CIRCE
counteracting accent discrimination practices in education

DATA MANAGEMENT PLAN

Silvia Calamai, Laurie Jane Anderson, Giorgio Carella, Letizia Cirillo, Rosalba Nodari, Duccio Piccardi, Claudia Soria, Paola Baroni, Riccardo Del Gratta, Valeria Quochi, Philipp Meer, Frauke Matz, Ricardo Römhild, Robert Fuchs, Luís Guerra, Jean Antunes, Lili Cavalheiro, Laura Melgão, Ricardo Pereira, Amna Brdarević-Čeljo, Vildana Dubravac

Co-funded by the European Union

Funded by the European Union under agreement n° 2022-1-IT02-KA220-SCH-000087602. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Datasets

Title: CIRCE

Template: Horizon 2020

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- To obtain information
- To develop a product

Data collection/generation is carried out mainly in work packages 2 (WP2 Raising awareness on accent discrimination) and 4 (WP4 Engaging students in self-reflective practices).
Specifically, the activities (As) of the two WPs involving data collection/generation are the following:

WP2

- A4> Collecting & analysing accent attitudes and beliefs I: Accents (regional & transnational) of the national language;
- A5> Collecting & analysing accent attitudes and beliefs II: Accents of the national language determined by immigration language backgrounds;
- A6> Collecting & analysing accent attitudes and beliefs III: Different accents of English as an L2.

WP4

- A1> Journalism: Creation of educational (scripted, non-fictional) podcasts on accent discrimination;
- A2> Linguistic autobiographies: collect personal experiences about linguistic discrimination promoting intercultural exchanges;
- A3> Contest: call for an online video contest (videos recounting real experiences of accent discrimination) and a writing contest (short stories/poems promoting ways of inclusive behaviours and actions to combat discrimination).

WP2 activities: Collecting & analysing accent attitudes and beliefs

Data on students' accent attitudes and beliefs will be collected via verbal guise test (VGT), involving the presentation of recordings of different speakers to listeners. The recordings represent specific accents and participants are asked to rate speakers on a range of attributes. Qualitative data will be collected using metalinguistic interviews with selected participants. The elicitation and analysis of students' language attitudes via VGT will yield evidence of accent prejudice and discrimination in education in Europe, with regard to both L1 regional and non-native accents and L2 English accents.

WP4 activities: Podcasts, autobiographies and contests

In this WP, students will be asked to recount their own experience and to survey peers' experiences of discrimination through the collection of interviews and linguistic autobiographies and the creation of

 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

podcasts and short videos, as well as to propose ways to eliminate discrimination through a short story writing contest. Multilingual podcasts on accent discrimination will be created by the partner university students on the topics of accent discrimination, linguistic insecurity, inferiority, and psychological trauma caused by such behaviours.

1.1.2 What are the types of the described generated/collected data?

- observational (e.g. sensor data, data from surveys)
- experimental (e.g. gene sequencing data)
- Other

The project activities will yield different types and formats of data, such as Audio/Video recordings (.wav /.mp4), autobiographies and short stories/poems (.txt/.doc/.pdf) as well as surveys and questionnaires (.txt/.doc/.pdf), as shown in Table 1:

Table 1: *Type and format of data generated/collected related to the relevant activities.*

Activity	Type	Format	Aim
2.4/2.5/2.6	VGT audio recordings	.wav	Present VGT participants with different accents.
	Socio-demographic surveys	.txt/.doc/.pdf	Collect data about VGT participants' background.
	VGT questionnaires	.txt/.doc/.pdf	Register VGT participants' judgments on the different accents.
4.1	Podcast Audio/Video recordings	.wav/.mp3/.mp4	Create podcast to spread awareness on accent discrimination.
4.2	Linguistic Autobiographies	.txt/.doc/.pdf	Collect students' personal experience on accent discrimination.
4.3	(Contest) Video recordings	.mp4	Collect real experiences of accent discrimination.
	(Contest) Short stories/poems	.txt/.doc/.pdf	Promote thinking, behaving, and acting in combating accent discrimination.

1.1.3 What are the formats of the described generated/collected data?

- Text files
- Multimedia

.txt - .doc - .pdf - .wav - .mp3 - .mp4

1.1.4 What is the origin of the described data?

Primary data

1.1.5 What is the expected size of the described data?

MB (megabyte)

 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

The relative majority of produced data (in terms of size) will consist of Video and Audio recordings, but scanned surveys and questionnaires will also represent a relevant portion of data.

Nevertheless, some issues remain to be decided, such as the balance between the quality of the recordings and the size of video/audio files, as well as the specific way in which surveys and questionnaires will be submitted and stored.

Considering an average quality/size ratio for audio/video recording and paper-based surveys/questionnaires subsequently scanned and stored in PDF format, the total amount of data produced in the project is estimated to exceed 22 GB, as shown in Table 2:[1]

Table 2: *Estimated data size.*

Activity	Type	Length	Size	Amount	Total size
2.4/2.5/2.6	VGT audio recording	30-40 seconds	5-7 MB	45	225-315 MB
	Socio-demographic survey	1-2 pages	15-30 MB	1000	15-30 GB
	VGT questionnaire	19 pages	285 MB	1000	285 GB
4.1	Podcast Audio/Video recording	10 minutes	Audio 110 MB Video 470 MB	25	Total size will depend on audio/video ratio.
4.2	Linguistic Autobiography	1-2 pages	?	200	Total file size will depend on submission method.
4.3	(Contest) Video recording	5 minutes	235 MB	50	11.75 GB
	(Contest) Short story/poem	500 words	25 kB	25	625 kB
Total					> 300 GB

1.1.6 To whom might it be useful ('data utility')?

- Researchers
- Research communities
- Decision makers
- Education
- The public

[1] Audio recordings size is calculated assuming .wav format (sampling rate 44,1 kHz; bit rate1411.2 kbps); the use of .mp3 format would considerably decrease total size but also audio quality. Video recording size is calculated assuming the common .mp4 (H.264/AVC) format. Surveys and questionnaires size is calculated assuming a 400 dpi, grayscale scan. All estimates refer to uncompressed files.

 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

Data collected/generated within the CIRCE project will be useful for the following stakeholders: research community, policy makers, teachers, students, and the general public.

Research community and scientific activity

CIRCE aims to generate comparative, scientific and systematic knowledge about language discrimination in the educational context and its conscious and unconscious effects on performance evaluation. Thus, these data will allow further research developments in the following fields, among others:

- sociolinguistics
- applied linguistics
- educational linguistics
- social psychology
- perceptual dialectology
- language teaching
- linguistic discrimination
- language rights

Policy makers

The project goal is to raise awareness of issues linked to linguistic discrimination, a phenomenon which is not yet at the centre of public debate. This increased awareness will encourage national and international lawmakers to address the issue of linguicism/ language discrimination (with particular attention to accentism), implementing specific legal devices against language-based discrimination, along the lines of the bill approved by the French National Assembly in November 2020.

Teachers

In contemporary European societies, teachers are confronted daily with regional and non-native accents of the national language and are at risk of unconsciously succumbing to prejudice and negative evaluations of non-standard varieties. Hence, CIRCE aims to provide teachers with the skills and competencies needed to avoid conscious and unconscious linguistic discrimination which can lead students from marginalized (linguistic) backgrounds to be judged more negatively in terms of academic achievement. It will also support teachers and educators by providing adequate materials that take into account linguistic discrimination in the classroom, so as to prepare students to live in increasingly multicultural settings.

Students and the general public

The school environment is a hotspot for addressing accentism: students are exposed to different accents, form attitudes towards them and reinforce their attitudes on the basis of peer pressure. However, in all contexts non-standard speakers, or speakers with ethnic backgrounds, may experience different forms of linguistic microaggressions and accent discrimination, such that they may be perceived as incomprehensible, mocked and shamed for how they speak, or become the object of normative comments that address their pronunciation.

Hence, the increased awareness spread by CIRCE about the topic of accent discrimination will help students, and more generally citizens, develop a greater tolerance towards accent variation, unhinging the mechanisms of this powerful discriminatory mechanism and piercing the social tolerance that surrounds it by promoting knowledge of how it works.

Finally, CIRCE aims to make people understand that language proficiency is not compromised by a non-standard accent (even if strong) and that the two concepts should be kept separate.

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

2.1 Reused Data

2.1.1 Are you re-using the described data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

In order to allow discovery, data produced will be provided with rich metadata using the Component Metadata Infrastructure (CMDI) developed within CLARIN.

One of the key properties of CMDI is its flexibility, which makes it possible to create metadata records closely tailored to the requirements of resources and tools/services.

Specifically, description building blocks ('components', which include field definitions) can be grouped into a ready-made description format (a 'profile'). All types of data generated/collected within CIRCE will be associated to an appropriate profile, if present, otherwise new specific and CMDI compliant profiles will be created ad hoc. Metadata will also include search keywords to optimize the possibility for discovery potential re-use.

3.1.1.2 Please provide URL/Location describing the used metadata schema

https://www.clarin.eu/content/component-metadata

ILC4CLARIN

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

Metadata will be represented in accordance with CLARIN standard and methodology using standard vocabularies common to researchers and institutions in the fields of sociolinguistics (e.g., ISO 639-2 for names of languages). Hence, no project specific ontologies or vocabularies will be used.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Metadata will be available free-of-charge.

3.1.1.6 Will your metadata be harvestable?

Yes

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

CIRCE's datasets will be hosted in a free-to-use repository based on the OAIPMH standard harvesting protocol.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

The CIRCE Project Reference Manual (§ 3.4) contains precise instructions regarding file naming and storage. Specifically, in order to facilitate the manual storage and retrieval of data, consistent naming conventions will be implemented, such as: · avoid excessively lengthy folder names and intricate hierarchical structures; · use information-rich filenames containing diverse elements (e.g., titles, dates, version numbers); · separate elements within filenames using visually clear characters (e.g. underscores, hyphens, etc.); · capitalize the first letter of each element in filenames for improved readability; · sequence time elements as year, month, and day for consistent ordering.

3.1.1.9 Will you provide clear version numbers for your data?

Yes

As for versioning, the following conventions will be implemented:

- integrate a version control element at the end of filenames (e.g., "v" or "rev" followed by version and file identifiers) for files intended for sharing;
- distinguish version identifiers between drafts and final releases (e.g., zero vs. non-zero digits);
- maintain coherent progression and judicious incrementing of version identifiers;
- avoid terms like "new version" or "final version" to prevent confusion, as final drafts can still be updated.

3.1.1.10 Will you provide persistent identifiers for the described data?

Yes

All resources produced within CIRCE will be assigned a persistent identifier via the Handle System (CNRI).

3.1.1.11 Persistent identifiers

Handle System

3.1.1.12 Will you provide searchable metadata for the described data?

Yes

CLARIN metadata are indexable and searchable by search engines based on OAI-PMH, such as the CLARIN Virtual Language Observatory, which harvests all CLARIN related repositories and makes their content identifiable by web crawlers.

3.1.1.13 What services will you use to provide searchable metadata?

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

Registry/Catalogue

3.1.1.14 Please provide URL/Name for the used searchable metadata

https://www.clarin.eu/content/component-metadata

ILC4CLARIN

3.1.1.15 Will you use standardised formats for the described data?

Yes

Raw text data will be stored in .txt files. Annotated data will be represented and stored in XML format following as far as possible standard models such as TEI.

The exact definition of the formats and models are part of the activities of the project and will thus be updated at a later stage.

3.1.1.18 Are the file formats you will use open?

Yes

In order to facilitate data interoperability and sharing, all released data (whenever possible) will conform to the most common non-proprietary, open formats. The exact definition of the formats and models are part of the activities of the project and will thus be updated at a later stage.

3.1.1.20 Do supported open-source tools exist for accessing the data? Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the described data?

Yes

As for personal data, CIRCE operates in accordance with articles 13 and 14 of EU's 2016/679 regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (also known as GDPR: General Data Protection Regulation). In compliance with GDPR, some datasets will not be made (entirely) accessible. Considering all the above, all data generated/collected within CIRCE which can be made public according to the GDPR will be made accessible through either the project website or the ILC4CLARIN repository.

3.1.2.2 Will the described data be openly accessible?

Yes

The final release of the data will be openly accessible. Intermediate releases will be covered by licenses that are being finalized

3.1.2.3 How will the data be made available?

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

Repository of Archive

CIRCE's datasets will be made available to the scientific community through ILC4CLARIN, the data centre hosted by CNR-ILC as a node of the CLARIN-IT consortium, the Italian component of the European CLARIN federation. ILC4CLARIN is a free-to-use repository based on the OAI-PMH standard harvesting protocol. In order to gain access to the project's datasets, a registration to the CLARIN website will be required in order to verify the identity of the person accessing the data.

The presence in the consortium of the repository host has allowed the establishment of specific and appropriate arrangements for data storage and accessibility, as well as the possibility to easily update them as needed.

3.1.2.4 Please provide URL/Name of used data repositories

https://ilc4clarin.ilc.cnr.it/en/

ILC4CLARIN

3.1.2.5 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

All data stored within the ILC4CLARIN archive will be protected according to the Code of Conduct for Service Providers, a common protection standard for the research and higher education sector. Moreover, ILC4CLARIN's storage plan offers the following features against data loss:

- all data in the repository will be kept in at least three copies, one of which off-site, at all times;
- all the copies will be checked regularly and replaced, should any of them become corrupted.

3.1.2.6 Are there any methods or tools required to access the described data?

No

3.1.2.9 Will you also make auxiliary data that may be of interest to researchers available?

no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for the described data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make the described data available for reuse?

after article publication

In some cases, the publication of data could be delayed in order to give time to publish research articles based on them. Nevertheless, CIRCE's members are aware of the importance of making research data available as soon as possible, and thus this practice will be avoided as much as possible.

3.1.4.4 What internationally recognised licence(s) will you use for the described data?

CIRCE's data will be made available under the CC BY-NC license, so as to permit the widest possible re-use, while at the same time retaining the intellectual property of data and maintaining the non-profit nature of the project.

3.1.4.5 Do you have documented procedures for quality assurance of the described data?

Yes

Data quality will be guaranteed by employing the most common data quality assurance methods.

3.1.4.7 Describe the data quality assurance processes

Set up of scientific and technical committee

3.1.4.8 Will you provide any support for data reuse?

Yes

In order to facilitate data analysis and re-use, readme files containing all necessary information (e.g., methodology, analyses, units of measurements, etc.) will be provided.

3.1.4.9 How long do you intend to support data reuse?

Less than 2 years

4.1 Allocation of resources

4.1.1 How will the cost of making the described data findable, accessible, interoperable and reusable be covered?

- Use of institution infrastructure
- Infrastructure Grant

The use of the ILC4CLARIN repository is free of charge and the FAIRification of the project data will be conducted by all members on the basis of a protocol elaborated by CNR-ILC.

On the other hand, the realization of informative and educational material in WP3 involves some costs, as summarized in Table 3:

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

Table 3: *Estimated costs.*

Activity	Type	Cost
3.1/3.4	Development of informative and educational material	7,000€
3.4	Publication of the handbook about accent discrimination	8,000€
3.5	Development and hosting of the interactive platform	5,000€
	Total	20,000€

All the abovementioned costs have been clearly listed in the project and relevant funds from the project budget have been allocated to cover them.

4.1.2 Will you identify a data manager to manage the described data? If not who will be responsible for the management of the data?

Yes

UniSi and CNR-ILC are responsible for the management of the data.

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

All accessible data will be findable for the period necessary to carry out the purposes of the project (see CIRCE's Joint Controller Agreement, Article 3) and for related scientific research, after which only metadata will be guaranteed to remain available.

The researchers involved in the CIRCE project intend to destroy all the participant personal and behavioural data (including their consent forms) as well as all researchers' written notes no later than 10 years after the start of the project (i.e., by 31 December 2032 at the latest).

However, personal data may be retained for statistical or scientific purposes even beyond the period necessary to achieve the purposes for which they were collected or subsequently processed, in accordance with Art. 5, § 1 letter e) of the GDPR.

5.1 Data Security

5.1.1 Where do you plan to keep the described data?

Kept on secure, managed storage for limited time

As already mentioned, CIRCE's datasets will be hosted on the ILC4CLARIN repository, which follows the Code of Conduct for Service Providers.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Ethical issues are fully considered in the project design. Indeed, specific activities in WPs 2 and 3 are devoted to Ethical and Legal issues, in order to collect data in accordance with GDPR and the member Universities' Ethical Committees.

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

CIRCE keeps ethical issues in very high regard and thus a Code of Conduct is being drafted in all members' languages, which will be approved in the next few months. The CIRCE Code of Conduct provides general ethical principles on how CIRCE Members should operate during fieldwork, meetings, publication, conferences, events, committee and other work, mentoring relationships, and in all online spaces.

The purpose of the Code is to provide the CIRCE members with a clear set of guiding principles during the life cycle of the project and, thereafter, for related scientific research and dissemination, as long as necessary.

For instance, students' activities will be designed directly involving school teachers, in order to minimise possible difficulties that can be related to Special Educational Needs.

Moreover, even if CIRCE does not directly address issues relating to sex and/or gender analysis, attention to gender issues is nevertheless duly taken into account. Needless to say, journalism documents may present a feminine or masculine point of view, and in these cases, care will be taken to avoid conveying any gender-dominant or homophobic view, in full respect of the EU policy. Additionally, linguistic discrimination is one of the possible axes on which intersectional discrimination is experienced, since language is, in fact, a personal attribute, but at the same time, it also says something about an individual's ethnic background, gender, sexual orientation, or socioeconomic status. Furthermore, all these individual categories are deeply intertwined and are mutually related and co-constructed socially and linguistically. In this respect, CIRCE activities are, in their own substance, inclusive because they target all the possible forms in which linguistic discrimination can manifest itself.

Finally, participants' data anonymity in WP2 will be ensured by A1 (M1-3). Interviewees/interviewers (or, if minors, their parents) will be asked to sign an informed consent on data protection, in order to allow CIRCE to keep and use all personal data for the purposes of the project and for any future research at the same institution that collected data, as laid out in European and National ethical guidelines.

The consent form will make participants aware of all manners and goals of personal data management and it will be drafted both in English and in all members' languages, so as to clearly describe national and European norms concerning informants' rights to privacy and the specific ways in which these can be exercised.

The informed consent will be published in the 'privacy policy' section of CIRCE's website.

6.1.2 Are the described data sensitive?

Yes

6.1.3 Are the described data personal?

Yes

6.1.4 What are the methods used for processing sensitive/personal data?

Anonymising data where necessary

Sensitive/personal data collected during fieldwork will be treated with utmost protection, using a dissociation process to separate it from informants' identities. A dissociation file will link participant identities with unique identifiers, kept separate from project data. Each partner will have their own

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

dissociation file, which will not be shared with any other partner. Both data and metadata, including participants' personal information, will be coded using participant numbers in tests and data tables so as to ensure anonymity, and publications will employ arbitrary subject numbers. All physical documents will be stored securely in locked closets and digital files will be kept in a controlled environment with authorized access. Researchers will sign confidentiality agreements. Personal data will be retained only for the period necessary to carry out the purposes of the project (see § 2.2), while anonymized data will be preserved for future research with participant consent.

As for managing data on CIRCE's website advanced measures to ensure the security of personal data and strict adherence to data protection guidelines are employed. Regarding data minimization, only essential information required for specific purposes are collected, ensuring that data is kept to a minimum. This not only safeguards user privacy but also reduces the risk of excessive data collection. To guarantee data quality, validation and data verification processes during registration are implemented, since each enregistration need to be approved by a data controller: this procedure ensures that data is accurate, complete, and up-to-date, and avoids the inclusion of outdated or inaccurate information. In terms of limited data retention, stringent unused data deletion and archiving policies are applied. Data is retained only for the period necessary for the stated purposes and is promptly deleted in compliance with applicable regulations.

AES-256 encryption in conjunction with SSL/TLS are used for data transfer and archiving, so as to protect data transmission during user registration and account access. Using password protected backup with AES-256 encryption is an advanced encryption standard that ensures a high level of data security during both transmission and storage. Users' rights to access their data at any time and request its deletion are respected. Procedures and tools to allow users to exercise these rights easily and directly through the personal profile page on the website are provided.

Physical access to data is strictly controlled and limited to authorized personnel, using badges for identification. All accesses and/or changes to data are meticulously tracked. This tracking is done through system logs (server logs) and the use of specific plugins that record the activities of registered users. The recorded activities are centrally stored in a separate database table, allowing constant surveillance of the environment and prompt response to any suspicious activities. In order to provide a secure and compliant computing environment, system configuration management, together with regular updates and advanced configuration, is regular applied.

In any case, as stated in CIRCE's Joint Controllers Agreement (JCA), all members will implement, in accordance with the provisions of Article 32 of the GDPR, all appropriate measures to protect the personal data processed within the project, with special consideration to the destruction, loss, modification, unauthorized disclosure of or access to personal data transmitted, stored or otherwise processed.

All personal data will be processed mainly within the European Economic Area. However, some Personal Data may be transferred outside the EEA, since one of the project's partner is from Bosnia and Herzegovina, namely, Visokoskolska Ustanova Internacionalni Burc Univerzitet - International Burch University (IBU). Nevertheless, all transfer or personal data to this partner will be regulated in accordance with the provisions of Chapter V of the GDPR.

Finally, according to Article 11 of the JCA, appropriate actions will be taken in the event of a personal data breach, in accordance with Article 33 of the GDPR.

CC-BY-NC-4.0 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023
Co-funded by the European Union

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Powered by

 Version 0.0 DOI: 10.5281/zenodo.10085133 29/10/2023

Co-funded by the European Union