FAITH

Fluency as faithfulness: A cognitive approach to an interpreting mega-corpus

DOI	10.6092/unibo/amsacta/7551
Licence statement	This work is licensed under Attribution 4.0 International. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Deliverable No.	1
Deliverable Full title	FAITH Data Management Plan Version 1.0
Lead beneficiary (extended name and acronym)	University of Bologna (UNIBO)
Authors (of the deliverable)	Nannan Liu
Planned delivery date	2024-02-22
Actual delivery date	2024-02-12
Dissemination level (PU = Public; PP = Restricted to other program participants; RE = Restricted to a group specified by the consortium; CO = Confidential, only for members of the consortium)	PU
Project website	NA
Project start date and duration	Start date of project: 2023-09-01 Duration: 24 months

Document History

Version	Date (DD/MM/YYYY)	Created/Amended by	Changes
1.0	14/01/2024	Nannan Liu	Initial commitment

Scheduled Data Management Plan (DMP) Updates

The DMP is a document that evolves during the project's lifespan and registers all relevant changes in the lifecycle of all research datasets. Updated versions of the DMP have been planned. Moreover, this document will be updated whenever important changes in the data or the data management policy occur.

Content

The Data Management Plan (DMP)

This DMP specifies all the research data collected and generated within the FAITH project. In particular, it explains how research data are handled, organised, licenced and made available to the public and how they will be preserved after the project is completed.

This DMP reflects the current state of the FAITH project. However, the details and the final number of datasets may vary during the project. The variations will be recorded in updated versions of this DMP.

1. Data Summary

The aim of FAITH is to unify existing interpreting corpora on a single-access platform and conduct three cognitive-linguistic studies on the unified corpus. The project will reuse pre-existing interpreting corpora (Work Package 1; WP1) and generate three datasets containing the analyses of: (1) asymmetrical structures between English and Mandarin Chinese (WP2A), (2) perception of political speeches (WP2B), and (3) explicitation features in interpreting (WP3).

The project will reuse datasets from the following sources: (1) the European Parliament Interpreting Corpora (EPIC) (Russo et al. 2005), (2) EPIC Ghent (Bernardini et al. 2018), (3) the Speech Corpus of Interpreted Premiers' Press Conferences (Liu 2023), and (4) interpreting corpora that will be provided by our prospective partners, such as the *Dolmetschen im Krankenhaus* ('Interpreting in Hospitals') corpus (Bührig & Meyer 2009) and the South African Sign Language Interpreting Corpus (Wehrmeyer 2019).

The reused data are textual and audiovisual in text, audio and video formats. They are restricted to non-commercial use for research and teaching purposes (see Bührig & Meyer 2009; Russo et al. 2005), and such licences are respected in WP1.

The project will produce the following types of data:

1. compressed data of the unified interpreting corpus by standardising the representation formats of existing corpora (WP1);
2. quantitative tabular data of extensive metadata from re-analysing and standardising the metadata formats of existing interpreting corpora (WP1);
3. quantitative tabular data from surveys and annotations made to textual, audio, and video samples of asymmetric structures and explicitation features (WP2–3);
4. qualitative data from focus groups about the perception of fluency (WP2B);
5. documentation and scripts of data analysis written for all four studies (WP1–3).

The research team agrees to convert data from proprietary to well-known, documented, and open formats to facilitate accessibility and reusability. A summary of the open formats is provided in Table 1.

Table 1 – Summary of data formats

Type of data	Formats used during data processing	Formats for sharing, reuse and preservation
Processed and compressed data from the re-analysis of existing interpreting corpora (textual, audio, and video data)	.gz, .txt, .xml, .mpg., .wav, .textgrid	.gz, .cmdi, .txt, .xml, .mpg., .wav, .textgrid
Quantitative tabular data of extensive metadata	.txt	.txt
Quantitative tabular data from surveys and annotations	.txt	.txt
Qualitative data from focus groups	.txt	.txt

Type of data	Formats used during data processing	Formats for sharing, reuse and preservation
Documentation and scripts of data analysis	.pdf, .r, .py	.pdf, .r, .py, .Rmd, .ipynb

The Python and R scripts will be preserved with thorough comments in original and markdown versions (in Jupyter Notebook [.ipynb] and RMarkdown [.Rmd] formats), version-controlled using *git*, a piece of open-source software to keep track of changes, and archived in AMS Acta and Zenodo with citations and licencing (see Section 2.1). This workflow makes it easy for interested readers to inspect and reuse the code, whether they have a technical background.

README files [1] explaining all relevant details regarding data collection, processing methodologies, and quality assurance will be deposited along with the datasets in .html, .md, or .txt formats.

The expected size of the data is 12 GB. Considering the early stage of the project, the effective size may vary from what is declared here. Potential variations will be addressed in further versions of this document.

The data produced can be of interest to different potential users. They may include interpreting researchers, educators, students, corpus linguists, computational linguists, language engineers, policy-makers, and the general public.

2. FAIR Data

2.1 Making data findable, including provisions for metadata

To ensure the findability of research data produced during FATIH, datasets will be deposited in the trusted institutional data repository the AMS Acta (https://amsacta.unibo.it/) and the general repository Zenodo (https://zenodo.org/) when appropriate. Whenever project results are published, we will deposit and describe the relative underlying datasets in AMS Acta and Zenodo to guarantee their discoverability, access and preservation beyond the end of FAITH.

The chosen repositories will attribute a unique persistent identifier (PID) to the deposited items. Both AMS Acta and Zenodo adopt the DOI system of PIDs. The identifiers are then used to cite the datasets in all research publications. More information about the repositories is provided in Table 2.

AMS Acta and Zenodo support standard descriptive metadata to ensure dataset indexing and discoverability. Both repositories require users to provide metadata about the funding information, description, licence, visibility, embargo, language, PIDs, creators, contributors, contact, related identifiers, etc. AMS Acta meets the requirements of Dublin Core, a widespread metadata standard in the library sciences. Zenodo complies with Dublin Core and DataCite Metadata Schema, as required by the OpenAIRE guidelines[2].

The repositories provide application programming interfaces (APIs) supporting the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a standard protocol for collecting metadata from repositories. Zenodo also supports REST API, a versatile way to collect the records.

[1] A 'README' file contains relevant information about dataset authorship, terms of reuse and responsibilities, explaining dataset content and structure, collection procedures and analysis, such as file specifics, methodologies, codebooks of variables, data sources, and further necessary notes. See Annex III for the suggested README file template.
[2] See https://guidelines.openaire.eu/en/latest/.

Specific keywords will be associated with each dataset to enhance semantic discoverability using controlled vocabularies registered in the Concept Registry of the Common Language Resources and Technology Infrastructure (CLARIN), a European research infrastructure in the domain of language resources. Example keywords include the registered concepts 'translation and interpreting, 'multimodal', 'multilingual, 'cognitive science', and those not yet registered, namely 'corpus standardisation', 'fluency', and 'explicitation'.

Research data are organised in datasets, i.e. named collections of data units with the same focus and scope. In this DMP, we set out the rules for dataset naming to improve data visibility, discoverability, citation, and permanent online tracking.

The recommended dataset title structure consists of the following:

> *FAITH. Task title or description. Additional information (if necessary). Version number.*

Example:

> *FAITH. Core metadata schema for interpreting corpora. Version 1.0.*

The version number of the dataset will be affixed to the end of the title in case of data revisions to help identify dataset updates, especially in repositories that do not automatically track versioning (see Annex I).

This DMP also recommends the following rules for file naming:

- for dataset file(s)

  > *FAITH_TaskNumber_Coverage or other content specifications_Date(YYYYMMDD)_VersionNumber.fileExtention*

  Example:

  > *FAITH_WP1_MetadataSchema_20240117_1.0.txt*

- for README file(s)

  > *FAITH_TaskNumber_Coverage or other content specifications_Date(YYYYMMDD)_VersionNumber_README.fileExtention*

  Example:

  > *FAITH_WP1_EPICmetadata_20240117_1.0_README.html*

2.2 Making data accessible

As a guiding principle, FAITH seeks to make all research data openly available as soon as possible and ensure open access via the repositories to allow the dissemination, validation, and reuse of research results.

To this purpose, all possible and legitimate actions and strategies are adopted to allow data sharing, including:

- converting the files to standard open formats;
- providing all relevant documentation and explanations for the data(sets);
- obtaining copyright permissions from third-party data owners to be allowed to reuse, reproduce and distribute the collected data;
- anonymising and aggregating the data;
- in case of the copyright on raw data derived, collected or elaborated from pre-existing databases or other sources (i.e. papers, journal articles, book chapters, reports, video and audio sources), collected data will be made available if the reproduction and sharing are

allowed by expressed permission of the right holders or by applicable copyright exceptions and exemptions. Otherwise, only aggregate data resulting from the analysis will be openly published. When the sources are freely available online in their original repositories, but direct reproduction is not allowed, a detailed account of how the dataset was created from the original data will be provided, together with the specification of open repositories from where the original datasets are available. Raw data in full texts will not be made available without the copyright holders' permission.

Restrictions to access are applied only in the following cases:

- collected data belong to third parties which have denied permission for sharing on account of confidentiality and proprietary issues;
- data anonymisation is not possible;
- data availability would jeopardise the project's main aim.

Annex I provides details on the accessibility of each dataset. In all cases, metadata will be made openly available and licenced under a 'No Rights Reserved' CC0 licence or equivalent per the Grant Agreement. The metadata will contain information on how to access the data.

The data repositories chosen guarantee long-term preservation and attribute PIDs to the archived datasets. They support open licences and different access levels and allow cross-linking between publications and the relevant datasets (see Table 2).

Table 2 – Summary of repositories.
The following table shows the repositories for dataset publication and preservation.

Repository name	Type	URL	PID	OpenAIRE compatibility?
AMS Acta	Institutional	https://amsacta.unibo.it/	DOI	Yes
Zenodo	General	https://zenodo.org/	DOI	Yes

2.3 Making data interoperable

All datasets will be described using standard descriptive metadata to ensure interoperability for indexing and discoverability. For each deposited dataset, relevant documentation explaining data collection procedures and analysis is made available along with the data to guarantee the intelligibility, reproducibility and validation of the project findings.

As mentioned, the team will convert all shareable data from proprietary to well-known and documented open formats (see Table 1). This allows data exchange and reuse among researchers, institutions, and countries.

Any publications based on the FAITH data will be mentioned in subsequent versions of the present DMP and contain a data availability statement.

To increase semantic interoperability beyond the FAITH project, we use the vocabularies of the CLARIN Concept Registry to describe the datasets created. We are creating a metadata schema for interpreting corpora, which maps to the CLARIN Concept and Component Registries, two commonly used ontologies in the field of language resources. The schema will be openly published for reuse, refinement, and extension.

The deposited documentation includes a full explanation and instructions for the data processing software used. Table 3 summarises the tools and software necessary to reuse our data.

Table 3 – Summary of tools and software for enabling the reuse of datasets

Tools/software	Open-access?
R	Yes
Python	Yes

Tools/software	Open-access?
Praat	Yes

2.4 Increasing data reuse

FAITH licences data of the unified corpus created in WP1 under the CC BY-SA licence while respecting the licencing of prospective corpus components, such as the non-commercial licences of EPIC (Russo et al. 2005) and EPIC-UdS at the University of Saarlandes (Przybyl et al. 2022). We will advocate using a permissive licence such as CC BY-SA to potential collaborators.

Nonetheless, we acknowledge that some corpus providers may be concerned about the commercial reuse and redistribution of interpreting data. Some language technology companies are alleged to infringe upon the intellectual property rights of human interpreters and falsely attribute information to data sources such as interpreters (cf. New York Times v. Microsoft and OpenAI 2023). The terms of service between interpreters and clients may also constrain permissive use. Therefore, we offer the option of restrictive licences to corpus providers.

The datasets created in WP2–3 will be licenced under the permissive Creative Commons Attribution International Public Licence (CC BY).

As per the Grant Agreement, metadata will be available under the Creative Commons 'No Rights Reserved' (CC0) licence or equivalent.

The data quality will be carefully assured by using the software listed in Table 3 to perform routine checks (e.g. for errors, duplicates, and inconsistencies) and backups of the datasets generated.

3. Allocation of resources

Making data FAIR requires monetary and time investments. Data processing (e.g. transcription and transfer) is expected to cost 5,000 euros and three months. Preparing data for sharing (e.g. anonymisation and cleaning) will take two weeks. Long-term preservation will be ensured without costs because the chosen repositories are free of charge. Data management (e.g. the preparation of this DMP) is expected to take one week throughout the project lifecycle.

Nannan Liu is responsible for data management, and her contact information is available in Table 4. Table 5 identifies all contributors participating in data management and specifies their roles.

Table 4 – Summary and contacts of people responsible for data management

Name	ORCID (if available)	Email address
Nannan Liu	https://orcid.org/0000-0003-2660-602X	nannan.liu@unibo.it

Table 5 – Summary of team members involved in dataset collection and management

Name	ORCID (if available)	Role
Nannan Liu	https://orcid.org/0000-0003-2660-602X	Data Collector
Mariachiara Russo	https://orcid.org/0000-0002-0904-2771	Data Collector
Marco Lobascio		Researcher

Annex I provides details about data management responsibilities related to each dataset.

4. Data security

During active data management (e.g. data collection and analysis), research data stored in computers, laptops, intranets or hard drives are accessible only after logging in with a username and password (periodically modified according to national law provisions for data security) and

are protected by updated antivirus software. They are regularly backed up to avoid accidental losses. None of the project data will be left inadvertently available. If external devices are used to store data files (e.g. backup files), they will be kept in a safe place accessible only to the researchers involved or will be encrypted with ad-hoc software.

The Microsoft OneDrive provided by UNIBO will be adopted for data sharing among team members. Regular backup of the data will be performed to ensure data recovery.

Long-term preservation of public data is ensured by the chosen data repositories with specific preservation policies.

5. Ethical or legal aspects

The Ethics Summary Report issued by the European Commission for FAITH states that the project was “ethics ready”. Data from focus groups and surveys conducted in WP2 will be collected anonymously; the study will not involve collecting and processing personal data.

References

Bernardini, S., Ferraresi, A., Russo, M., Collard, C. & Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In M. Russo, C. Bendazzoli & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies*. Singapore: Springer, 21–42.

Bührig, K. & Meyer, B. (2009, January 5). Dolmetschen im Krankenhaus (DiK). Dataset, https://www.fdr.uni-hamburg.de/record/8308.

Liu, N. (2023). Speaking in the first-person singular or plural: A multifactorial, speech corpus-based analysis of institutional interpreters. *Interpreting* 25 (2), 239–273.

New York Times v. Microsoft and OpenAI. No. 1:23-cv-11195 (United States District Court Southern District of New York 27 December 2023).

Przybyl, H., Karakanta, A., Menzel, K. & Teich, E. (2022). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In M. Kajzer-Wietrzny, S. Bernardini, A. Ferraresi & I. Ivaska (Eds.), *Mediated discourse at the European Parliament: Empirical investigations*. Berlin: Language Science Press, 191–218, https://doi.org/10.5281/zenodo.6977050.

Russo, M., Bendazzoli, C., Monti, C., Sandrelli, A., Baroni, M., Bernardini, S., Mack, G., Piccioni, L., Zanchetta, E., Ballardini, E., and others (2005, May 12). European Parliament Interpretation Corpus (EPIC). https://catalogue.elra.info/en-us/repository/browse/ELRA-S0323/.

Wehrmeyer, E. (2019). A corpus for signed language interpreting research. *Interpreting* 21 (1), 62–90.

Annex I: Datasets

Each expected dataset of the FAITH project is described in this Annex.

Dataset number	**Ready at month of project**	***Dataset title***
1	**7**	***FAITH. Mega-corpus. Version 1.0***.
Status		***not yet available***
ID [ID type]		A DOI will be provided when the dataset is available.
Version		1
Creator/s		Liu, Nannan (nannan.liu@unibo.it)
Contributor/s		Russo, Mariachiara (mariachiara.russo@unibo.it)
Researcher/s		Lobascio, Marco (marco.lobascio2@gmail.com)
Contact Person/s		Liu, Nannan (nannan.liu@unibo.it)
Contents		Merged corpus of EPIC, SCIPPC, and EPICG in a unified format
Data format		txt, xml, cmdi, mpg, wav, textgrid
Data volume		8 GB
Accessibility		open to academic users; this is because the EPIC dataset is licenced under Non-Commercial Use – ELRA END USER open only to academic users
Related publication/s		

Dataset number	Ready at month of project	*Dataset title*
2	**10**	***FAITH. Structural asymmetry. Version 1.0.***
Status		***not yet available***
ID [ID type]		A DOI will be provided when the dataset is available.
Version		1
Creator/s		Liu, Nannan (nannan.liu@unibo.it)
Contributor/s		Russo, Mariachiara (mariachiara.russo@unibo.it)
Contact Person/s		Liu, Nannan (nannan.liu@unibo.it)
Contents		Annotated datasets of asymmetrical structures in samples of Chinese–English interpreting and analytical workflows in R
Data format		txt, r, RMarkdown
Data volume		500 KB
Accessibility		openly available
Related publication/s		

Dataset number	Ready at month of project	*Dataset title*
3	**16**	***FAITH. Explicitation. Version 1.0.***
Status		***not yet available***
ID [ID type]		A DOI will be provided when the dataset is available.
Version		1
Creator/s		Liu, Nannan (nannan.liu@unibo.it)
Contributor/s		Russo, Mariachiara (mariachiara.russo@unibo.it)
Contact Person/s		Liu, Nannan (nannan.liu@unibo.it)
Contents		Annotated datasets of explicitation features in samples of Chinese/Italian–English interpreting and analytical workflows in R
Data format		txt, r, RMarkdown
Data volume		800 KB
Accessibility		openly available
Related publication/s		

Dataset number	Ready at month of project	*Dataset title*
4	**24**	***FAITH. Fluency perception. Version 1.0.***
Status		***not yet available***
ID [ID type]		A DOI will be provided when the dataset is available.
Version		1
Creator/s		Liu, Nannan (nannan.liu@unibo.it)
Contributor/s		Defranq, Bart (bart.defrancq@ugent.be)
Contact Person/s		Liu, Nannan (nannan.liu@unibo.it)
Contents		Annotated datasets of fluency features in samples of Chinese/French/Italian–English interpreting and survey results and analytical workflows in R
Data format		txt, r, RMarkdown
Data volume		700 KB
Accessibility		openly available
Related publication/s		

Annex II: Open Access status of project publications

In the following table, we will describe the open access status of the project publications and the underlying datasets. Because no publications have been produced, we leave this table unpopulated in the current version of the DMP.

Table 7 – Publications and related datasets.

Publications	
Bibliographic citation of the publication	
Link to copy archived in repository	
Related dataset/s	
Bibliographic citation of the publication	
Link to copy archived in repository	
Related dataset/s	
Bibliographic citation of the publication	
Link to copy archived in repository	
Related dataset/s	
Bibliographic citation of the publication	
Link to copy archived in repository	
Related dataset/s	

Annex III: "README" file template

This is the template of the README file that we will use.

README file

Dataset Title: "[insert title as defined in the DMP]"
Dataset Author/s: Name Surname (Affiliation), ORCID (if available);
Dataset Contributor/s: Name Surname (Affiliation), ORCID (if available);
Dataset Contact Person/s: Name Surname (Affiliation), ORCID (if available), email;
Dataset Licence: this dataset is distributed under a [insert LICENCE]
Publication Year: [insert YEAR]
Project Info: FAITH Fluency as faithfulness: A cognitive approach to an interpreting mega-corpus, funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101108651. https://cordis.europa.eu/project/id/101108651

Dataset Contents

The dataset consists of:
EXAMPLE 1

- 1 textual qualitative file saved in .txt format: "FAITH_WP2B_BeligiumFocusGroup_20251221_v01.txt"
- 1 README file: "README_FAITH_WP2B _BelgiumFocusGroup_20251221_v01.html"

EXAMPLE 2

- 1 tabular quantitative file saved in .csv format: "FAITH_WP2B_Survey_20260105.csv"
- 1 README file: "README_ FAITH_WP2B_Survey_20260105.html"]

Dataset Documentation

Abstract:
[Insert dataset abstract]

Content of the files:

- file [Insert filename] contains ...
- file [Insert filename] contains ...
- …

File specifics
[Please indicate instruction/technical info to allow potential users to correctly visualize and reuse your data (e.g. specific software, …). In the case of data converted into open formats, it could be useful to provide some further information. For example, if you deposit a .csv file derived from Excel for long-term preservation, you can describe the conversion. Here is an example description of conversion using Libre Office calc software:
To create the .csv files, Python 3.9 was used, with the following specifics:

- Character set UTF-8
- Field delimiter «, » (comma)
- Text delimiter «" » (quotes)]

Notes
[Related to the whole dataset or single files of a multi-file dataset (Optional)]

Data sources
[Optional]

Methodologies
[If necessary to understand how to reuse data]

Codebook of variables
[If necessary to understand the meaning of the variables]