ITACA
Corpus Handbook

Edited by

Arianna Bienati, Jennifer-Carmen Frey, Lorenzo Zanasi, Chiara Vettori

Institute for Applied Linguistics
Eurac Research

Version 1.0
December 2024

Contents

Corpus figures and sample description

The ITACA dataset comprises 636 students who completed at least one of the three tasks (writing task, language assessment test or questionnaire). One student did not complete the writing task, so that the corpus contains only 635 texts. 14 students did not complete the questionnaire and 24 did not complete the language assessment test. Therefore, the students who completed both questionnaire and writing task are 621 and those who completed all tasks were 602.

Population description and sampling procedures

For the purpose of analysis, a representative sample of the South Tyrolean students' population enrolled in schools with Italian as the medium of instruction in the school year 2021/2022, stratified per school type (vocational vs. high school) was drawn from the entire corpus of 635 texts. These texts are marked with the metadata `sample_repr_south_tyrol=True`.

The sample was calculated considering the strata related to students enrolled in vocational schools (*istituti*) and in high schools (*licei*). It adds up to a total of 569 students, of which 554 have completed both the writing task and the sociolinguistic questionnaire.

A total of 1308 students were enrolled in the fourth year of vocational (*istituti* tecnici e professionali) and high schools (*licei*) in the Province of Bozen/Bolzano in the school year 2021/2022. Of these, 43.4% were enrolled in vocational schools and 56.6% in high schools. A representative sample of this student population results in 588 valid participants (and texts) to be drawn from all educational institutions in South Tyrol. Therefore, accounting for potentially ur unusable texts with a coefficient known from previous research conducted in the territory, a sample of 659 students was drawn. Knowing the number of classes and the average class size per educational institution, 38 classes were extracted, respectively 18 classes in vocational schools and 20 classes from high schools, to reflect the distribution of students enrolled in different types of schools. In the extraction, a criterion of proportionality was applied: if an institution covered the 8% of students enrolled in high schools, an appropriate number of classes from that institution was extracted to cover approximately 8% of the high school sample.

The sample to be drawn was made up as follows:

	N students	**%**	**** 588 sample****	**Corrective surcharge**	**Classes to extract**
Istituto	567	43.4%	255	286	18
Liceo	741	56.6%	333	373	20
	1308	**100.0%**	**588**	**659**	**38**

Starting from these calculations, educational institutions have been chosen for data collection. However, given the difficulties in convincing some schools to participate, the return on the type

of educational institution was as follows: 90% of the vocational schools in the province participated with at least one class (9 out of 10 schools) and 82% of the high schools in the province participated with at least one class (9 out of 11 high schools).

	Planned sample	**Actual sample**	**Sample coverage**
Istituti	10	9	90%
Licei	11	9	82%

Within those 9 *licei* and *istituti* 18 and 20 classes have been chosen for data collection. From the texts collected in these classes, 554 students have been randomly chosen for creating a sample representative of the South Tyrolean school reality. This sample maintains almost perfectly the proportion between school types envisaged in the beginning and it has been thus deemed representative of the student population under investigation.

	Collected sample	**Sample proportions**	**Classes in the sample that could be used**
Istituto	243	43.9%	18
Liceo	311	56.1%	20
	554	**100.0%**	**38**

Corpus metadata

Text and person-related metadata

The following text and person-related metadata for the ITACA corpus can be found in `core_metadata.csv` (in the folder metadata).

Text and sample identifiers

`text_id`
Unique and anonymous identifier for each student essay composed of a combination of a six-character long anonymous identifier for the class and a six-character long identifier for the respective student that are joined with an underscore (e.g., `AC20PA_BAZ14M`).

`sample_repr_south_tyrol`
Information, whether the essay was part of the statistically drawn representative sample. Note that `core_metadata.csv` contains reference to some essays that were collected but had to be filtered out for various reasons at a later stage. Texts pertaining to the representative sample are marked with `TRUE`, all other texts are marked with `FALSE`.

`sample_curated`

Information, whether the essay was part of the curated sample. Curated texts represent the ITACA gold standard. It comprises a set of 40 texts (10% circa of the whole annotated corpus), of which 10 are annotated by all three annotators who took part in the annotation process and the remaining 30 are annotated by couples of two annotators.

All these texts have been reviewed by the project team members during the curation phase, a moment in which discrepancies between annotations made by different annotators have been discussed and resolved to arrive at a consensus and build the gold standard.

`sample_annotated`

Information on whether the essay was annotated for cohesion and coherence features according to the annotation schema described below. The sample consist of 388 texts. Of which 40 texts belong to the sample of curated texts, annotated by all three annotators and the remaining 348, which have been annotated by only one out of three annotators.

`author_id`
Unique and anonymous identifier for each student, i.e. the author of the essay, composed of six alphanumeric characters.

`class_id`
Unique and anonymous identifier for the school class in which the essay has been collected, composed of six alphanumeric characters.

`school_id`
Unique and anonymous identifier for the school in which the essay has been collected, composed of 2 characters.

`school_type`
The type of school in which the essay was collected.

- `liceo`, i.e. grammar school
- `istituto`, i.e. technical high school

Note:
Each author contributed with one essay. Authors are nested in classes which are again nested in schools of different school types. Missing values are encoded with `'NA'`.

Author metadata

Person-related metadata on study participants have been collected in an online questionnaire and can be identified by the metadata prefix `author_*` These include general sociodemographic information, metadata regarding the language background of the student as well as metadata on reading and writing habits (that are, however, recorded in the file `rw_habits.csv`.

Sociodemographic metadata on participating students

The following items were collected in the project:

`author_birthyear`
The year in which the student was born, elicited by the question: "In che anno sei nato/a? (quattro cifre)". Possible responses include 4-digit birthyears between 2001-2005.

`author_gender`
The students self-defined gender, elicited by the question: "Sei un/a". Possible responses include: `Maschio | Femmina | Preferisco non dirlo | Altro`

`author_dsa`

Encodes the presence of learning disabilities and special needs (Disturbi Specifici dell'Apprendimento), elicited by the questionnaire item: "Hai una diagnosi di DSA (ad es: dislessia, disgrafia, discalculia, ecc.)?". Possible responses include: `Sì | No`

`author_legge_104`
Encodes the presence of other disabilities defined in the Italian legislation (legge 104/1992), elicited by the questionnaire item: "Sei portatore/portatrice di handicap ai sensi della legge 104/1992?". Possible responses: `Sì | No`

`author_SES`
The student's socio-economic background. This value was calculated as the first dimension of a factor analysis using questionnaire items on the student's home possessions, the parents' highest educational degree and the parents' job category (see section Calculation of socio-economic status variable).

The following questionnaire items have been regarded.

For the highest educational degree of the parents:

- 'Indica il titolo di studio più alto conseguito dai tuoi genitori. Seleziona l'opzione "non si applica" in assenza della figura genitoriale.'

For the highest job classification of the parents:

- 'I tuoi genitori lavorano? Seleziona l'opzione "Non si applica" in assenza della figura genitoriale.'
- 'Tra le seguenti categorie, indica quale descrive meglio la professione dei tuoi genitori. Se in pensione, non più abile al lavoro, in cassa integrazione o disoccupato/a, fai riferimento all'ultima professione svolta.'
- 'Puoi descrivere la professione dei tuoi genitori? Denominazione del lavoro (ad es. operaio tornitore qualificato, magazziniere, avvocatessa, insegnante di scuola secondaria di I grado, ragioniera in un ufficio amministrativo, ecc.). Se deceduto/a, in pensione, in cassa integrazione o disoccupato/a fai riferimento all'ultima professione svolta, mentre nel caso in cui la professione sia sconosciuta indica non lo so. Ti preghiamo di essere il più preciso/a possibile.'

For students' home possessions:

- 'A casa tua ci sono:'
    - Una camera per te
    - Un computer che puoi usare per lo studio
    - Un collegamento Internet
    - Libri di letteratura classica (ad es. Dante)

    - Opere d'arte (ad es. quadri)
    - Una lavastoviglie
    - Un televisore a schermo piatto
- 'Quante di ciascuna delle seguenti cose ci sono a casa tua?'
    - Smartphone
    - Televisore/i
    - Computer/tablet
    - Automobile/i
    - Locali con vasca da bagno/doccia
- 'Quanti libri ci sono a casa tua? Considera che ogni metro di scaffale contiene circa 40 libri. Non includere nel calcolo le riviste, i giornali o libri scolastici.'
    - 0-10 libri
    - 11-25 libri
    - 26-100 libri
    - 101-200 libri
    - 201-500 libri
    - Più di 500 libri

Calculation of socio-economic status variable

For the purposes of the study, we furthermore constructed an Economic, Social and Cultural Status index to measure the effect of family background on the learning outcomes of the students participating in the ITACA project. This research largely follows the methodology proposed by OECD-PISA (PISA, 2018), albeit with some modifications dictated by the specificity of the data taken into consideration.

The ESCS index is based on the following variables:

- parents' employment status (HISEI)
- parents' level of education expressed in years of formal education (PARED)
- possession of certain material goods understood as proximity variables of an economic-cultural context favourable to learning (HOMEPOS).

The information used for the construction of these variables was obtained from the sociolinguistic questionnaire administered to the students (see survey.pdf), more specifically from the information gathered with the questions 40-50. The questionnaire data contained information for 622 of the 635 students. For the remaining 13 students, who did not fill in the questionnaire, no ESCS could be calculated resulting in a missing value.

The employment status of parents (HISEI)

The information used for the construction of the index of the occupational status of the student's parents (HISEI) was obtained from the questions 40-42 (with respect to parent 1) and the questions 43-45 (with respect to parent 2) of the socio-demographic part of the questionnaire administered to the pupils. The occupational data of the student's parent 1 and parent 2 were obtained from the answers to the open-ended questions (Q42 and Q45). The answers were coded with the four-digit ISCO codes (ILO, 2007) and then mapped to the International Socioeconomic Employment Status Index (ISEI) using the R package ISCO08ConveRsions (Schwitter, 2023). Based on this information, three indices were calculated: the employment status of parent 1; the employment status of parent 2; and the

highest parental employment status (HISEI), which corresponds to the highest ISEI score of either parent or the ISEI score of the only available parent. For all three indices, higher ISEI scores indicate higher levels of employment status. Information on the occupation of both parents were missing in 57 students of the 635 students in the ITACA corpus.

Parents' level of education (PARED)

Another variable related to the student's family background that was taken into account was the parents' level of education. Based on the information given in Q39 which asked for the educational qualification of both parents (codified in the seven ISCED levels), the PARED index provides an estimate of years of schooling according to the highest educational qualification of both parents. In 12 cases the information on the educational qualification was missing for both parents and could thus not be calculated.

Economic-cultural well-being (HOMEPOS)

In this study, economic wellbeing was explored through information derived from questions 49 and 50, in which students reported the quantity of certain goods (Q49) and the estimated amount of books present in the home (Q50). In this case, data was missing just for those students who did not fill in the questionnaire (N=13). The synthetic variable HOMEPOS was calculated using the R package TAM which offers various types of Item Response Models. We based our HOMEPOS variable on the weighted likelihood estimates of a Partial Credit Model (PCM) (Masters G.N.& Wright B.D., 1997). The PCM model represents a form of Rasch Model for polynomial data.

The Economic, Social and Cultural Status index (ESCS)

This index has been defined and consequently calculated following the methodological indications that have found greater acceptance in the international context. The ESCS was constructed on the basis of the three indices illustrated above: the parents' employment status (HISEI), their level of education (PARED) and whether or not they own particular capital and cultural assets (HOMEPOS).

The ESCS composite index is derived by means of a principal component analysis (PCA) of the three indices HISEI, PARED and HOMEPOS. As a preliminary step to the PCA, missing values were imputed with the R package missMDA (Josse & Husson, 2016). The results of the analysis report that the first extracted component is able to explain approximately 64% of the overall variance and that it is associated with a Eigen value greater than 1 (1.91) and with the following factor loadings: parents' employment status (HISEI) 0.85; their level of education (PARED) 0.82; family background (HOMEPOS) 0.73. The ESCS of each student was then calculated by centering the coordinates of the first dimension, so that a student with a strictly positive individual ESCS value is a student with a more favourable socio-economic-cultural background than the average of the study participants. The size of the deviation from the average of the students participating in the study can be assessed as a function of the standard deviation.

Metadata regarding the sociolinguistic background of the participating students

The language background of the students is documented for the following aspects.

Language used in different contexts

The questionnaire collected detailed information on the language(s) that is/are present in a student's environment and thus used and heard most frequently across different contexts. These are recorded in the items starting with `author_context_language_*` and were elicited with the following prompt:

> "Indica, per favore, se il tuo ambiente familiare, scolastico, la tua cerchia di amicizie, ecc. sono/sono stati unicamente o prevalentemente italiani oppure se l'ambiente in cui vivi, vai a scuola, coltivi le tue amicizie è "altro", ovvero arricchito da altre lingue e culture che possono essere quelle "tipiche" altoatesine - tedesca e ladina -, oppure lingue e culture diverse come quella spagnola, inglese, albanese, cinese, araba, eccetera. Per rispondere, pensa soprattutto agli ultimi 10 anni della tua vita (più o meno dall'inizio della scuola primaria) e leggi bene quali sono le possibili risposte."

- `author_context_language_family`: l'ambiente familiare in senso allargato (ovvero includendo parenti e amici frequentati abitualmente dalla tua famiglia) è/è stato
- `author_context_language_school`: Complessivamente, l'ambiente scolastico è/è stato
- `author_context_language_friends`: Complessivamente, il contesto delle amicizie è/è stato
- `author_context_language_free_time`: Complessivamente, il contesto del tempo libero (es. associazioni sportive, parrocchia, centri giovani, associazioni culturali, ecc.), è/è stato

Possible responses were:

- `Unicamente italiano`
- `Più italiano che altro`
- `Italiano e altro in egual misura`
- `Più altro che italiano`
- `Unicamente altro`

Habitual language

The questionnaire collected detailed information on the language(s) that is/are used with different people. These are recorded in the items starting with `author_habitual_language_*` and were elicited with the following prompt:

> In quale lingua parli abitualmente con... Scegli l'opzione "non si applica" in caso di assenza della relativa figura familiare.

- `author_habitual_language_parent1`: Genitore 1
- `author_habitual_language_parent2`: Genitore 2
- `author_habitual_language_siblings`: Fratelli/sorelle
- `author_habitual_language_relatives`: Parenti
  `author_habitual_language_friends`: Amici

Possible responses were:

- `Unicamente in italiano`
- `Più italiano che in altra/e lingua/e`

- In egual misura in italiano e in altra/e lingua/e
- Più in altra/e lingua/e che in italiano
- Unicamente in altra/e lingua/e
- Non si applica

Language of education

The questionnaire also collected data on the languages primarily used in educational contexts for different phases of a student's educational path. These are recorded in the metadata items `author_schooling_language_*` and were elicited with the following prompt:

> Nella tabella sottostante indica, per favore, quali scuole hai frequentato finora:

author_schooling_language_kindergarten	Scuola dell'infanzia
author_schooling_language_primary	Scuola elementare
author_schooling_language_middle	Scuola media
author_schooling_language_high	Scuola superiore

Possible responses were:

- in lingua italiana
- in lingua tedesca
- in lingua ladina
- in altra lingua (ad es. in un altro Paese)

First language (L1)

Students have been asked to self-declare their first language, and the first language of their parents, based on which language(s) they know best. Metadata regarding the student's and parents' first language are recorded in the items: `author_L1`, `parent1_L1` and `parent2_L1`. This information has been elicited with the following prompt:

> Nella tabella sottostante indica, per favore, quale consideri essere la tua prima lingua, ovvero la lingua che conosci meglio. Fai lo stesso pensando ai tuoi genitori. Scegli l'opzione "non si applica" in caso di assenza della figura genitoriale.
>
> Se hai indicato "Altra lingua" o "Due o più lingue", ti preghiamo di specificare quale/i:

Possible responses included: `Italiano | Tedesco | Ladino | Altra lingua | Due o più lingue`

Since this question prompted a multitude of different language combinations, the final L1 labels have been simplified recording only the official languages of South Tyrol, Italian (`ITA`), German (`DEU`) and Ladin (`LLD`), and binning all other languages with a placeholder `OTHER` to maintain student's anonymity. Students indicating more than one language as their or their parents' first language are marked with a `+` indicating which languages they combine. Occasionally information could not be reconciled leading to missing values.

Language environment

Moreover, a number of questions aimed at recording the language environment a student was surrounded with, asking for their residence and their or their parents' birthplace as well as when they arrived in South Tyrol in case of people being born abroad or in another Italian region.

author_residence	In quale comune altoatesino vivi attualmente? (Se NON abiti in Alto Adige, seleziona la voce "Non si applica")"
author_birthplace	Dove sei nato/a? Se sei nato/a all'estero, ti preghiamo di specificare il Paese.
author_age_arrival_southtyrol	Se non sei nato/a in provincia di Bolzano, a che età sei arrivato/a in Alto Adige?
parent1_birthplace **parent1_birthplace_abroad**	Dove sono nati i tuoi genitori? Seleziona l'opzione "non si applica" in assenza della figura genitoriale. GENITORE1
parent2_birthplace **parent2_birthplace_abroad**	Dove sono nati i tuoi genitori? Seleziona l'opzione "non si applica" in assenza della figura genitoriale. GENITORE 2

Language assessment score

Finally, a language assessment test (TVI, Test di Verifica in Ingresso from University of Bergamo) was conducted with participating students. The final score is recorded in the metadata item `tvi_score`.

Metadata on reading and writing habits

During the ITACA project, information on reading and writing habits of study participants has been collected in an online questionnaire and can be found in `rw_habits.csv` (in metadata). The collected items are detailed below.

Reading attitudes

`author_reading_attitudes_*` items are intended to capture the reading attitudes of students. They were translated from Schutte et al. (2007).

They were administered with the following request:

> Le seguenti affermazioni riguardano il tuo rapporto con la lettura. Indica, per favore, il tuo grado di accordo o disaccordo con ciascuna affermazione, cercando di prendere posizione e di limitare il più possibile la risposta 'né d'accordo né in disaccordo'.

The items are recoded as:

author_reading_attitudes_1	Se un libro o un articolo è interessante, non mi importa quanto è difficile da leggere.
author_reading_attitudes_2	Senza la lettura la mia vita non sarebbe la stessa.
author_reading_attitudes_3	I miei amici/le mie amiche a volte sono sorpresi da quanto leggo.
author_reading_attitudes_4	A me e ai miei amici/alle mie amiche piace scambiarci libri o articoli che ci piacciono particolarmente.

author_reading_attitudes_5	Per me è molto importante passare del tempo a leggere.
author_reading_attitudes_6	Rispetto ad altre attività, per me la lettura è importante.
author_reading_attitudes_7	Se ho bisogno di informazioni dal materiale che leggo, finisco di leggere molto prima di quando devo studiare e imparare ciò che leggo.
author_reading_attitudes_8	I miei voti a scuola rispecchiano quanto è efficace il mio modo di leggere.
author_reading_attitudes_9	Attraverso il mio rapporto con la lettura rappresento un buon esempio per gli altri.
author_reading_attitudes_10	Leggo velocemente.
author_reading_attitudes_11	La lettura mi aiuta a dare un senso alla mia vita.
author_reading_attitudes_12	Per me è importante ricevere complimenti per le conoscenze che acquisisco leggendo.
author_reading_attitudes_13	Mi piace che gli altri mi facciano domande su ciò che leggo, perché così posso mostrare la mia conoscenza.
author_reading_attitudes_14	Non mi piace leggere materiale tecnico (manuali, libretti di istruzioni, ecc.). [REVERSED ITEM]
author_reading_attitudes_15	Per me è importante che gli altri notino quanto leggo.
author_reading_attitudes_16	Mi piacciono i libri o gli articoli difficili e impegnativi.
author_reading_attitudes_17	Non mi piace leggere materiale con un vocabolario difficile. [REVERSED ITEM]
author_reading_attitudes_18	Faccio tutte le letture previste per la scuola.
author_reading_attitudes_19	Sono sicuro di poter capire libri o articoli difficili.
author_reading_attitudes_20	Sono un buon lettore/una buona lettrice.
author_reading_attitudes_21	Leggo per migliorare il mio rendimento a scuola.

The Likert scale has not been recoded.

Writing attitudes

`author_writing_attitudes_*` items aim to capture the writing attitudes of students. They were translated from Troia et al. (2013). Original English items are not published with the article. They were shared by the authors with the ITACA corpus compilers.

Items for this category were administered with the following request:

> Le affermazioni che seguono riguardano il tuo rapporto con la scrittura. Indica quanto ciascuna delle affermazioni è vera per te.

The items are recoded as:

author_writing_attitudes_1	Mi piace scrivere testi impegnativi che mi fanno riflettere.
author_writing_attitudes_2	Credo di saper scrivere dei buoni temi.
author_writing_attitudes_3	Preferisco più leggere che scrivere. [REVERSED ITEM]
author_writing_attitudes_4	È importante per me che il mio insegnante dica agli altri che so scrivere bene.
author_writing_attitudes_5	Mi piace scrivere, così posso imparare di più sugli argomenti che mi interessano.
author_writing_attitudes_6	Cerco di scrivere il meno possibile. [REVERSED ITEM]

author_writing_attitudes_7	Quando scrivo un tema, mi risulta facile trovare degli argomenti da includere in esso.
author_writing_attitudes_8	Scrivere può essere molto divertente.
author_writing_attitudes_9	Mentre rileggo e modifico quanto ho scritto, trovo difficile individuare tutti gli errori. [REVERSED ITEM]
author_writing_attitudes_10	Scrivo soprattutto perché devo farlo per la scuola. [REVERSED ITEM]
author_writing_attitudes_11	Se prendo un buon voto in un compito di scrittura, è perché ho avuto fortuna. [REVERSED ITEM]
author_writing_attitudes_12	Mi piace scrivere su argomenti nuovi.
author_writing_attitudes_13	Nei temi cerco di ottenere il voto più alto della classe.
author_writing_attitudes_14	Se un compito di scrittura è interessante, non mi importa quanto è difficile.
author_writing_attitudes_15	Mi piace ricevere complimenti per come scrivo.
author_writing_attitudes_16	Quando in classe ci viene chiesto di scrivere un saggio, una relazione o una storia, il mio testo è uno dei migliori.
author_writing_attitudes_17	So scrivere dei buoni temi perché scrivere per me è facile.
author_writing_attitudes_18	Imparare a scrivere bene mi aiuterà a diventare un adulto di successo.
author_writing_attitudes_19	Quando scrivo un testo, è difficile per me decidere cosa va al primo, secondo, terzo posto eccetera. [REVERSED ITEM]
author_writing_attitudes_20	Se prendo un buon voto in un tema, è perché mi sono impegnato molto. [REVERSED ITEM?]
author_writing_attitudes_21	Quando scrivo un tema, è facile per me scrivere le mie idee usando frasi ben strutturate.
author_writing_attitudes_22	Mi piace scrivere.
author_writing_attitudes_23	Se non devo correggere quello che ho scritto, ne sono felice. [REVERSED ITEM]
author_writing_attitudes_24	Quando scrivo un tema, faccio fatica a trovare le parole giuste per esprimere quello che voglio dire. [REVERSED ITEM]
author_writing_attitudes_25	Mi sento realizzato se vedo che le mie capacità di scrittura sono migliorate molto.
author_writing_attitudes_26	Per me è importante mostrare agli altri che sono in gamba, attraverso ciò che scrivo.
author_writing_attitudes_27	Preferirei non avere compiti scritti (temi, relazioni, ecc.) da fare a casa. [REVERSED ITEM]
author_writing_attitudes_28	Riuscire a esprimere per iscritto un concetto davvero difficile o confuso mi dà molta soddisfazione.
author_writing_attitudes_29	Non mi piace dover svolgere compiti di scrittura difficili. [REVERSED ITEM]
author_writing_attitudes_30	Non prendo spesso buoni voti negli scritti perché non sono abbastanza intelligente. [REVERSED ITEM]

Writing habits

`author_writing_habits_*` items intend to capture the types of writing most written by students. The items were taken from questionnaires used by the Institute for Applied Linguistics for the project “KoKo: Eine korpusunterstützte Analyse der Sprachkompetenzen bei Lernenden im deutschen Sprachraum unter besonderer Berücksichtigung des Deutschen in Südtirol“.

They were administered with the following request:

> In una settimana, mediamente con quale frequenza scrivi i seguenti testi? Una risposta per singola voce.

The response items are recoded as:

author_writing_habits_email	E-mail
author_writing_habits_social	Post sui Social Network (Facebook, Instagram, Tumblr, ecc.)
author_writing_habits_blog	Blog post
author_writing_habits_writers_forum	Testi su siti/social network per scrittori (Wattpad, The Incipit, Penne Matte, ecc.)
author_writing_habits_whatsapp	Messaggi istantanei (WhatsApp, Telegram, Messenger, Viber, ecc.)
author_writing_habits_twitter	Tweet (Twitter)
author_writing_habits_diary	Pagine di diario personale
author_writing_habits_literature	Poesie, racconti
author_writing_habits_songs	Canzoni
author_writing_habits_other	*Se ti dedichi alla scrittura di altri generi testuali, ti preghiamo di specificare a che genere di testo fai riferimento:

The frequency scale has not been recoded.

The single item `author_writing_publish` was meant to ask whether the student ever published one of his writings either online or in print.

It was administered with the following request:

> Hai mai scritto un testo (ad es. racconto, poesia, saggio, romanzo, ecc.) da pubblicare a stampa oppure online su siti ad hoc o su un blog personale?

Reading habits

`author_reading_habits_*` items and `parents_reading_habits_*` items intend to capture the reading habits of the students and their parents: Whether they read outside school, which kind of books, online or printed, and whether they read newspapers, magazines, etc.

All items are taken from "Indagine multiscopo sulle famiglie - I cittadini e il tempo libero 2015" by ISTAT (Istituto Nazionale di Statistica) and were adapted by Chiara Vettori to fit the target population.

Outside school (extracurricular)

The `author_reading_habits_no_school` is a yes/no question that asks whether a student reads outside the school. To this question a logic is applied.

- YES: the student continues answering to the `author_reading_habits_no_school_*` items.

- NO: the students is required to give a reason why he does not read outside the school, continuing to the `author_reading_habits_no_school_why_*`

The `author_reading_habits_no_school` was administered with the following request:

> "Negli ultimi 12 mesi hai letto dei libri, in formato cartaceo o elettronico o ascoltato degli audiolibri? Considera solo i libri letti per motivi non strettamente scolastici (= escludi i romanzi e/o le e opere che hai letto per “obbligo scolastico”)."

If NO, provide reasons, prompted with the following request:

> "Puoi indicare i motivi per cui negli ultimi 12 mesi non hai letto libri per motivi non strettamente scolastici?"

The response items are recoded as columns with these names. Values in the columns are the same as the original response items, except for the last one, i.e. 'other (please specify)', for which open answers have been collected from students.

author_reading_habits_no_school_why_1	I libri costano troppo
author_reading_habits_no_school_why_2	Non ho un posto tranquillo dove leggere
author_reading_habits_no_school_why_3	Ho poco tempo libero
author_reading_habits_no_school_why_4	I libri sono scritti in modo difficile
author_reading_habits_no_school_why_5	Sono troppo stanco/a dopo avere studiato/fatto i compiti
author_reading_habits_no_school_why_6	Mi annoia, non mi appassiona
author_reading_habits_no_school_why_7	Preferisco altri svaghi
author_reading_habits_no_school_why_8	Al giorno d’oggi non serve più leggere
author_reading_habits_no_school_why_9	Preferisco altre forme di comunicazione (televisione, radio, computer, cinema)
author_reading_habits_no_school_why_10	Ci vuole troppo tempo, ho bisogno di stimoli più veloci
author_reading_habits_no_school_why_11	È sufficiente essere informati (attraverso giornali, settimanali, riviste)
author_reading_habits_no_school_why_12	Leggo già molti libri per obbligo scolastico
author_reading_habits_no_school_why_13	Altro (ti preghiamo di specificare)

Languages

If YES, they are asked whether they also read in other languages: `author_reading_habits_no_school_languages`, with the following question:

> Hai letto o ascoltato anche libri in una lingua diversa dall'italiano?

`author_reading_habits_no_school_languages` is a yes/no question.

If yes, they are asked which languages: `author_reading_habits_no_school_which_languages`, with the following request:

> Se sì, in quale lingua? Sono possibili più risposte.

And with the following items recodings:

author_reading_habits_no_school_which_languages_de	Tedesco
author_reading_habits_no_school_which_languages_lad	Ladino
author_reading_habits_no_school_which_languages_en	Inglese
author_reading_habits_no_school_which_languages_fr	Francese
author_reading_habits_no_school_which_languages_sp	Spagnolo
author_reading_habits_no_school_which_languages_al	Albanese
author_reading_habits_no_school_which_languages_ar	Arabo
author_reading_habits_no_school_which_languages_ru	Russo
author_reading_habits_no_school_which_languages_other	Altro (ti preghiamo di specificare)

Formats

If NO, they go directly to `author_reading_formats_no school_*` which is a question that asks whether online formats (e-books, online books and audiobooks) have been read in the past 12 months.

> Se pensi al solo formato elettronico/digitale, negli ultimi 12 mesi ti è capitato di fare uso dei seguenti formati di lettura? Nel rispondere, considera sempre solo i libri a cui ti sei avvicinato/a per piacere/interesse personale (= escludi i romanzi e/o le opere che hai letto per "obbligo scolastico").

Items are recoded into following columns with yes/no answers:

author_reading_formats_no school_ebook	E-book
author_reading_formats_no school_onlinebooks	Libri online
author_reading_formats_no school_audiobooks	Audiolibri

The same options are given to answer `author_reading_or_listening_no_school`, which is prompted by:

> In generale preferisci leggere o ascoltare?

Genres

Literary and technical prose

`author_reading_habits_no_school_genre_*` items intend to capture the genres (high culture) most read by students.

They were administered with the following request:

> Quali dei seguenti generi di libri ti è capitato di leggere o ascoltare, per motivi non strettamente scolastici, negli ultimi 12 mesi? Sono possibili più risposte.

The response items are recoded in separate columns that correspond to:

author_reading_habits_no_school_genre_novel	Romanzi, racconti, poesia, teatro
author_reading_habits_no_school_genre_romance	Romanzi rosa
author_reading_habits_no_school_genre_thriller	Gialli, noir
author_reading_habits_no_school_genre_scifi	Fantascienza

author_reading_habits_no_school_genre_fantasy	Fantasy, horror
author_reading_habits_no_school_genre_socialsci	Libri di scienze sociali o umane (filosofia, sociologia, politica, psicologia, storia, pedagogia, ecc.)
author_reading_habits_no_school_genre_natsci	Libri di scienze naturali, esatte, applicate, di tecnica
author_reading_habits_no_school_genre_art	Arte
author_reading_habits_no_school_genre_religion	Religione
author_reading_habits_no_school_genre_music	Musica
author_reading_habits_no_school_genre_instant	Libri di attualità (instant book)
author_reading_habits_no_school_genre_other:	Altro (ti preghiamo di specificare)

Popular genres

`author_reading_habits_no_school_genre_other_*` items intend to capture the genres (pop culture) most read by students.

They were elicited with the following request:

> Oltre a quelli già elencati, di quali altri generi letterari hai letto o ascoltato delle opere, per motivi non strettamente scolastici, negli ultimi 12 mesi? Sono possibili più risposte.

The response items are recoded in separate columns that correspond to:

author_reading_habits_no_school_genre_other_humor	Umoristici
author_reading_habits_no_school_genre_other_hobby	Hobby e tempo libero
author_reading_habits_no_school_genre_other_astro	Astrologia, magia, esoterismo
author_reading_habits_no_school_genre_other_tech	Libri di informatica
author_reading_habits_no_school_genre_other_tourism	Guide turistiche
author_reading_habits_no_school_genre_other_health	Libri sulla salute
author_reading_habits_no_school_genre_other_home	Libri per la casa (cucina, bricolage, maglia, cucito, ecc.)
author_reading_habits_no_school_genre_other_manual	Manuali pratici (ad es. la collana 'For Dummies' [per principianti])
author_reading_habits_no_school_genre_other_photo	Fotografia, cinema
author_reading_habits_no_school_genre_other_comics	Libri a fumetti, graphic novel
author_reading_habits_no_school_genre_other_animals	Libri sugli animali
author_reading_habits_no_school_genre_other_illustr	Albi illustrati
author_reading_habits_no_school_genre_other_bio	Biografie
author_reading_habits_no_school_genre_other_none	Nessun altro genere letterario oltre a quelli già indicati
author_reading_habits_no_school_genre_other_other	Altro (ti preghiamo di specificare)

Frequency, quantity and purchase

The following items give general information on the extracurricular reading habits of students:

`author_reading_habits_no_school_frequency` prompted with:

> Con quale frequenza leggi o ascolti libri per motivi non strettamente scolastici?

`author_reading_habits_no_school_quantity` prompted with:

> Complessivamente, quanti libri hai letto o ascoltato negli ultimi 12 mesi per motivi non strettamente scolastici?

`author_reading_habits_no_school_purchase` prompted with:

> Come sei venuto/a in possesso dell'ultimo libro che hai letto o ascoltato?

Responses have not been recoded.

At school (curricular)

The `author_reading_habits_school` is a yes/no question that asks whether a student reads for the curriculum. This question is formulated as:

> Considera gli ultimi 12 mesi: hai letto o ascoltato libri per motivi scolastici (esclusi i libri di testo obbligatori)?

To this question a logic is applied.

- YES: the student answers the `author_reading_habits_school_quantity` item. This question is formulated as:

> Quanti libri hai letto o ascoltato?

Answers are not recoded.

- NO: the student goes directly to the newspaper section `author_reading_habits_newspaper_*`

Newspapers

The `author_reading_habits_newspaper_*` items are about newspaper reading habits by the students.

The `author_reading_habits_newspaper` item asks whether and how often students read newspapers. It is prompted as:

> Leggi quotidiani cartacei o online almeno una volta alla settimana?

A logic is applied:

- If not "NO", then students answer to the `author_reading_habits_newspaper_genre_*`items

> Che tipo di articoli leggi prevalentemente sui quotidiani?

The response items are recoded in separate columns that correspond to:

author_reading_habits_newspaper_genre_natpol	Politica nazionale
author_reading_habits_newspaper_genre_internatpol	Politica internazionale
author_reading_habits_newspaper_genre_economy	Economia, finanza

author_reading_habits_newspaper_genre_report	Cronaca
author_reading_habits_newspaper_genre_culture	Culturali
author_reading_habits_newspaper_genre_showbiz	Spettacoli
author_reading_habits_newspaper_genre_localnews	Notizie locali
author_reading_habits_newspaper_genre_latnews	Attualità
author_reading_habits_newspaper_genre_sport	Sport
author_reading_habits_newspaper_genre_tech	Tecnologia, scienze, ambiente
author_reading_habits_newspaper_genre_comments	Approfondimenti sulla cronaca
author_reading_habits_newspaper_genre_editorials	Editoriali
author_reading_habits_newspaper_genre_horoscope	Oroscopo
author_reading_habits_newspaper_genre_post	Rubriche di posta
author_reading_habits_newspaper_genre_other	Altro (ti preghiamo di specificare)

- If no, they go directly to the questions about library services (`author_reading_habits_library_*`)

Library services

The `author_reading_habits_library_*` items are about use of library services by the students.

The `author_reading_habits_library` item asks whether students have been to or used library services in the last year. It is prompted as:

> Negli ultimi 12 mesi sei stato/a in una biblioteca o hai usufruito dei servizi di una biblioteca?

A logic is applied:

- If YES, then students answer to the `author_reading_habits_library_services_*` items, prompted as:

> Per quali delle seguenti attività sei stato/a in biblioteca o hai usufruito dei servizi di una biblioteca? (Sono possibili più risposte)

The response items are recoded in separate columns that correspond to:

author_reading_habits_library_services_catalogues	Per consultare cataloghi
author_reading_habits_library_services_booking	Per prenotare prestiti
author_reading_habits_library_services_loan	Per prendere in prestito libri
author_reading_habits_library_services_audiovisual	Per prendere in prestito materiale audio-visivo (CD-rom, DVD, ecc.)
author_reading_habits_library_services_mlol	Per scaricare e-book, quotidiani, riviste, ecc. (ad es. attraverso la rete di digital lending MLOL)
author_reading_habits_library_services_other	Altro (ti preghiamo di specificare)

- If NO, they fill out the sociolinguistic questionnaire and then answer to the reading habits related to their parents (`parents_reading_habits_*` items).

Reading habits - parents

The `parents_reading_habits_*` items are about the reading habits of the parents.

The `parents_reading_habits` item asks whether the parents of the student read books. It is prompted as:

> I tuoi genitori leggono libri (cartacei o elettronici) o ascoltano audiolibri?

A logic is applied:

- If YES, then students answer the `parents_reading_habits_genre_*` items, prompted as:

> Che genere di libri leggono o ascoltano i tuoi genitori? Nel rispondere riferisciti al genitore che legge di più fra i due.

The response items are recoded in separate columns that correspond to:

parents_reading_habits_genre_novel	Romanzi, racconti, poesia, teatro
parents_reading_habits_genre_romance	Romanzi rosa
parents_reading_habits_genre_thriller	Gialli, noir
parents_reading_habits_genre_scifi	Fantascienza
parents_reading_habits_genre_fantasy	Fantasy, horror
parents_reading_habits_genre_socialsci	Libri di scienze sociali o umane (filosofia, sociologia, politica, psicologia, storia, pedagogia, ecc.)
parents_reading_habits_genre_natsci	Libri di scienze naturali, esatte, applicate, di tecnica
parents_reading_habits_genre_art	Arte
parents_reading_habits_genre_religion	Religione
parents_reading_habits_genre_music	Musica
parents_reading_habits_genre_instant	Libri di attualità (instant book)
parents_reading_habits_genre_humor	Umoristici
parents_reading_habits_genre_hobby	Hobby e tempo libero
parents_reading_habits_genre_astro	Astrologia, magia, esoterismo
parents_reading_habits_genre_tech	Libri di informatica
parents_reading_habits_genre_tourism	Guide turistiche
parents_reading_habits_genre_health	Libri sulla salute
parents_reading_habits_genre_home	Libri per la casa (cucina, bricolage, maglia, cucito, ecc.)
parents_reading_habits_genre_manual	Manuali pratici (ad es. la collana 'For Dummies' [per principianti])
parents_reading_habits_genre_photo	Fotografia, cinema
parents_reading_habits_genre_comics	Libri a fumetti, graphic novel
parents_reading_habits_genre_animals	Libri sugli animali
parents_reading_habits_genre_illustr	Albi illustrati
parents_reading_habits_genre_bio	Biografie
parents_reading_habits_genre_other	Altro (ti preghiamo di specificare)

- If NO, they directly end the survey.

Coherence ratings

Manual coherence ratings were obtained for the texts in the ITACA corpus and recorded in `ratings.csv` (see metadata). In the following we describe the individual items of the coherence rating.

Note that `ratings.csv` can contain various ratings per text. To calculate inter-rater agreement and ensure rating quality, most texts received two ratings. However, some of the texts were rated by only one rater, because one of the raters decided that those texts were too short to be rated. Ten of the texts were rated in a pilot by 6 raters. Finally, one of the texts (CA31TO_LAR14V) has 4 ratings, because it was one of three (next to ST23TO_PER21M and GR25ME_NAR22A) that has been used for the training of the raters.

Questions were administered with the following request:

> Indica, per favore, il tuo grado di accordo o disaccordo con ciascuna affermazione.

The items to rate are recoded as:

likert_001	Il testo è di tipo narrativo o espositivo. [REVERSED ITEM]
likert_002	Il testo rispecchia il genere testuale previsto dalla consegna (lettera).
likert_003	Il registro è generalmente adeguato al tipo di task assegnato (lettera formale a un rappresentante delle istituzioni).
likert_004	La trattazione del tema è aderente al task assegnato (ovvero non c'è confusione tra DAD e DDI).
likert_007	Gli incisi sono di norma correttamente segnalati.
likert_008	Nel testo sono presenti molte virgole splice. [REVERSED ITEM]
likert_005	All'interno del paragrafo la punteggiatura è scarsamente segnalata. [REVERSED ITEM]
likert_006	L'uso del punto nel testo è sempre funzionale alla suddivisione degli enunciati.
likert_009	Il testo è suddiviso in modo incoerente o non è affatto suddiviso in sezioni. [REVERSED ITEM]
likert_010	Il testo è suddiviso in paragrafi coerenti con i cambi di argomento (movimenti testuali).
likert_011	La segmentazione è funzionale all'articolazione del testo.
likert_012	Le devianze nella punteggiatura sono tali da compromettere la coerenza e/o la comprensione del testo. [REVERSED ITEM]
likert_013	Nel testo vengono rispettate le regole di interpunzione.
likert_014	Di norma i connettivi sono impiegati coerentemente con la loro semantica.
likert_015	I connettivi sono spesso impiegati in modo sintatticamente improprio. [REVERSED ITEM]
likert_016	Le devianze nell'uso dei connettivi sono rilevanti. [REVERSED ITEM]
likert_017	Il collegamento tra la ripresa e il suo referente è sempre ricostruibile.
likert_018	I riferimenti (incapsulatori) a contenuti già menzionati nel testo a volte sono difficili da ricostruire. [REVERSED ITEM]

likert_019	Le parti del testo sono per lo più giustapposte. [REVERSED ITEM]
likert_020	L'uso dei connettivi e delle anafore è funzionale alla coesione del testo.
likert_021	Nel testo è riconoscibile un'idea di fondo.
likert_022	Il testo è strutturato sulla base dell'idea di fondo.
likert_023	La tesi è riconoscibile.
likert_024	La tesi è sostenuta da argomenti.
likert_025	Non tutte le argomentazioni sono sviluppate. [REVERSED ITEM]
likert_026	Le conclusioni derivano dagli argomenti presenti nel testo.
likert_027	Le idee presenti nel testo sono ben collegate tra loro.
likert_028	Il testo è costruito a elenco, ovvero è privo di una gerarchia riconoscibile. [REVERSED ITEM]
likert_029	Sono presenti delle devianze nella sequenza logica, tematica o temporale degli argomenti. [REVERSED ITEM]
likert_030	Gli argomenti introdotti nel testo sono ridondanti. [REVERSED ITEM]
likert_031	Il testo è stato progettato in modo efficace.
likert_032	Il contenuto è contraddittorio/incoerente. [REVERSED ITEM]
likert_033	Il contenuto è convincente.
likert_034	Gli argomenti introdotti nel testo sono funzionali allo scopo.
likert_035	Gli argomenti introdotti nel testo vanno fuori tema (off-topic). [REVERSED ITEM]
likert_036	Gli argomenti introdotti nel testo sono generici. [REVERSED ITEM]
likert_037	I termini utilizzati provocano fraintendimenti. [REVERSED ITEM]
likert_038	I termini utilizzati sono coerenti con la loro semantica e il contesto in cui appaiono.
likert_039	Nel testo ci sono informazioni implicite che rendono difficoltosa la comprensione. [REVERSED ITEM]
likert_040	Tutte le informazioni necessarie alla piena comprensione del testo sono ben esplicitate.
likert_041	Il lettore deve sforzarsi per comprendere quanto scritto. [REVERSED ITEM]
likert_042	Sono necessarie diverse riletture per comprendere il testo o alcune sue parti. [REVERSED ITEM]
likert_043	Il testo si comprende facilmente.
likert_044	Lo scopo/gli scopi del testo è/sono chiaro/i.

Ratings were done on a Likert scale recoded as:

"3": "Del tutto d'accordo"
"2": "Un po' d'accordo"
"1": "Un po' in disaccordo"
"0": "Del tutto in disaccordo"

The rating scale for reversed items is recoded as:

"0": "Del tutto d'accordo"
"1": "Un po' d'accordo"
"2": "Un po' in disaccordo"
"3": "Del tutto in disaccordo"

References

Josse J., Husson F. (2016). "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis." Journal of Statistical Software, 70(1), 1–31. doi:10.18637/jss.v070.i01.

Masters, G.N., Wright, B.D. (1997). The Partial Credit Model. In: van der Linden, W.J., Hambleton, R.K. (eds) Handbook of Modern Item Response Theory. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-2691-6_6

PISA 2018 Results (Volume III): What School Life Means for Students' Lives, Annex A1. Construction of indices, https://www.oecd-ilibrary.org//sites/0a428b07-en/index.html?itemId=/content/component/0a428b07-en#s131

Schwitter N. (2023). "ISCO08ConveRsions: Converts ISCO-08 to Job Prestige Scores, ISCO-88 and Job Name", https://CRAN.R-project.org/package=ISCO08ConveRsions

Schutte, N. S., & Malouff, J. M. (2007). Dimensions of Reading Motivation: Development of an Adult Reading Motivation Scale. *Reading Psychology*, *28*(5), 469–489. https://doi.org/10.1080/02702710701568991

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: effects of grade, sex, and ability. *Reading and Writing*, *26*(1), 17–44. https://doi.org/10.1007/s11145-012-9379-2