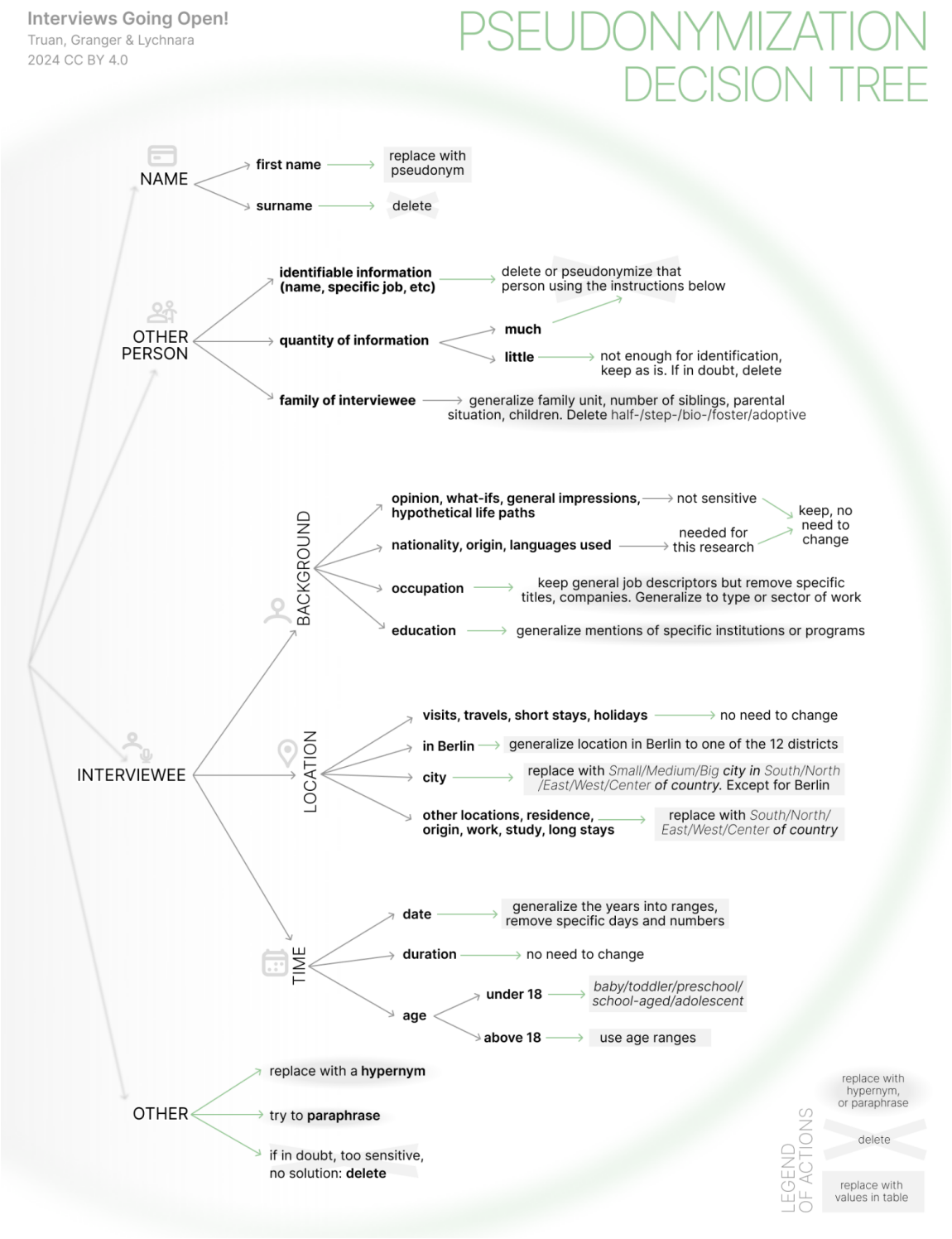

Fig 1: Overview of the pseudonymization scheme 'Interviews Going Open!'

Figure by Sophie Granger

The goal of this **technical document** detailing the pseudonymization scheme is to outline the **specific methods and processes used to pseudonymize data** in a way that ensures privacy while maintaining data utility for research or analysis.

We define the **types of identifiers removed or masked**, describe the techniques applied to reduce the risk of re-identification, and provide guidelines on managing indirect identifiers.

A separate document (Truan, Granger & Lychnara 2024) presents a step-by-step description of the research data workflow with time stamps: ⟨halshs-04743263⟩.

Cite as: Truan, Granger & Lychnara 2024, Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility.
Last update: 17.12.2024

Project Overview

Pseudonymizing Interviews in the Humanities & Social Sciences

Title: Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility

Keywords: qualitative research, interviews, sensitive information, pseudonymization, anonymization, guidelines, funding, time management, grants

Key values: open science, open data, FAIR principles

Project (origin, grants): This document has been produced as part of the Leiden University Centre for Digital Humanities Small Grant 2024 for the project "Interviews Going Open! Developing Interdisciplinary Guidelines on How to Publish Qualitative Interview Data as Open Data".

Authors: Project leader: Dr Naomi Truan; Research Assistants: Sophie Granger and Jo Lychnara. See below for Author contribution CRediT statement.

Abstract: The project *Interviews Going Open!* addresses the challenge of sharing sensitive interview data within the humanities and social sciences. Interviews are valuable sources of information, but concerns about privacy and ethical considerations often hinder their open accessibility. Another common challenge is how to make "thick description" (a property of qualitative fieldwork) suitable for metadata (which is often criticized as oversimplification in qualitative research). This handout documents the process, including the number of hours needed for each task and our questions as they arise. In doing so, we identify challenges faced by qualitative researchers and offer lowkey, realistic, practical, and tailored solutions based on a corpus of 25 interviews in a small and possibly easily recognizable community.

Author contribution CRediT statement: (see Allen, O'Connell & Kiermer 2019)
* denotes equal contribution
Conceptualization: Naomi Truan
Project Administration and Supervision: Naomi Truan
Writing - Original Draft Preparation: Naomi Truan, Sophie Granger, Jo Lychnara
Writing - Review & Editing: Naomi Truan, Sophie Granger, Jo Lychnara
Pseudonymizing: *Sophie Granger, *Jo Lychnara
Data management: *Naomi Truan, *Jo Lychnara
Data formatting: *Sophie Granger, *Jo Lychnara
Coding: Sophie Granger *Figures*: Sophie Granger
Acknowledgments: Andrew Hoffman & Femmy Admiraal advised us during the process.

Table of Contents

Pseudonymizing Interviews in the Humanities & Social Sciences

Pseudonymization Scheme

Pseudonymizing Interviews in the Humanities & Social Sciences

The interview data

The data used to experiment with this workflow originates from a research project on the language ideologies of French speakers who also use German in Berlin[1] (Truan 2024; Truan forthcoming). The corpus consists of 25 semi-structured interviews in French (mean length: 50') which were audio- or video-recorded. The corpus amounts to around 30 hours of recorded interviews, transcribed in 160,333 words or 477 pages (A4, Word).

The interview data in a nutshell:

- 25 semi-structured interviews in French
- mean length: 50'
- around 30 hours of recorded interviews
- 160,333 words or 477 pages

The data collection and transcription were handled before this project, while Naomi Truan was working at the University of Leipzig (grant: Leipzig Flexible Fund; student assistants: Jun An Chen, Matilde Mondo). The interviews were transcribed thoroughly (word level), but may still present minor inaccuracies (e.g. inaudible words).

Disclaimer: *The interview data is provided in Word format rather than an interoperable format such as XML or TEI due to time constraints and a limited budget. While XML or TEI would offer enhanced flexibility and long-term preservation benefits, the resources required to implement these formats were not available during the project. Prioritizing the timely completion of the pseudonymization, we opted for a widely accessible and manageable format that meets the immediate needs of the project. However, meeting interoperability standards remains the goal for future data management.*

The second phase, which is the object of this workflow, was made possible thanks to the LUCDH Small Grant 2024, with the pseudonymization workflow as a target.

Micro and macro pseudonymization

Our workflow targeted small pieces of sensitive information, **at the sentence level**. Being short, these could conveniently be put into rows and cells, streamlining the workflow.

[1] Hence the special status of Berlin when it comes to pseudonymizing locations (see decision tree, p. 1).

***Micro edits* (one line, on sentence level)**

These edits do not affect the general reading flow of the text.

- **"Deleted"**: information that is too sensitive and cannot be easily pseudonymized; information about people mentioned by the interviewees.

Transcribed: You can contact me further on @sarahgermanylife.
Pseudonymized: You can contact me further on {sensitive data deleted}.

Transcribed: My colleague at the bakery comes from Luxembourg and lives now with a French roommate in Potsdam.
Pseudonymized: My colleague {sensitive data deleted}.

- **"Replaced"**: following a rule or a table, the sensitive data was substituted with an equivalent information point.

Transcribed: John moved there in the Fall.
Pseudonymized: Denis moved there in the Fall.

Transcribed: I'm 36 years old.
Pseudonymized: I'm 35-41 years old.

- **"Paraphrased"**: not a simple substitution, outside of the scope of the replacement tables, overall sentence structure had to be changed.

Transcribed: I moved with my mother after her divorce.
Pseudonymized: I moved with my mother after a change in her personal situation.

Transcribed: He dislikes he's called Smith now at work.
Pseudonymized: He dislikes he's called an english family name now at work.

Macro edits (multiline, on paragraph level)

In other projects, pseudonymization edits could happen on a bigger scale, e.g. a **paragraph containing an abundance of sensitive information**. These situations require other approaches. DeLacey (2024) described two pseudonymization methods to deal with sensitive data on a larger scope:

- **“Summarize”**: take a bigger part that is sensitive and make a summary of it. The summary only contains the non-sensitive information, or vaguely explains the sensitive information without the details. This approach can be used when the texts become incomprehensible due to the removed parts or if the editing will be too time consuming.

- **“Quote”**: for sensitive information that for the research purpose has to be kept, some obfuscation can be created by separating the sensitive elements. One person could be split into two distinct pseudonyms, given the impression that we are dealing with two different interviewees. Instead of targeting the sensitive information, remove the information that makes it possible to connect the two parts to each other. Make sure that the metadata does not show this truncation (atypical length, same date and place). The smaller part, despite its sensitive nature, can now be used in the paper with less risks of identification, being smaller it contains less context.

We opted not to create a summary of full paragraphs, because we wanted to stay as close to the data as possible, believing that while summarizing would preserve the content, it would sacrifice the nuanced wording essential for linguistic analysis.

Categories

This part shows the specific guidelines we used for our project the language ideologies of French speakers who also use German in Berlin (Truan 2024; Truan under review) , for example with a category about places in Berlin.

Outliers always occur, sometimes we had to deviate a bit from the instructions (e.g. paraphrase instead of replace) to be able to deal with that sensitive information. In the examples some of these difficult cases are shown.

People

Names

Refers to names mentioned in the transcript whether they are of the interviewees or other individuals.

Instruction: Keep an alternative name and stay consistent throughout the document. Recurrent names should be replaced with pseudonyms. Remove the family names. Note the new names in a key file, ideally kept on separate server.

Key file table:

original name	pseudonym
John	Denis
Sarah	Emma
Smith	[none]

Tip: See under "Automated tasks" the subsection "Search and replace" for tips on how to prevent problems when changing the names.

Example with key file replacement:

Transcribed: John and Sarah Smith met in Berlin. John moved there in the Fall, and Sarah followed a year later.
Pseudonymized: {Denis: *sensitive data replaced*} and {Emma: *sensitive data replaced*} met in Berlin. {Denis: *sensitive data replaced*} moved there in the Fall, and {Emma: *sensitive data replaced*} followed a year later.

Example with deletion:

Transcribed: To continue the story, the Smith couple met in Berlin.
Pseudonymized: To continue the story, the {sensitive data deleted} couple met in Berlin.

In some cases, finding the right micro-edit to keep the context or sense can be tricky:

Example with paraphrasing:

Transcribed: The Smith couple met in Berlin.
Unclearly pseudonymized: The {sensitive data deleted} couple met in Berlin. (which people in the story?)
Pseudonymized: The {interviewee's friends, previously mentioned in 11:54:59: *sensitive data paraphrased*} couple met in Berlin.

Example with hypernym replacement:

Transcribed: He dislikes he's called Smith now at work.
Unclearly pseudonymized: He dislikes he's called {sensitive data deleted} now at work. (called what?)
Pseudonymized: He dislikes he's called {an English family name: *sensitive data paraphrased*} now at work.

Place of origin

Helps define a person's origin and identity prior to migration. Although this can play into the identifiable data, this is a biographical point that a lot of people have in common.

Instruction: Use generic regions (North/South/East/West/Central of country X) instead of specific places of origin unless otherwise stated.

Transcribed: I grew up in Mainz.
Pseudonymized: I grew up in a {small town in West Germany: *sensitive data replaced*}.

Transcribed: Due to the crisis in Libya, we quit the city.
No change incorporated

Nationality

Ties to the cultural aspects of identity.

Instruction: Keep all instances where nationality is mentioned, as this information is relevant to analyze the data and key to the project on migration and multilingualism.

Transcribed: My mother is German
No change incorporated

Languages used

Reflects cultural and linguistic background.

Instruction: Keep all instances of languages used since it is a key aspect of the (in our case, sociolinguistic) analysis.

Transcribed: I speak German, English, and French fluently.
No change incorporated

Subjective data

Non-factual data are not considered sensitive.

Instruction: Keep all instances of opinions, what-ifs, impressions about the person, hypothetical life paths.

Other people

Refers to mentions of individuals in the narrative who were not interviewed, individuals mentioned by the interviewees, but these did not give their consent.

Instruction: Often these people have little identifiable information in the interviews compared to interviewees. Thus in some cases, information should be deleted without extra thought, but often it is not sensitive and can be left as is.

Name Give pseudonym	Example with pseudonym replacement:

Recurrent name

Put pseudonym in the key file.

Personal information (specific job, place of living)

Delete

Very little sensitive information:

Not enough for identification, keep as is. If in doubt, delete.

Much sensitive information

The person could be traced back by using a search engine.

There are enough details for them to be recognized by common acquaintances.

Delete or pseudomize that person using the instructions below for interviewees.

Transcribed: My neighbor, John, never moved out
Pseudonymized: My neighbour, {Jack: *sensitive data replaced*}, never moved out.

Example with no deletion:

Transcribed: My friend from London.
No change incorporated

Example with deletion:

Transcribed: My colleague at the bakery comes from Luxembourg and lives now with a French roommate in Potsdam.
Pseudonymized: My colleague {sensitive data deleted}.

Family

The type of familial citation can be very specific to a person. Can influence personal motivations and experiences.

Instruction: Generalize family unit, number of siblings, parental situation, children.

my brother and sister → *my siblings*
half-sibling, step-sibling → *sibling*
adoptive parent, step-parent, foster parent, bio-parent → *parent*
I have one daughter → *I have a child*

Occupation

Defines the person's professional life.

Instruction: Keep general job descriptors but remove specific titles, companies or institutions. Generalize to type or sector of work.

Transcribed: He worked as a software engineer at Google.'
Pseudonymized: He worked as a software engineer at {a tech company: *sensitive data replaced*}.

Transcribed: I work in a consulting firm for the local municipality.
Pseudonymized: I work in {the public sector: *sensitive data paraphrased*}

Transcribed: As an aircraft engineer
Pseudonymized: As {a technical role: *sensitive data replaced*}

Transcribed: I'm making a living from dancing.
Pseudonymized: I'm making a living from {entertainment domain: *sensitive data paraphrased*}

Education

Key for professional and intellectual background.

Instruction: Pseudonymize mentions of specific institutions or programs.

Replacement micro edit table:

education corresponding to life phase	**replaced data**
0-6	Pre-school
6-12	Primary school
12-18	Secondary school
18+	Hypernym*

* For words such as *technical university*, *bachelor*, MBA, and *traineeship*, you may use general words such as *training* or *education*.

Example with replacement:

Transcribed: I discovered that at the Kita Zebrablume.
Pseudonymized: I discovered that at the {pre-school: *sensitive data replaced*}.

Example with hypernym replacement:

Transcribed: We studied linguistics at UvA.
Pseudonymized: We studied linguistics at {a university in the Netherlands: *sensitive data replaced*}

Example with hypernym replacement:

Transcribed: When I was doing my bachelor's propedeuse,
Pseudonymized: When I was doing {my study's first year: *sensitive data replaced*},

Example with paraphrasing:

Transcribed: For this job I needed a LLM.
Pseudonymized: For this profession I needed a {specific degree in law: *sensitive data paraphrased*}.

Time indicators

Dates

Significant for understanding the timeline of life events.

Instruction: Keep but generalize the years into ranges. Remove specific days numbers.

Note: Dates about other persons should be treated differently, as they did give consent about their personal information, see “other people”.

Replacement micro edit table:

original data	replaced data
xxx0	Early x0s
xxx1 to xxx5	Early decade
xxx6 to xxx9	Late decade

Example with replacement:

Transcribed: My birthday is 5 June 2000. I moved again in September 2007.
Pseudonymized: My birthday is {early 2000s: *sensitive data replaced*} I moved again in September {late 2000s: *sensitive data replaced*}.

Age Groups

Age is an important factor of an understanding of the interviewees’ life events.

Instruction: Generalize age into categories.

Underage groups

The age ranges for underage people are based on the language acquisition phases

Replacement micro edit table:

original data	replaced data
0 to 1	baby
1 to 2	toddler
2 to 6	preschool
6 to 12	school-aged
12 to 17	adolescent

Example with replacement:

Transcribed: I was five when we moved.
Pseudonymized: I was {preschool aged: *sensitive data replaced*} when we moved.

Adult groups

Provides context for life experience.

Instruction: Categories made in an arbitrary manner per decennia; use age range.

Replacement micro edit table:

original data	replaced with range
18 to 24	18-24
25 to 34	25-34
35 to 44	35-41
45 to 54	45-54
55 to 64	55-64
65 and after	65+

Example with replacement:

Transcribed: Today I'm 35 years old.
Pseudonymized: Today I'm {35-41: *sensitive data replaced*} years old.

Example with paraphrasing:

Transcribed: When I became 18, I could finally handle that paperwork.
Pseudonymized: When I became {adult: *sensitive data paraphrased*}, I could finally handle that paperwork.

Ages of non-interviewees

The interviewees also share other stories than their own, such as their siblings or children's experience with languages.

Instruction: See part "Other people". If the age of the other person is needed, follow the above tables.

Other biographical information

Details about a person's life that help describe their background, experiences, and identity, not covered in the other pseudonymization rules.

Instruction: Pseudonymize sensitive information by generalizing personal life details.

Transcribed: I moved with my mother after her divorce.
Pseudonymized: I moved with my mother after {a change in her personal family situation: *sensitive data paraphrased*}.

Geographic locations

Regions, municipalities

Ties biographical information to a location. Areas can be linked with linguistic variants.

Instruction: Replace the region with *South/North/East/West/Center* of *Country*. Possibly combine two, such as *North-East*. No need to pseudonymize visits and holidays.

Example with replacement:

Transcribed: Each summer we went to my family in Dordogne.
Pseudonymized: Each summer we went to my family in {South of France: *sensitive data replaced*}.

Cities, region

Indicates past and current location of the interviewees. Can illustrate information regarding the socioeconomic background of the people.

Instruction: Replace specific city or region names with population-based descriptions or geographic generalizations. Replace with *type of city* in *South/North/East/West/Center* of *Country*. No need to pseudonymize visits and holidays.

Replacement micro edit table:

inhabitants	**type**
Less than 100,000	Small city
More than 100,000	Medium city
More than 1,000,000	Big city

* For the purposes of this project the city of Berlin is not pseudonymized as it is where the interviews occurred.

Example with replacement:

Transcribed: I live in Paris.
Pseudonymized: I live in {a large city in the West of France: *sensitive data replaced*}.

Area of interest for research project

This degree of locational detail is needed for socioeconomic precision.

Instruction: Generalize location in Berlin to one of the 12 districts.

Fig 2: Map of Berlin, accessed on 03.09.2024
Creative Commons CC0 1.0 Universal Public Domain Dedication
https://commons.wikimedia.org/wiki/File:Berlin_Subdivisions.svg

Example with replacement:

Transcribed: My office was in Kottbusser Damm.
Pseudonymized: My office was in {Neukölln: *sensitive data replaced*}.

Replace with:
Charlottenburg-Wilmersdorf
Friedrichshain-Kreuzberg
Lichtenberg
Marzahn-Hellersdorf
Mitte
Neukölln
Pankow
Reinickendorf
Spandau
Steglitz-Zehlendorf
Tempelhof-Schöneberg
Treptow-Köpenick

Practical tips

Pseudonymizing Interviews in the Humanities & Social Sciences

File formatting for the pseudonymization process

The following working method, where each sensitive part of the text becomes a row in a table, works particularly well for edits on a micro level:

Phase 1: Explore the data

Decide if you want to consolidate the corpus into one large document, or have multiple interview files. The first makes it easier to keep track of the most recent version, the other allows it to work in a more flexible manner.

For our project we found it handy to do a mix of both methods: split the corpus in two and each student assistant had one large document containing their half of transcripts.

See: Truan, Granger & Lychnara 2024a, Interviews Going Open! How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps). ⟨halshs-04743263⟩

Phase 2: Systematically annotate sensitive details

`File: original transcript`

00:03:43
Anamaria
I never liked school
When I was 5 years old
Everything was difficult

Highlight and add a label.

`File: annotated transcript`

00:03:43
Anamaria
I never liked school
when I was 5 years old[! age]
everything was difficult

Tip: see further subsections "Labeling sensitive information" and "Annotation marks"

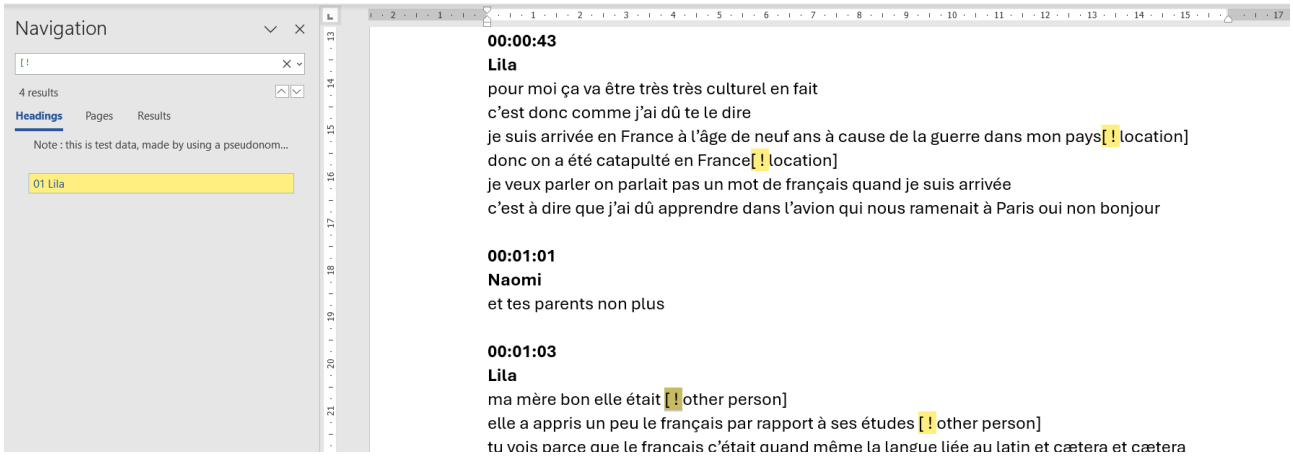

Figure 2: screenshot of the transcript with annotated sensitive information and type labels (here replaced with bogus data to protect the interviewee)

`File: sensitive data spreadsheet`

Speaker	ID	Timestamp	Original
Anamaria	5	00:03:43	when I was 5 years old[! age]

Tip: see further subsection "Annotation marks"

Use the *column split* function and *search and replace* all to remove the remaining brackets

Speaker	ID	Timestamp	Original	Type of data
Anamaria	5	00:03:43	when I was 5 years old	age]

Tip: see further subsection "Text transcription and tabular data"

	A	C	D	E	F	G	H	I
1	Interview	Time stamp	Sensitive Info, original text	Type of sensitive Info				
2	1	0:00:43	je suis arrivée en France à l'âge de neuf ans à cause de la guerre dans mon pay	année d'arrivée				
3	1	0:00:43	donc on a été catapulté en France	biographie				
4	1	0:00:43	ma mère bon elle était danceuse	famille				
5	1	0:00:43	elle a appris un peu le français par rapport à ses études de théâtre	famille				
6	1	0:00:43	donc on est arrivé en France quand moi j'avais neuf ans	année d'arrivée				
7	1	0:01:50	ce passage de ma langue maternelle qui était le serbo-croate	langue maternelle				
8	1	0:01:50	dehors on jouait dehors on parlait serbo-croate	langue maternelle				
9	1	0:01:50	heu le fait d'arriver à neuf ans à apprendre une autre langue	année d'arrivée				
10	1	0:01:50	mon frère a 5 ans de plus	famille				
11	1	0:03:34	c'est à dire que moi j'ai eu neuf ans pendant la guerre	biographie				
12	1	0:03:34	j'avais à peine huit ans quand j'ai arrêté d'apprendre ma langue	biographie				
13	1	0:03:34	donc en fait ils ont passé peut-être les deux premières années où j'étais à l'éco	biographie				
14	1	0:05:01	en France toutes mes études je les ai faites en France	études				
15	1	0:05:01	c'est à dire que je lisais pas j'allais pas à des cours spécial pour réviser le serbo-	langue maternelle				
16	1	0:05:01	et arrivée en Bosnie ce que je faisais quand j'allais en vacances	biographie				
17	1	0:05:01	donc j'le baragouine oui ça fait 20 ans	ans				
18	1	0:08:45	alors justement 1994 revenons 18 ans en arrière	année d'arrivée				
19	1	0:08:49	donc ça fait 18 ans que t'es ici non y a eu une interruption	année d'arrivée				
20	1	0:08:52	la première fois que j'ai découvert Berlin 1954	année d'arrivée				
21	1	0:08:52	je viens d'être acceptée dans la huitième meilleure prépa de France	études				
22	1	0:09:08	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	études				

Figure 3: screenshot of the spreadsheet with potential sensitive sentences (here modified to protect the interviewee)

During the process, we used the following columns:

- ID: The interview number was recorded in this column to track which interview the sensitive information originated from.
- Speaker: This column noted the pseudonymized name of the speaker (not displayed on Fig. 3 for privacy reasons).
- Timestamp: The time within the interview where the sensitive information was found was recorded here.
- Type of sensitive Information: This column indicated the category of sensitive information, which could vary according to the different corpora.
- Original text with sensitive information: The full utterance containing the sensitive information was documented in this box, rather than just the sensitive information itself. This approach facilitates the automatic replacement of text in the interview later on.
- Keywords: (see subsection "Flag words") This column was created to note various recurring words and expressions, allowing for a later search to ensure no sensitive information was missed. Due to time constraints, this task was not fully completed but is recommended for future projects.

Phase 3: Develop and implement the pseudonymization scheme

Split column, replace name with a pseudonym

`File: sensitive data spreadsheet`

Speaker	ID	Timestamp	Original	Type of data
Antoinette	5	00:03:43	when I was 5 years old	age

Duplicate text column and edit

Files: `sensitive data spreadsheet`
`pseudonymization spreadsheet`

Note: because the “years old” disappears with the replacement with “preschool aged”, we choose the label “paraphrased” instead of “replaced”.

Speaker	ID	Timestamp	Original	Pseudonymized	Type of data	Type of edit
Antoinette	5	00:03:43	when I was 5 years old	when I was {preschool aged	age	paraphrased

Copy paste the content of the text column, to go faster and minimize differences made by typing. Lock the column with the original texts, to prevent accidentally changing something.

Tip: see further subsection “Unicode disparities”

The new columns are:

- Pseudonymization: For this step, a coding scheme based on the categories under 'type of sensitive information' was agreed upon. The pseudonymization strategy—whether paraphrasing, replacing, or deleting information—was indicated. Curly brackets '{ }' were used to mark the changes, as this symbol was not used in the transcription.
- Type of edit: A separate column indicated whether the data was replaced, paraphrased, deleted, or not changed. Although this did not directly contribute to the pseudonymization process, it provided a means to calculate statistics for the corpus if required.

	A	B	C	D	E	F	G	H	I
1	Interview	Speaker	Time stamp	Type of sensitive i	Sensitive Info, original text	Pseudonymized	Type pseud.	Keywords	notes
2	1	Lila	0:00:43	année d'arrivée	je suis arrivée en France à l'âge de neuf ans à cause de la guerre dans mon pay	not changed	not changed		
3	1	Lila	0:00:43	biographie	donc on a été catapulté en France	not changed	not changed		
4	1	Lila	0:00:43	famille	ma mère bon elle était danceuse	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced		
5	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études de théâtre	elle a appris un peu le français par rapport à ses études {sensiti	deleted		
6	1	Lila	0:00:43	année d'arrivée	donc on est arrivé en France quand moi j'avais neuf ans	not changed	not changed		
7	1	Lila	0:01:50	langue maternelle	ce passage de ma langue maternelle qui était le serbo-croate	not changed	not changed		
8	1	Lila	0:01:50	langue maternelle	dehors on jouait dehors on parlait serbo-croate	not changed	not changed		
9	1	Lila	0:01:50	année d'arrivée	heu le fait d'arriver à neuf ans à apprendre une autre langue	not changed	not changed		
10	1	Lila	0:01:50	famille	mon frère a 5 ans de plus	mon grand frère {sensitive data deleted about other person}	deleted	frère	
11	1	Lila	0:03:34	biographie	c'est à dire que moi j'ai eu neuf ans pendant la guerre	not changed	not changed		
12	1	Lila	0:03:34	biographie	j'avais à peine huit ans quand j'ai arrêté d'apprendre ma langue	not changed	not changed		
13	1	Lila	0:03:34	biographie	donc en fait ils ont passé peut-être les deux premières années où j'étais à l'éco	not changed	not changed		
14	1	Lila	0:05:01	études	en France toutes mes études je les ai faites en France	not changed	not changed		
15	1	Lila	0:05:01	langue maternelle	c'est à dire que je lisais pas j'allais pas à des cours spécial pour réviser le serbo-	not changed	not changed		
16	1	Lila	0:05:01	biographie	et arrivée en Bosnie ce que je faisais quand j'allais en vacances	not changed	not changed		
17	1	Lila	0:05:01	ans	donc j'le baragouine oui ça fait 20 ans	not changed	not changed		
18	1	Lila	0:08:45	année d'arrivée	alors justement 1994 revenons 18 ans en arrière	alors justement Berlin {début années 2000: sensitive data repla	replaced		
19	1	Lila	0:08:49	année d'arrivée	donc ça fait 18 ans que t'es ici non y a eu une interruption	not changed	not changed		
20	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin 1954	la première fois que j'ai découvert Berlin {début années 2000: s	replaced		
21	1	Lila	0:08:52	études	je viens d'être acceptée dans la huitième meilleure prépa de France	je viens d'être acceptée dans {une bonne formation en France:	paraphrased		
22	1	Lila	0:09:08	études	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	not changed	not changed		

Figure 4: screenshot of the pseudonymization spreadsheet with extra column for the types of edits

Phase 4: Reformate transcriptions and make final files

Use script to generate an edited transcript

`Files:` `pseudonymization spreadsheet`
`automation scripts`

> 00:03:43
> Anamaria
> I never liked school
> when I was {preschool aged: sensitive data paraphrased}
> everything was difficult

Tip: see further subsection "Handy coding skills"

Correct errors and apply formatting

`Files:` `pseudonymization spreadsheet`
`error log`

> **00:03:43**
> **Antoinette**
> I never liked school
> when I was {preschool aged: sensitive data paraphrased}
> everything was difficult

Tip: see further subsection "Search and replace problems" and "Advanced search functions"

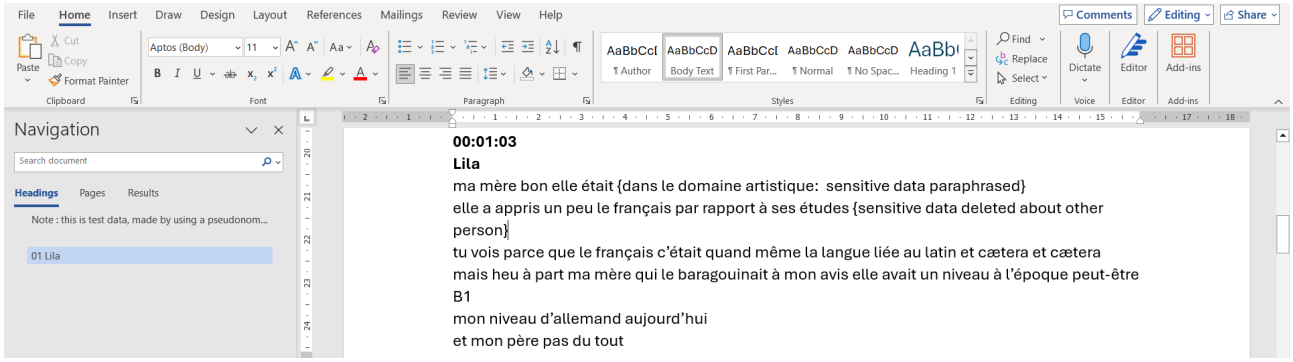

Figure 5: screenshot of the transcript with the pseudonymization edits

Make a copy of the spreadsheet and delete the original text column so that there is no sensitive information anymore, the original spreadsheet becomes a key file.

`File:` `pseudonymization spreadsheet`

Speaker	ID	Timestamp	Pseudonymized	Type of data	Type of edit
Antoinette	5	00:03:43	when I was {preschool aged	age	paraphrased

	A	B	C	D	E	F	G
1	Interview	Speaker	Time stamp	Type of sensitive I	Sensitive Info, original text	Pseudonymized	Type pseud.
2	1	Lila	0:00:43	famille	ma mère bon elle était danceuse	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced
3	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études de théâtre	elle a appris un peu le français par rapport à ses études {sensiti	deleted
4	1	Lila	0:01:50	famille	mon frère a 5 ans de plus	mon grand frère {sensitive data deleted about other person}	deleted
5	1	Lila	0:08:45	année d'arrivée	alors justement 1994 revenons 18 ans en arrière	alors justement Berlin {début années 2000: sensitive data repla	replaced
6	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin 1954	la première fois que j'ai découvert Berlin {début années 2000: s	replaced
7	1	Lila	0:08:52	études	je viens d'être acceptée dans la huitième meilleure prépa de France	je viens d'être acceptée dans {une bonne formation en France:	paraphrased
8	1	Lila	0:09:08	études	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	not changed	not changed

Figure 7: screenshot of the key file table

	A	B	C	D	E	F	G
1	Interview	Speaker	Time stamp	Type of sensitive I	Pseudonymized	Type pseud.	
2	1	Lila	0:00:43	famille	ma mère bon elle était {dans le domaine artistique: sensitive da	replaced	
3	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études {sensiti	deleted	
4	1	Lila	0:01:50	famille	mon grand frère {sensitive data deleted about other person}	deleted	
5	1	Lila	0:08:45	année d'arrivée	alors justement Berlin {début années 2000: sensitive data repla	replaced	
6	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin {début années 2000: s	replaced	
7	1	Lila	0:08:52	études	je viens d'être acceptée dans {une bonne formation en France:	paraphrased	
8	1	Lila	0:09:08	études	not changed	not changed	

Figure 8: screenshot of the pseudonymized metadata table

Optional: Convert into XML-TEI files (see Puren & Cafiero 2024)

Note: Due to time constraints, we did not convert the pseudonymized transcript into XML-TEI files.

```
<u who="#interv5">I never liked school
  When I was
  <gap reason="sensitive data paraphrased">
    <desc>preschool aged</desc>
  </gap>
  everything was difficult
</u>
```

Note that instead of directly using the pseudonym, an utterance is used to link to <profileDesc> <person xml:id=#interv5> <persName>Antoinette</persName> (pseudocode)

Besides the encoding of pseudonymization, other elements of the transcripts should be incorporated, such as the timestamps and paralinguistic information. Besides the transcripts, for this project we had an overview table with extra information, this metadata can be specified in the TEI header.

Editing schemes

Labeling sensitive information

The labels for the annotations were not a prescribed set, but made up on the moment to reflect what we encountered. In phase 2 each annotator created their own set of labels, giving two different and individual perspectives on the data. The flexibility of this approach made it possible to adapt our annotation process to the particular nature of

the content of our transcription.

These were the most common labels we assigned to our data:
Nationality Ethnicity Origin Language Work Duration Name Time Biographical Family Other person Age Birth place Year of move to Berlin School Main language Social media presence Vocation Education Metacommentary Place of residency Location

This also allowed us to better capture the degrees of sensitivity subjective (a piece of data could be interpreted as different types) and multilevel (general label location, versus specific label birthplace and label place of residence). In a nutshell let the data speak for itself, before putting it in rigid boxes. In phase 3 these served as the basis to determine the categories and instructions.

Naming the labels

Think about the exact meaning of the label names, but also that they sound differently, to limit confusion. This applies to all the types of categories and labels.

Too similar, prone to errors:
Replaced
Reformulated
Removed

Less confusing:
Replaced
Paraphrased
Deleted

Annotation marks

Avoid the use of formatting (text color, highlighting, background color, bold, italic, underline), as these get "lost" easily. Formatting can disappear when editing the text, exporting the file, when moving the content to another software. Ellis and Leek (2018) suggest using plain text instead.

Note: when working with docx documents, it is possible the query these formatting properties with the advanced functions ("Find and Replace" → "More >>" → "Format" → "Highlight" or "Font" → "bold" or "font color"). This function however is not available in most other text editors, and also not on the online versions of Word. It is also problematic when the document is exported to PDF.

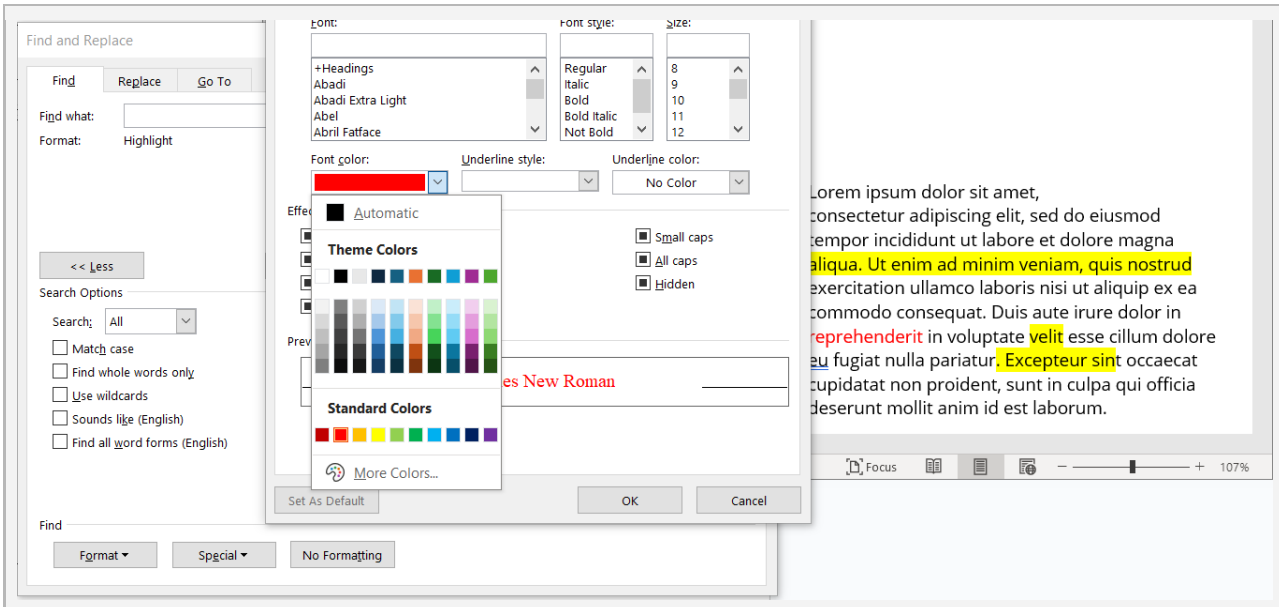

Figure 9: screenshot by authors of the advanced functions in Word

Special characters on international QWERTY keyboard:
` - = [ ] \ ; ‘, . / ~ ! @ # $ % ^ & * ( ) _ + { } | : “ < > ?

Instead of using colors, delimiters can be used to type inside the transcription, while making clear which parts are original or added. This differentiation remains clear in plain text.

Special characters

When choosing these delimiters, keep in mind the overlap with characters in programming languages that will be used later on in the process. Eventually the automatic changes can be constrained with escape characters.

Punctuation in words:
’ _

It is useful to consider a transcription system that doesn’t incorporate much punctuation, so that these symbols could be used for annotation purposes.

Punctuation in typography:
- ‘ ’ . , « » “ ” : ; % $
(also in social media content # @)

Special characters in Markdown:
* ** _ __ > >>> - ` ``` ~ = ! []() |

Special characters in HTML / XML / TEI:
< > / ! = "
(also with CSS : ; { })

Special characters in regex;
* + . [] () {} ^ $ - \ | ? ! < >

Special characters in Python strings;
" ' \

Using a combination of characters will prevent accidental matches.
Eg. Using - as a sequence --- will not be conflated with the - in ex-husband
Or using [? will not also match [...]

Automated tasks

Search and replace problems

When using the replace function for pseudonyms, use advanced options such as case matching and word boundaries to prevent replacing words that contain the same letter sequence as the name.

name: Anne
P<u>anne</u>aux
l'<u>année</u> (some software conflate e and é)

name: Claire
<u>Claire</u>ment, c'est
voix <u>claire</u> (homograph)

name: Marie
de <u>marie</u>r
le <u>marié</u> (some softwares conflate e and é)

Handy coding skills

Even if we do not want to use computational pseudonymization, because when dealing with qualitative data, human eyes are very important to deal with subtleties, some parts can be automated. In particular the formatting and menial but time-consuming tasks. For those, the project assistants could have following skills:

To automate repetitive actions (at least one, all of these can do the data manipulations):
Python, R, Javascript, command line.

Familiar with data formats (at least one, the others can be learned easily after):
HTML, XML, TEI, markdown

Understand the concepts of:
- Tidy data
- Different encoding systems (ASCII, Unicode/UTF)
- Version control
- Wildcards, regular expressions (Regex)

Unicode disparities

Some bugs were caused by differences in Unicode encoding. These changes probably happened when working in multiple software (Excel, Sheets, Word).

Not discernible to the naked eye, there were some non-breakable spaces (NBSP, U+00A0). While the texts seemed identical, they did not contain the same type of space U+0020.

Another problem was inconsistencies between apostrophes and single quotes, ' versus ’ (U+0027 versus U+2019)

Going from text to tabular formats, caused some linebreaks to appear or disappear, which caused some problems when merging the edited texts with the original transcriptions

Hybrid workflows for GUI and coding

As we want include the people that have expertise but not the digital skills, hybrid working methods (graphical user interface + programming) can be an option:

- Tabular data
    - Tech person can convert data to a CSV, which can be imported into Excel for Non-Tech person.
    - Non-Tech person can export Excel to CSV, which can be handled automatically by Tech person.
- Text data
    - Tech person can convert plaintext to a Docx, which can used by Non-

Tech person
    - Non-Tech person's Docx can be converted into plain text or markdown by Tech person, using a tool such as Pandoc.
    - Note on markdown: this system (for headings, italics, underline) is implemented in a lot of software, for example in Whatsapp and Reddit. Non-Technical persons might already be familiar with it. The markdown system was also invented as a bridge for non-Technical people so that they don't have to deal with more cumbersome markup tags.

Text transcription and tabular data

Moving between spreadsheet and transcripts, both formats have their advantages, the first to organize and have an overview, the second to have the full context.

- From text data to tabular data
    - Use a consistent annotation scheme (see section Annotation marks)
    - Extract annotations using pattern matching, Regex
    - Split the line into text and label using delimiters (simple GUI example below)

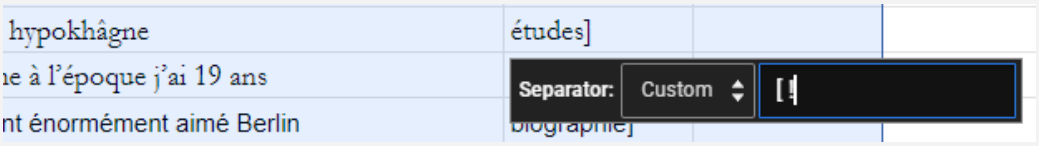

Figure 10: screenshot by authors of the split function in Sheets

- From tabular data to text data
    - Use a script to replace parts
    - The lack or addition of a line break often can cause problems
    - Pay attention to text formatting

Advanced search functions

Timestamps in documents:

```
Regex (?<!^|\d\d:\d\d:\d\d\n)Anamaria
Word ^#^#:^#^#:^#^#Anamaria
```

Dates (the content of our interviews concerned the 21th and 20th centuries):

```
Regex 20\d\d    19\d\d
Word 20[0-9][0-9]   19[0-9][0-9]
```

Age:

```
Regex \s\d\d\s
Word wildcards [!:]<[0-9][0-9]>[!:]
```

Word, add a new page between two interviews

```
Find what: ((fin de l'enregistrement))^p
Replace with: incompréhensible, ^&^m
```

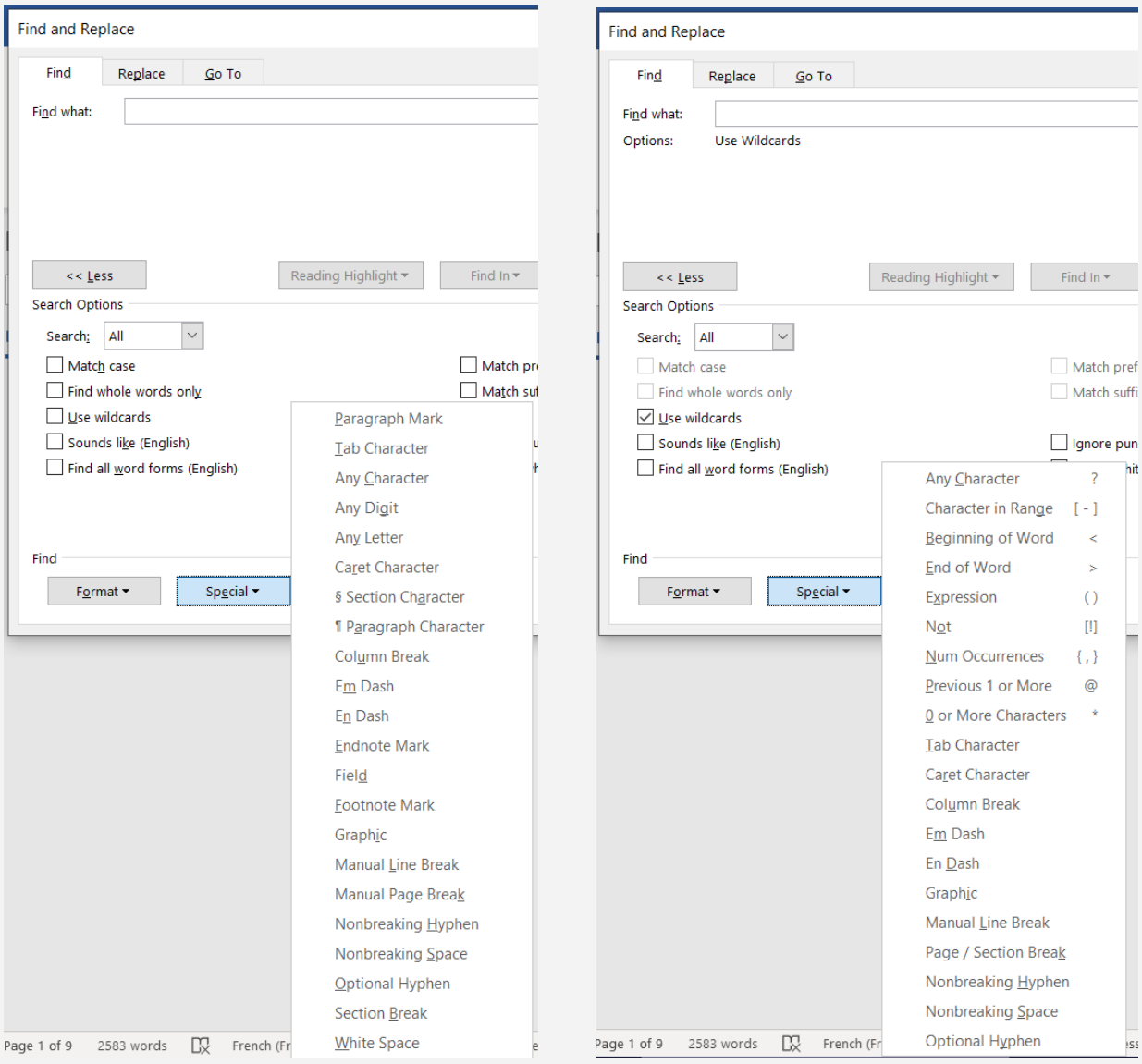

Figure 11: screenshot by authors of the advanced search and replace function in Word

Code snippet

Quickly written to make some parts faster, the code was not refined further. More details in the README.md file. Not intended to be used as is, but to serve as inspiration. For another project, the methods might differ greatly, a new script must be made to match the particularities of the initial documents.

What the script does:

0) prerequisites: an original transcript (docx) and a table with the changes to apply (a column with original text and another with pseudonymized text)

1) standardize differences in Unicode

2) it looks in the transcript for the table's original text and replaces it with the pseudonymized text.

```
# "error_log.txt" should be emptied

import pandas as pd

path = "\."

table_filename = f"{path}testtable.xlsx"
text_filename = f"{path}testtranscript.docx"
newtext_filename = f"{path}testresults.docx"

if __name__ == '__main__':

    with open("error_log.txt", "w") as tempfile:
        tempfile.write("") # make empty

    df = pd.read_csv(table_filename, encoding='utf8')
    df['Sensitive Info, original text'] = df['Sensitive Info,
                                            original text'].str.replace(u"\u00A0", " ")
    df['Pseudonymized'] = df['Pseudonymized'].str.replace(u"\u00A0", " ")

    df['Sensitive Info, original text'] = df['Sensitive Info, original text'].str.replace("'",
                                                                                          u"\u2019")
    df['Pseudonymized'] = df['Pseudonymized'].str.replace("'", u"\u2019")


    with open(text_filename, 'r', encoding='utf-8') as file:
        text = file.read()
        text = text.replace("'", u"\u2019")

        for index, row in df.iterrows():
            if row['Pseudonymized'] != 'not changed':

                if text.find(row['Sensitive Info, original text'].strip()) > 0:

                    text = text.replace(row['Sensitive Info, original text'].strip(),
                                "//CHANGED://"+str(row['Pseudonymized']))

                else:
                    with open("error_log.txt", "a", encoding='utf-8') as tempfile:
                        tempfile.write(
```

```
                        f"""{row['Interview']} {row['Speaker']} {row["Time stamp"]}\n"""
                        f"""{row['Sensitive Info, original text'].strip()}"""
                        f"""\n{row['Pseudonymized'.strip()]}\n\n"""
                        )

    with open(newtext_filename, 'w', encoding='utf-8') as newfile:
        newfile.write(text)
```

It can be improved in many ways, for example by implementing: better filepaths, better logging, fuzzy text matching, dealing with inconsistencies in line breaks (missing or new ones), and dealing with differences in markup (italics, bold, underline).

Flag words (in French)

When the same information occurs multiple times, it could be handy to deal with them using the search function. Some words, such as *lived* and *hometown*, also tend to co-occur with sensitive data.

```
frère, sœur, mère, père
m'appelle
\d+ ans
actuellement
job poste travail
je vis
Aujourd'hui
20\d\d   19\d\d
janvier février mars avril mai juin juillet août septembre octobre novembre décembre
Région
Profession
Ici
je suis
né[e]?
vécu déménagé
Venir venais \sven\w+\sde
nationalité
je suis
Grandi
ville vivais vivre
Origine
Vivre vivais \sviv\w+\s
Famille
Allé aller \sall\w+\s
Spécifique
Scolarisé
```

Famille
emploi employeur

References cited

Allen, Liz, Alison O'Connell & Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing* 32(1). 71–74.

DeLacey, Hannah. 2024. Pseudonymizing Data. Presented at the Centre for Digital Scholarship: Summer Training Week, Leiden University. https://www.digitalscholarshipleiden.nl/articles/cds-summer-training-week-2024. (2 December, 2024).

Ellis, Shannon E. & Jeffrey T. Leek. 2018. How to Share Data for Collaboration. *The American Statistician* 72(1). 53–57. https://doi.org/10.1080/00031305.2017.1375987.

Puren, Marie & Florian Cafiero. 2024. InTEIrviews: An ODD for Qualitative Interviews in the Humanities. *Journal of the Text Encoding Initiative* Issue 15. https://doi.org/10.4000/jtei.5007.

Truan, Naomi. 2024. Whose language counts? Native speakerism and monolingual bias in language ideological research: Challenges and directions for further research. *European Journal of Applied Linguistics* 12(1). 34–53. https://doi.org/10.1515/eujal-2024-0006.

Truan, Naomi. forthcoming. Becoming a Speaker of German as an L1 French Speaker: Elite Multilingualism as a Means of Distinction in a Globalized World. *International Journal of Multilingualism.*

Truan, Naomi, Sophie Granger & Jo Lychnara. 2024. Interviews Going Open! How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps). ⟨halshs-04743263⟩