Plan Overview

A Data Management Plan created using DMPonline

Title: Corpus of Dutch Verb Selection by Monolingual and Bilingual Football Coaches

Creator: Wieteke Windrich

Principal Investigator: Wieteke Windrich

Affiliation: Other

Template: DCC Template

ORCID iD: 0009-0004-8146-9835

Project abstract:

Investigates the verb selection by German and Dutch soccer coaches in their first languages and their second languages. Transcriptions will be taken from interviews pre- and post-football matches. The coaches' monolingual verb selections will be used to investigate monolingual present tense verb selections in spontaneous speech. This will then be compared with L2 verb selection to assess whether CLI of present tense constructions occurs.

ID: 214288

Last modified: 05-11-2026

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan 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

Corpus of Dutch Verb Selection by Monolingual and Bilingual Football Coaches

Data Collection

What data will you collect or create?

A dataset of present tense verb selections from publicly broadcast interviews and press conferences given by professional football coaches. There are four speaker groups: Dutch monolingual, German monolingual, Dutch coaches working in Germany, and German coaches working in the Netherlands. Each group has 5 to 10 speakers. The collection target is a minimum number of target verb tokens per group, estimated at approximately 20 hours of speech. Outputs are a token dataset in CSV format (one row per target verb token), speaker metadata, a codebook and frequency summaries. Original recordings will not be retained or redistributed.

How will the data be collected or created?

Interviews will be identified from publicly accessible sources (broadcaster archives, official club and league channels) and logged with source URL, broadcaster, date and timestamps. Relevant passages will be transcribed following a documented convention. If automatic speech recognition is used, every transcript will be manually corrected. Target verb tokens will be extracted and coded according to a written coding scheme, and a subset will be recoded to check consistency.

Documentation and Metadata

What documentation and metadata will accompany the data?

The dataset will be accompanied by a README describing the methods, transcription conventions and coding procedure, and a codebook defining every category and abbreviation in the CSV files. Speaker metadata will be limited to a speaker ID, L1, the language of the interview, and whether that language is the speaker’s L1 or L2. Metadata will follow the OLAC standard, which extends Dublin Core for language resources, and languages will be identified with ISO 639-3 codes. The README will state the version and license.

Ethics and Legal Compliance

How will you manage any ethical issues?

The project has no institutional ethics review, so I will apply the ALLEA European Code of Conduct for Research Integrity as the reference standard. Speakers are public figures speaking in a professional, public capacity. Data processing falls under the GDPR, with legitimate interests (Article 6(1)(f)) as the lawful basis and the research safeguards of Article 89. Each speaker will receive a privacy notice under Article 14 within one month of their data being obtained, sent through their club’s press office or a published professional contact address. The notice will set out the purpose of the research, the lawful basis, the categories of data, the source, the retention period, the speaker’s rights and the right to complain to the Data Protection Commission. Speakers may object to inclusion before a stated deadline, in which case their data will be removed. If no objection is received, processing will continue. The shared dataset is pseudonymised. It contains no names, verbatim sentences, source references or detailed biographical information, which limits the risk of reidentification.

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

The recordings belong to the broadcasters, clubs or leagues, and the speakers may hold copyright in their utterances. Copies will be made and analysed only under Ireland’s exception for computational analysis for noncommercial research, which does not require the rights holders’ permission. Any rights reservations for text and data mining published by broadcasters will be respected where they apply, and broadcasters will be informed of the research as a courtesy. No recordings, transcripts or verbatim excerpts will be shared. The shared dataset contains only coded variables and isolated verb forms, which do not reproduce protected expression. I own the annotations, codebook and metadata, and will release them under CC0.

Storage and Backup

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

Data will be stored on an encrypted laptop drive, with automatic backup to an encrypted EU-based cloud service and a monthly copy on an encrypted external drive kept at a separate location.

How will you manage access and security?

I will be the only person with access to the transcripts and the source log. The source log linking speaker IDs to interviews will be stored separately from the token dataset. Devices and backups will be encrypted and password protected.

Selection and Preservation

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

The token dataset, codebook, README, frequency summaries and analysis scripts will be preserved and shared. The source log will be retained privately to allow verification. Transcripts will be retained only as long as necessary to verify published findings and then deleted.

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

Shareable data will be deposited in Zenodo with a DOI, in open formats (CSV, plain text, PDF/A). Zenodo commits to retaining records for the lifetime of the repository. New versions will be added as the dataset grows.

Data Sharing

How will you share the data?

The token dataset, codebook, README and frequency summaries will be openly available on Zenodo under CC0. Other researchers can use them to test hypotheses about verb selection by L1 and L2 status.

Are any restrictions on data sharing required?

Yes. Recordings, transcripts and the source log will not be shared openly, for copyright and data protection reasons. On request, the source log may be shared with researchers for verification purposes only, under a written agreement. No data will be deposited until the objection deadline stated in the privacy notice has passed, and speakers who object before then will be excluded from all shared versions. Objections received after publication will be honoured in subsequent versions, but earlier versions on Zenodo will remain accessible. The privacy notice will state this.

Responsibilities and Resources

Who will be responsible for data management?

Wieteke Windrich as sole and primary investigator.

What resources will you require to deliver your plan?

Transcription and coding time, encrypted cloud storage (a small monthly cost), an external drive, and free tools such as Microsoft, R and Zenodo. The project is self-funded.