READ ME
GramAdapt Social Contact Dataset V1.0.0

This document provides information on how the GramAdapt Social Contact dataset V1.0.0 is structured, and outlines points of importance to be aware of when using this data set.

Created 30-12-2022
Eri Kashima & Francesca Di Garbo

Files

The GramAdapt Social Contact dataset V1.0.0 comprises
- V1.0.0: The combined dataset of all 34 contact sets (csv)
- 34 Individual contact sets (csv)
- GA_SocialContactDataset_Bibliography_V1.0.0 x2: Bibliography of combined dataset (bibtex and formatted rtf)
- GA_V1.0.0_Metadata: Basic metadata about individual contact sets (csv)
- ReadMe file x2 (txt and pdf)

Terminology

Respondent: The expert who filled out the questionnaire. There may be more than one person per set.

Editor: A member of the GramAdapt project team who collected data by corresponding with respondents, and curated the set.

Reviewer: A member of the GramAdapt project team who checked respondent responses, and corresponded with them to finalise answers for publication. All reviewers are also editors of the dataset.

Particularities of the dataset

This dataset is a first attempt at collecting socio-cultural-demographic data about contact scenarios, including attempts at quantification as well as qualitative elaborations by respondents.

Respondents were approached as academic collaborators, not as volunteers in an experiment. Respondents were not chosen at random. Respondents were chosen based on their published research and fieldwork experience with either or both Focus and Neighbour language communities.

Each contact set is unique in terms of the timeframe they respond for. We urge researchers who use this dataset to read the Comments column for questions CID P1, P2, and P3 carefully for each set, to get a sense of the heterogeneity of timeframes represented in each set, as well as the whole dataset. See section "About the timeframe" further on.

Table structure of the dataset

The following is a description of the column names and their functions, from left to right.

Version: The version of the questionnaire. The first release version published in January 2023 is version 1.0.0

Set: The unique identifier of a contact pair. The two digit IDs were assigned based on order of completion. Sets that are linked by language communities, but represent different time slices, contain a Roman alphabet symbol, i.e. Set06a and Set06b are for the contact scenario between Maltese and Sicilian, but for different time periods of contact.

Old_Set_ID: Legacy identifiers of contact pairs. These IDs have been included as they were used for project internal purposes, publications, and presentations created until December 2022. For instance, Di Garbo et al (2021) and Di Garbo & Napoleão de Souza (accepted) make reference to contact pairs using these old set IDs.

CID: Abbreviation of “Correspondence ID”. This ID identifies sections within the Domains Questionnaire, as well as the explanatory factor questions that are asked across the different social domains. The specific realisation of the question in each domain is identifiable by the “Question ID” (QID, see below). The symbol correspondences are: ‘P’ = Preamble; ‘D’ = Domain characterisation; ‘S’ = Social network; ‘B’ = Behaviour affecting biases; ‘O’ = Linguistic output of Focus group people; ‘I’ = Linguistic input of Focus group people, i.e. the output of Neighbour group people; ‘T’ = Language transmission to children of Focus group people; ‘E’ = Ending questions about data source and confidence.

For the Overview Questionnaire, the CIDs and QIDs are treated as one in the same.

QID: Abbreviation of “Question ID”. This ID gives a unique identifier to each question in the questionnaire, regardless of its CID. The first letter indicates whether the question is part of the Domains Questionnaire (‘D’), or the Overview Questionnaire (‘O’). The three letter IDs beginning with ‘D’ indicate which social domain the question is part of: ‘DEM’ = Social exchange and marriage; ‘DFK’ = Family and kin; ‘DKN’ = Knowledge; ‘DLB’ = Labour; ‘DLC’ = Local community; ‘DTR’ = Trade.

For the Overview Questionnaire, the CIDs and QIDs are treated as one and the same. The character following the ‘O’ indicates that those overview questions are thematically linked. The symbol correspondences are: ‘OD’ = Demographics; ‘OG’ = Language geography; ‘OI’ = Language and identity; ‘OL’ = Literacy; ‘OS’ = Social structure; ‘OH’ = History; ‘OE’ = Respondent fieldwork experience (self report); ‘OC’ = Response confidence (respondent self report).

Sub-ID: IDs that link certain questions together by virtue of being asked together in the questionnaire. IDs of questions that were originally asked as part of a single question. Questions linked by Sub-IDs are thematically related.

Wording: The wording of the question.

Response: Data. The often pre-determined response to a question chosen by the collaborator. For possible data types, see “DataType” in this section further on.

Comment: Qualitative data. Clarifications, elaborations, and qualifications from the respondent regarding their response. Editors may also add their comments and clarifications, identifiable by square brackets.

Respondent: Name of collaborator who answered the particular question in the questionnaire.

Dom: Indicating the domain to which the responses apply. Possible options are the overview questionnaire (OV) and social domains (DEM = Exchange and Marriage; DFK = Family and Kin; DKN = Knowledge; DLB = Labour; DLC = Local Community; DTR = Trade).

DomOrder: The order in which the respondent answered this particular questionnaire. i.e. DomOrder ‘1’ means a question belongs to the first questionnaire to which the respondent answered; DomOrder ‘2’ the second questionnaire, and so on. Domain orders were randomised per respondent.

DataType: Indicates the possible answer form to a question. The possible data types are as follows:

- Binary-YesNo: A binary answer of either ‘Yes’ or ‘No;
- Comment: Not preset, just a comment field, i.e. free response.
- Scalar: A Likert 5 point scale. The response is in textual form, but can be numeracised on a scale of 1-5 (e.g. “Neither positive nor negative” -> 3).
- Types: A list of preset answers where only one can be chosen (e.g. “FL, NL, Some other language, This is highly contextual”)
- Types-Multiple: A list of preset answers, where multiple can be chosen
- Value: A numerical value
- B: B for “blank”. The question is relevant to the domain in question, but the respondent chose not to answer the question. This is qualitatively different from an NA response (“not applicable”), which indicates the question is not applicable for a particular sample set. NAs only arise in the dataset 1) if a social domain is

assessed as a “no social contact” domain by the respondent, and/or 2) if a “no” response to questions O1, I1, or T1, renders subsequently dependent questions irrelevant (i.e. a “no” response to O1 means questions O2 and O3 should be skipped.)

Answer1-8: List of possible responses for that particular question. For example, for a Binary type question, Answer 1 column will have “Yes”, Answer 2 column will have “No”, the columns Answer 3-8 will be NA

Surname: Surname of the respondent.

[q2o1answer]/FLang: The Focus Language name, e.g. “Maltese”

F_ISO: Focus language ISO code, e.g. “mlt”. For languages that do not have an ISO code, the Glottocode is used instead.

F_Glottocode: The Glottocode of the Focus Language, e.g. “malt1254”.

[q2o2answer]/NLang: The Neighbour Language name, e.g. “Sicilian”.

N_ISO: Neighbour language ISO code, e.g. “scn”. For languages that do not have an ISO code, the Glottocode is used instead.

N_Glottocode: The Glottocode of the Neighbour Language, e.g. “sici1248”.

ContactPair: The name of the contact pair, spelled out. e.g. “Maltese-Sicilian”

ContactPair_ISO: The ISO codes of the contact pair, e.g. “mlt-scn”

ContactPair_Glottocode: The Glottocodes of the contact pair, e.g. “malt1254-sici1248”

AArea: Information about the geographical location of each contact set based on Autotyp areal classification

Reviewer: The GramAdapt team member responsible for checking the responses. A1 (Author 1), A2 (Author 2), A3 (Author 3).

Respondent comments vs Reviewer comments

Comments are directly written by the authors of the respective set. For multiauthor sets, the specific respondent can be identified by the name associated with the particular question by looking at the 'Respondent' column.

Any comments in [square brackets] are those included by the editorial team for clarification.

About the timeframe

As mentioned in the "Particularities of this dataset" section, each contact set is unique in terms of the timeframe they respond for. We urge researchers who use this dataset to read the Comments column for questions CID P1, P2, and P3 carefully for each set, to get a sense of the heterogeneity of timeframes represented in each set, as well as the whole dataset.

The timeframes given in CIDs P2N and PN3 are broad and coarse approximations, often negotiated between the respondent and reviewer. These timeframes are to be used with caution. Always read the associated comments in the Comment column.

An end date of 2020 indicates that social contact is ongoing at the time of data collection in 2021.

Idiosyncratic contact sets

Set26 "Garifuna - Galibi" only contains responses for the Overview.

Set10 "FLNA - NLNA" see "Data Sensitivity" section below.

Set22 "Muak Sa-aak - Tai Lue" see "Data Sensitivity" section below.

Data Sensitivity

The respondents of sets 10 and 22 have requested access restrictions to their respective datasets.

The respondent of set10 has requested to have community identifying names anonymised. The respondent name for set 10 has also been anonymised. If you wish to access the community identifying names for set10, please contact Kaius Sinnemäki at the University of Helsinki, and he will get in contact with the author of set10.

The respondent for set22 has requested to make certain comments publicly invisible, due to their potentially sensitive nature. If you wish to access the invisible comments of set 22, please contact the author directly.

##Bibliography

References cited in the dataset are combined and available in the files named "GA_SocialContactDataset_Bibliography_V1.0.0". There is a Bibtex and formatted txt version available.

##Citing the dataset

The whole dataset should be cited using the format on the Zenodo page:

Eri Kashima, Francesca Di Garbo, Oona Raatikainen, Rosnátaly Avelino, Sacha Beck, Anna Berge, Ana Blanco, Ross Bowden, Nicolás Brid, Joseph M Brincat, María Belén Carpio, Alexander Cobbinah, Paola Cúneo, Anne-Maria Fehn, Saloumeh Gholami, Arun Ghosh, Hannah Gibson, Elizabeth Hall, Katja Hannß, Hannah Haynie, Jerry Jacka, Matias Jenny, Richard Kowalik, Sonal Kulkarni-Joshi, Maarten Mous, Marcela Mendoza, Cristina Messineo, Francesca Moro, Hank Nater, Michelle A Ocasio, Bruno Olsson, Ana María Ospina Bozzi, Agustina Paredes, Admire Phiri, Nicolas Quint, Erika Sandman, Dineke Schokkin, Ruth Singer, Ellen Smith-Dennis, Lameen Souag, Yunus Sulistyono, Yvonne Treis, Matthias Urban, Jill Vaughan, Deginet Wotango Doyiso, Georg Ziegelmeyer, Veronika Zikmundová. (2023). GramAdapt Crosslinguistic Social Contact Dataset. (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7508054

The citing of individual sets should take the following format, for example, set28:

Kashima, Eri & Schokkin, Dineke. Set28: Nen and Idi. InEri Kashima, Francesca Di Garbo, Oona Raatikainen, Rosnátaly Avelino, Sacha Beck, Anna Berge, Ana Blanco, Ross Bowden, Nicolás Brid, Joseph M Brincat, María Belén Carpio, Alexander Cobbinah, Paola Cúneo, Anne-Maria Fehn, Saloumeh Gholami, Arun Ghosh, Hannah Gibson, Elizabeth Hall, Katja Hannß, Hannah Haynie, Jerry Jacka, Matias Jenny, Richard Kowalik, Sonal Kulkarni-Joshi, Maarten Mous, Marcela Mendoza, Cristina Messineo, Francesca Moro, Hank Nater, Michelle A Ocasio, Bruno Olsson, Ana María Ospina Bozzi, Agustina Paredes, Admire Phiri, Nicolas Quint, Erika Sandman, Dineke Schokkin, Ruth Singer, Ellen Smith-Dennis, Lameen Souag, Yunus Sulistyono, Yvonne Treis, Matthias Urban, Jill Vaughan, Deginet Wotango Doyiso, Georg Ziegelmeyer, Veronika Zikmundová. (2023). GramAdapt Crosslinguistic Social Contact Dataset. (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7508054

##References used in this Read Me

Di Garbo, Francesca & Kashima, Eri & Napoleão De Souza, Ricardo & Sinnemäki, Kaius. 2021. Concepts and methods for integrating language typology and sociolinguistics. Atti del Workshop SLI “Sociolinguistica e tipologia: verso un approccio integrato allo studio della variazione”. 5. 143–176. (doi:10.17469/O2105SLI000005)

Di Garbo, Francesca and Ricardo Napoleão de Sousa. Aaccepted. A sampling technique for worldwide comparisons of language contact scenarios. To appear in Linguistic Typology.