Plan Overview

A Data Management Plan created using DMPonline

Title: Cohesion and Coherence: A corpus-driven analysis of graduating theses from Vietnamese Bachelor and Masterstudents

Creator: Van Ho

Principal Investigator: Thi Bao Van Ho

Data Manager: Thi Bao Van Ho

Project Administrator: Thi Bao Van Ho

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0002-8788-8932

Project abstract:

Cohesion and coherence are two fundamental dimensions of textuality that are strongly interconnected and thus are viewed as two complementary components of the internal text structure. Cohesion refers to grammatical-syntactical connections on the text surface, while coherence refers to logical-thematical relations between discourse objects in the deep structure of a text (cf. Fix et al. 2003: 16-32). This research uses the corpus-based, computer-assisted analysis method to study the different characteristics of these two textual dimensions in scientific writings of Vietnamese students. The aim of the study is to describe the use of selected cohesive phenomena and the structuring of arguments in graduating theses of Bachelor and Master students. The research data comes from the VieLKo learner's corpus (Vietnamesisches Lernerkorpus) and includes graduating theses from students studying German Linguistics at the University of Languages and International Studies - Vietnam National University (ULIS-VNU). These texts are in digital raw form and first have to be processed using a multi-level annotation scheme. Annotated data are then exported to multiple output formats and used in different statistical and qualitative analysis steps to provide the desired results.

ID: 74538

Start date: 01-10-2020

End date: 30-09-2024

Last modified: 02-06-2021

as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Cohesion and Coherence: A corpus-driven analysis of graduating theses from Vietnamese Bachelor and Masterstudents

Data Collection

What data will you collect or create?

This research involves the following groups of data:
1. Raw data:

- 10 Bachelor theses (about 15.000 tokens each, format: .doc/.docx/.pdf)
- 10 Master theses (about 25.000 tokens each, format: .doc/.docx/.pdf)

2. Base data:
- Annotation files

- Annotated texts with cohesion phenomena (tokenisation, lemmatisation, POS tagging, sentence span, paragraph, sentence complexity, sentence types) in EXMARaLDA (Dulko) (format: .exb)
- Annotated texts with coreference tagging in MMAX2 (format: .mmax)
- Annotated texts with rhetorical structure in RSTTool (format: .rs3)

- Output files

- Output for cohesion annotations in EXMARaLDA (Dulko): plain text files with annotation tiers for tokenisation, lemmatisation, POS tagging and sentence span (format: .txt); plain text files with annotation tiers for tokenisation, sentence span, paragraph, sentence complexity and sentence types (format: .txt); HTML files with all annotation tiers (format: .html)
- Output for coreference annotations: image (format: .jpeg/.png); XML document (format: .xml)
- Output for RST annotation: image (format: .jpeg/.png); XML document (format: .xml)

3. Analytic data

- Excel tables with quantified data to be used in R/JASP (format: .csv)
- compiled images of coreference chains and RST trees (format: .jpeg/.png)

All data currently take up less than 500 MB of storage. The final data pool would not exceed 1 TB.

How will the data be collected or created?

The raw data was externally requested from the *VieLko* learners' corpus. The base and analytic data are created within this research project. All data are stored in structured folders according to data groups and file types.

Examples for folder structure and file naming syntax:

C:|...|Project|Raw Data|BA|Word|2.2016.3.BA.S.1.09.doc

C:|...|Project|Working Copy|Phase II|00.ZS_exb|00.ZS_2.2013.3.MA.S.2.06.exb

Documentation and Metadata

What documentation and metadata will accompany the data?

Accompanying the data are the following documents and protocols:

- Metadata for the raw data (Excel file, procured from *VieLko*)
- Data overview (Excel)
- Guidelines for raw text manipulation
- Annotation templates for .exb files
- Annotation guidelines (Excel)
- Versioning logs (TortoiseSVN)

Ethics and Legal Compliance

How will you manage any ethical issues?

The raw data procured from the *VieLko* learners' corpus has been anonymised prior to this research.
The use of the metadata accompanying the raw data in this research has been approved by *VieLko* through a written agreement of confidentiality. These metadata are only embedded in the back-end annotation file(s) corresponding to each learner are will not be available in the public version of the annotations.
Any other processes/analyses in this research do not involve any kind of personal information.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Aside from the raw data from *VieLko*, all new data generated within this project will be copyrighted by the research author using Open Data Commons Attribution (ODC-By) 1.0 license.

Storage and Backup

How will the data be stored and backed up during the research?

The data body is stored in a local TortoiseSVN repository with versioning support and regularly backed up on OneDrive. A GitHub repository is also under development as a third data storage location.

How will you manage access and security?

Data access for collaborators and third-parties will be provided only through the GitHub repository.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

All data of the research is to be retained.

What is the long-term preservation plan for the dataset?

The dataset will be maintained on all three designated storage locations.

Data Sharing

How will you share the data?

Data sharing will be available through the GitHub repository.

Are any restrictions on data sharing required?

There are no restrictions on data sharing.

Responsibilities and Resources

Who will be responsible for data management?

The research author will be responsible for data management.

What resources will you require to deliver your plan?

There are no requirements for additional resources, since the local repository and OneDrive location are private property of the research author, and GitHub is free of charge.