A mixed-methods perspective on blue collar workplaces: the multilingual and multi-varietal linguascape of the Port of Antwerp

A Data Management Plan created using DMPonline.be

Creators: Kim Schoofs https://orcid.org/0000-0003-4645-8511, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n.

Grant number / URL: 3H240018

ID: 204680

Start date: 01-01-2024

End date: 31-12-2028

Project abstract:

Despite the boom in studies on communication in the workplace, blue-collar workplaces –particularly in non-transient multilingual settings– have received little attention. Hence, we aim to explore the language practices of a complex blue-collar community of practice situated in the multilingual as well as multi-varietal Flemish linguascape – i.e. the Port of Antwerp. Two linguistic groups can be identified among this community of port workers, whose communication sometimes falters: native speakers of Dutch –who typically speak Antwerp dialect– and workers with an immigration background who may draw on a variety of languages, among which Colloquial Belgian Dutch. To study these groups' interactions, a corpus will be compiled of their various forms of communication, supplemented with ethnographic observations. By relying on a mixed-methods approach – using quantitative and qualitative approaches – with the interface of pragmatics and interactional sociolinguistics at its core, we aim to inventory the features of Antwerp dialect spoken by the port workers, explore how their language practices can be related to the construction of their community of practice and various identities, and map the multimodal strategies – e.g. gesture, gaze – they employ to cope with communicative challenges in this multilingual, multi-varietal workplace. Finally, we use a quasi-experimental design to explore how AI-mediated communication tools affect language practices in this blue-collar community.

Last modified: 06-06-2024

A mixed-methods perspective on blue collar workplaces: the multilingual and multi-varietal linguascape of the Port of Antwerp

FWO DMP (Flemish Standard DMP)

---

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
Radio communication	Radio communication between port workers (both before and during use of the AI app which will be implemented in the final phase of the project)	Generate new data	Digital	Observational	Audio data	<100GB	
Team meetings	Team meetings among port workers	Generate new data	Digital	Observational	Video & audio data	<100GB	
Demographic survey	Short questionnaire in which demographic information and a few details about workers' career trajectories are probed for.	Generate new data	Physical				+/- 30 surveys
Ethnographic interviews and observations	The researchers will follow some of the workers and will conduct ethnographic interviews with them to gain a thorough understanding of the workings of this complex workplace, while also keeping track of ethnographic observations (both before and during use of the AI app which will be implemented in the final phase of the project)	Generate new data	Digital + Physical	Observational	Video data	<100GB	4 weeks' worth of observational notes
Automatic loggings AI app communication	Automatic loggings of AI app communication: e.g. spoken/written/semi-automatic responses, as well as the metadata generated by the app.	Generate new data	Digital	Observational	Screenshots + metadata generated by the app (data format unknown)	<100GB	
Pretest/posttest survey	A pretest/posttest survey to gauge how workers rate their (ease of) communication with colleagues and their CofPs' cohesion before and after AI app use.	Generate new data	Physical				+/- 30 surveys

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

No existing data will be reused.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

The data collection entails ethical issues, as personal data will be collected for all datasets and these will be collected in non-pseudonymized form. If the data were to be leaked, participants could thus be identified which could have an impact on those affected.

Ethical approval was already obtained from SMEC (file G-2022-5332) for the datasets consisting of radio communication, team meetings, demographic surveys, and ethnographic interviews and observations. Hence, the approved procedure will be followed during the collection of these data sets, which entails the following:

Prior to data collection participants will receive information about which data will be collected and how the data will be processed and used during and after the research. During an information session participants will receive a cover letter and informed consent form (ICF) detailing this information, after which they will be asked to sign the ICF. Secondary participants, namely the interlocutors with whom the participant interacts, will receive the same letter and will also be asked to sign an ICF.
Following data collection, data will be encrypted to protect participants' identity, by using aliases for all proper names and place names. The file linking these aliases to participants will be kept separately in a secure folder and not together with the data itself. In the case of video data, footage will also never be distributed in its raw form: when an analysis of nonverbal language is necessary, anonymized pencil-drawings will be used. Fragments for analysis will always be chosen with attention to anonymity: fragments that would somehow allow for identification are excluded from analysis. As for the survey, personal data will also be pseudonymized.
For the datasets consisting of automatic loggings of the AI app communication and the related pretest/posttest survey, ethical approval will be obtained later on in the project. As the specific AI app that will be implemented will only be chosen at a later stage, concrete details with regards to data collection are currently unknown.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

The following will be collected from all participants:
- audio and/or video recordings of linguistic interactions in the workplace.

The following will be collected from all primary participants:
- video recordings and notes of ethnographic interviews and observations
- personal data: name, age, gender, nationality, information about workers' career trajectories
- personal data collected through implementation of the AI app (exactly which personal data this will be, will become clear once the relevant app has been chosen).
- In some cases: contact information: email address

As already explained above, all personal data will be encrypted following a procedure that was already approved by SMEC.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

As part of respectful research practice, our findings will be communicated back to the participants of the study and the companies where our data was collected. This will take the form of information sessions in which we will not only show the results, but will also present some advice for how interaction and communication may be improved.
Furthermore, we will disseminate our generalized findings through e.g. a press release, workshops for other companies, publications in professional and
popularising journals aimed at our target audience... However, commercial valorization is not the aim of our project. The project primarily intends to make observations that can contribute to the relevant academic field. While the project can also generate some recommendations that may be relevant for the business world, this is of secondary importance.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Data collection will be organized solely by researchers within the MIDI research group at KU Leuven.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

There are no legal issues to be managed regarding the data collection as it is currently envisioned. However, this issue will be revisited once the AI app which will be implemented during the final phase of the project has been chosen, as legal issues may play a role during that part of data collection.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

The corpus will consist of a main folder in which the data will be accompanied by a README.txt file that describes:

- General information: dataset title, DOI of dataset, investigators (names, roles, contact information, ORCID) & grant information.
- Data collection: collection date/period, location of collection, methods and equipment used for data collection & problems and limitations.
- Organization: folder structure, file naming system, relationships and dependencies between files and data(sub)sets, a short description of files' contents and date of creation & other documentation files of interest within the dataset.
- Processing: methods used for data processing, software used for data processing (including version numbers), file formats used in the dataset & dataset changelog.
- Codebook: definition of codes and abbreviations & transcription conventions used for different datasets.

The main folder will contain subfolders for each dataset (as described in '1. Research Data Summary'). Additionally, the main folder will also include an Excel file containing metadata about the whole corpus, whereby each dataset will be represented by a sheet in the Excel file. A sheet in this Excel file will thus provide information about each of the individual files (e.g. file name, format, date, location, duration, participant information, interviewer information, …) in a specific subfolder. Each of the subfolders will contain master versions of the specific dataset's data files (i.e. video data (.mp4), audio data (.mp3), survey data (.csv), ethnographic notes (.docx) or AI app loggings), digitalized IC-forms as well as a Word file containing the encrypted transcriptions of the dataset if relevant.

For safety reasons, a key linking aliases to proper and place names will be kept separately in a secure folder and not together with the data itself.

In addition to the main folder and subfolders described above, which will store project-level documentation together with master versions of the data files, the encrypted data will be imported into NVivo for further processing. In particular, the data will be coded using a data-driven approach. Linguistic phenomena and themes will be annotated as they are observed, which entails that the codebook will be documented and continuously updated throughout the project to guarantee consistent data processing. Given the features of the NVivo software, it will be possible to let patterns emerge in the data through coding and thus create relationships between data sources. Hence, important data-level descriptive documentation will be added to the dataset throughout the processing in NVivo.

Finally, the encrypted transcriptions of video and audio data that will be imported into NVivo, will be transcriptions that reflect the original data as literally as possible without the use of conventions. Once imported in NVivo, these transcriptions will be transcribed in more detail using the Jeffersonian conversation analytic transcription conventions, complemented by the multimodal Mondada conventions and CHAT-conventions.

All documentation compiled during the NVivo project cycle will be exported in the form of an RTF-file for archiving at the end of the project.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

As a metadata standard is not commonly employed in the specific field of study, this project will also not use an existing standard. Furthermore, as the data will only be used within the research group, and since it will not be possible to share the data for (re)use both during and after the project, the need for a metadata standard is somewhat less essential.

Nevertheless, metadata will be created to make the data easier to manage and search. As described above, an Excel file containing metadata about the whole corpus will be put together. In addition, the data-driven coding that will be performed during processing in NVivo, will allow for dynamic and (semi-)automatic explorations of the data.

3. Data storage & back-up during the research project

Where will the data be stored?

The storage solution stipulated in the already approved PRET application (file G-2022-5332) will be used for the datasets consisting of radio communication, team meetings, demographic surveys, and ethnographic interviews and observations. Hence, during the project, the data will be stored on OneDrive linked to a KU Leuven account and on an external hard drive protected by BitLocker.

Very little paper data will be collected. However, ICFs and surveys that are signed and filled out on paper will be stored by the project's PhD researcher in a locked drawer or cupboard that is only accessible to him/herself. During the project all paper data will be digitalized and stored in the same way as the data described above and the paper versions will be destroyed in a responsible way.

The datasets consisting of automatic loggings of the AI app communication and the related pretest/posttest survey will likely be stored in the same way. However, as the specific AI app will only be chosen at a later stage, a final decision about this will also be made later on.

How will the data be backed up?

As the digital data will be stored on both OneDrive linked to a KU Leuven account and an external hard drive protected by BitLocker, the latter will provide a backup of the data.

Physical data will be digitized as soon as they are collected and stored in the same way as digital files in order to prevent loss.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

As OneDrive linked to a KU Leuven account offers up to 2TB in storage capacity, this far exceeds the amount of data that is expected to be collected throughout the project. As described in '1. Research Data Summary', the corpus will remain well under 400GB. Additionally, two external hard drives with sufficient storage capacity will be purchased at the start of the project. One will be used as back up during the project, and one for extra storage back up after the project.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The multifactor authentication required to access files on OneDrive linked to a KU Leuven account, forms a first important measure to avoid access by unauthorized persons.

Additionally, access rights to the datasets will be strictly managed. Data will in principle only be accessible to the supervisor and PhD researcher. These access rights will regularly be checked throughout the project.

Furthermore, only the supervisor and PhD researcher will have the password needed to access the external hard drive protected by BitLocker.

The data will not be made accessible to third parties, with the exception of students working on it as job students or as part of their BA or MA thesis. When a student uses the data collected by the researchers, the student first signs a confidentiality clause before gaining access to the data. When students work with the data, they work with the encrypted transcriptions as often as possible. Only when they collect data themselves (see below) or when it is completely impossible to answer the research questions with transcriptions do students come into contact with the raw data.

When a student collects their own data as part of their BA or MA thesis, the student signs the ICF and a confidentiality clause. Specifically, it is stipulated that the student is not to use the data except for his/her research and to follow the procedure for anonymization approved by SMEC. The students commit to storing the data only in a folder on OneDrive secured with KU Leuven authenticator.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

As OneDrive linked to a KU Leuven account is a free service for KU Leuven staff and students, this will not incur costs during the project.

In order to purchase external hard drives, funding has been acquired through the project application.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data collected throughout the project will be retained permanently for the purpose of potential historical linguistic research on language change processes.

Where will these data be archived (stored and curated for the long-term)?

The supervisor will keep the data on two external hard drives protected by BitLocker, following the project's conclusion. As the data is not to be shared with or reused by researchers outside the research group, given its personal and confidential nature, the benefits of data preservation in a data repository do not directly apply for this dataset. However, the dataset will be registered in KU Leuven RDR, which will enhance the citability of the dataset. In RDR the dataset's location and availability within the department will be described.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Data storage following the project's conclusion will not incur costs as the external hard drives will have been purchased with funding acquired through the project.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- No (closed access)

As all datasets contain personal data, the data will be kept under closed access both during and after the project due to privacy concerns for participants.
While pseudonymized transcripts of the audio and video data could be shared, these would not allow for replication of the research, since multimodal linguistic analyses cannot be made without the availability of the raw data. Similarly, sharing pseudonymized versions of the survey data and AI app loggings would not allow for replication of the research, as these datasets are intended to be analyzed in comparison to the audio and/or video datasets.
Given this, and in the best interest of the participants' privacy, the data will be held under closed access. This is also in line with the application that was already approved by SMEC for the datasets consisting of radio communication, team meetings, demographic surveys, and ethnographic interviews and observations. In the ICF participants will be reassured that confidentiality of data will be ensured at every stage of the research.

If access is restricted, please specify who will be able to access the data and under what conditions.

The data will in principle only be accessible to the supervisor and PhD researcher. As already described in '3. Data storage & back-up during the research project', BA or MA students may also come into contact with the data or even collect data. In the former case students will sign a confidentiality clause and in the latter case they will sign both a confidentiality clause and the ICF. When students work with the data, they work with the encrypted transcriptions as often as possible. Only when they collect data themselves or when it is completely impossible to answer the research questions with transcriptions do students come into contact with the raw data. Students will be required to follow the procedure for anonymization approved by SMEC and to store the data on OneDrive secured with KU Leuven authenticator.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects

As all the datasets contain personal data, and since only a non-anonymized or non-pseudonymized version of the data would allow for full replicability of the research, the data will be kept under closed access both during and after the project due to privacy concerns for participants

(see above for further explanation).

Where will the data be made available? If already known, please provide a repository per dataset or data type.

As already described in '4. Data preservation after the end of the research project', the dataset will be registered in KU Leuven RDR, where the dataset's location and availability within the department will be described. The supervisor will preserve the actual data on two external hard drives protected by BitLocker, following the project's conclusion.

When will the data be made available?

The data will not be made available, as it will be held under closed access (see above for further explanation).

Which data usage licenses are you going to provide? If none, please explain why.

Instead of a license, the 'Custom KU Leuven' option will be chosen, as the data will be held under closed access.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

As already described in '4. Data preservation after the end of the research project', the dataset will be registered in KU Leuven RDR, where the dataset's location and availability within the department will be described. Hence, the data will receive a DOI.

What are the expected costs for data sharing? How will these costs be covered?

Since the data are not intended to be shared, this will not incur any costs.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The PhD researcher, under supervision of Dorien Van De Mieroop

Who will manage data storage and backup during the research project?

The PhD researcher, under supervision of Dorien Van De Mieroop

Who will manage data preservation and sharing?

Dorien Van De Mieroop

Who will update and implement this DMP?

The PhD researcher, under supervision of Dorien Van De Mieroop