IJCoL
Italian Journal of Computational Linguistics

9-2 | 2023
Italian Journal of Computational Linguistics vol. 9, n. 2 december 2023

The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German

Aivars Glaznieks, Jennifer-Carmen Frey, Andrea Abel, Lionel Nicolas and Chiara Vettori

Electronic version
URL: https://journals.openedition.org/ijcol/1210
DOI: 10.4000/ijcol.1210
ISSN: 2499-4553

Publisher
Accademia University Press

Electronic reference
Aivars Glaznieks, Jennifer-Carmen Frey, Andrea Abel, Lionel Nicolas and Chiara Vettori, "The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German", *IJCoL* [Online], 9-2 | 2023, Online since 01 March 2024, connection on 11 October 2024. URL: http://journals.openedition.org/ijcol/1210 ; DOI: https://doi.org/10.4000/ijcol.1210

The Kolipsi Corpus Family: Resources for Learner Corpus Research in Italian and German

Aivars Glaznieks*
Eurac Research, Bolzano

Jennifer-Carmen Frey**
Eurac Research, Bolzano

Andrea Abel†
Eurac Research, Bolzano

Lionel Nicolas‡
Eurac Research, Bolzano

Chiara Vettori§
Eurac Research, Bolzano

This article describes the Kolipsi Corpus Family (KCF), a collection of eight related resources for learner corpus research in German and Italian. The KCF supports the study of second language (L2) acquisition of Italian and German in upper secondary schools. It subsumes four L2 corpora with comparable corpus design (with respect to data collection, writing tasks, additional metadata, annotation and processing), portraying two homogeneous learner groups and their learner varieties. The corpora are representative of language learners in the multilingual Italian province of South Tyrol, where both languages are taught daily. The L2 corpora were collected at two different points in time, in 2007 (Kolipsi-1) and 2014 (Kolipsi-2), and all texts were labeled with CEFR levels to allow comparisons of proficiency levels across time. L2 German texts were collected in schools with Italian as the main language of instruction, whereas L2 Italian texts were collected in schools with German as the main language of instruction. Additional resources within the KCF allow researchers to compare the students' language competences in their L2 with the language competences in their first language (L1) in a different task (Kolipsi-Matura) and with similarly aged L1 writers performing the same task (Kolipsi-1-L1). All texts are freely available to the scientific community. Access to the data is granted via an ANNIS search interface and via the Eurac Research CLARIN Repository, from which corpus data can be downloaded in various formats.

* Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy.
E-mail: aivars.glaznieks@eurac.edu
** Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy.
E-mail: JenniferCarmen.Frey@eurac.edu
† Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy. E-mail: andrea.abel@eurac.edu
‡ Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy.
E-mail: lionel.nicholas@eurac.edu
§ Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy.
E-mail: chiara.vettori@eurac.edu

1. Introduction

Learner corpus research (LCR), a young branch of empirical research in linguistics and educational sciences, has been constantly growing over the last 40 years. Over time, more and more learner corpora have emerged as practical resources for investigating learner varieties produced by learners at different ages, at different proficiency levels and with different linguistic backgrounds. However, most of the corpora target English as a second (L2) or foreign language (FL), with significantly fewer resources available for other languages (Granger, Gilquin, and Meunier 2015, p. 1-2). Nevertheless, the number of learner corpora for languages other than English has been recently increasing, allowing researchers and language teachers to study learner varieties on a broader empirical basis and to use this multitude of resources for pedagogical purposes.

This article contributes to this development by introducing the Kolipsi Corpus Family (KCF), a collection of learner corpora for German and Italian L2 writing of upper secondary school students living in the multilingual province of South Tyrol in northern Italy. Italian, as the language of the nation state, and German, the language of 70% of the inhabitants of the province, are the two officially recognized languages of South Tyrol. In addition, Ladin, a Romance minority language, is an official language in two Dolomite valleys where the Ladin language community (4% of the region's inhabitants) mainly resides. While most members of the Italian language group live in urban settlements and in the south of the province, rural areas are mainly inhabited by members of the German language community. To ensure mutual understanding between the language groups while protecting their languages, there are Italian and German monolingual schools where the other language is taught as a second language from the first grade of primary school.

In Section 2 we provide a brief overview of LCR in two non-English contexts, namely Italian and German, and introduce openly available L2 learner corpora for these two languages before comparing these to the KCF resources. We will then provide background information about the two Kolipsi projects, its writers and the data collection method (Sections 3.1-3.4), and will describe the resources of the KCF in more detail in Sections 3.5-3.9. We will conclude by outlining the potential scientific, pedagogical, and didactic applications of the KCF (Section 4).

2. Learner Corpora for German and Italian

2.1 L2 corpora for the study of Italian learner language

Corpus linguistic approaches to L2/FL learner varieties of Italian date back to the late 1980s and have become increasingly widespread in recent years. In their overview about Italian L2 corpora, Giacalone Ramat, Chini, and Andorno (2013) list several spoken, written and mixed corpora documenting the efforts of research in corpus-based linguistics of Italian as L2/FL. The oldest L2 Italian corpus is a collection of spoken L2 Italian known as the Corpus di Pavia/Banca Dati di Italiano L2 (described, e.g., in Andorno and Bernini (2003)), which consists of conversations between L2 learners and an L1 Italian interlocutor. It has been the basis for many research articles on the characteristics of Italian L2 interlanguage varieties of learners with an L1 language background that is typologically distant from Italian, like Arabic, Chinese and Tygrinya (Giacalone Ramat 2003). Since then, other spoken corpora have been created, most notably the LIPS

corpus (Lessico Italiano Parlato di Stranieri) and the VIP corpus (Pallotti et al. 2010).[1] In addition to corpora collected from L2 learners in Italy, there are also corpora that were collected outside of Italy, such as the A.Ma.Dis corpus of conversations between Spanish-speaking FL learners of Italian living in Spain (Cacchione and Borreguero Zuloaga 2018).

Mixed corpora combining spoken and written language data are ADIL2 - Archivio Digitale Italiano L2 (Palermo 2009) and Co.Cer.It (used in Ambroso and Bonvino (2008)). Neither corpus is available online. To the best of our knowledge, the only available mixed corpus is COLI (Corpus of Chinese Learners of Italian), a collection of texts (ca. 83,300 tokens) written by 30 Chinese students of Italian (B1-C1) who finished two different written tasks in 2009-2010 and structured oral interviews (ca. 15 hours).[2]

Among the written corpora listed by Giacalone Ramat, Chini, and Andorno (2013), the well-known VALICO corpus (Corino and Marello 2009) and the Corpus Italiano Scritto L2 (described in Turco and Voghera (2010)) are accessible online.[3] VALICO is a written corpus of 2,500 texts (ca. 567,000 tokens), written by Italian-as-a-foreign-language students from ten different L1 language backgrounds, and collected in different locations around the world (Corino and Marello 2009). Although writer backgrounds were diverse, all texts were based on one of five comic strip inputs eliciting students' narrative text competences. The Corpus Italiano Scritto L2 consists of 152 texts (18,123 tokens) collected from 41 foreign language learners (levels A1-C1) at the Greenwich University of London. Participants had diverse L1s and were all speakers of English as L2. Texts were collected during a language course. They were part of the students' language portfolio and belonged to narrative, descriptive and argumentative text types (Turco and Voghera 2010, p. 147-148).

In recent years, more corpora have been created and released, including CAIL2 (used in Bratánková (2015)), LOCCLI (Siyanova-Chanturia and Spina 2020) and CELI (Spina et al. 2022) of the Università per Stranieri di Perugia,[4] and CORITE - Corpus del Italiano de los Españoles of Spanish L1 speakers (Bailini and Frigerio 2018).[5] CAIL2 (Corpus di Apprendenti di Italiano L2) consists of 400 texts (ca. 237,000 tokens) written by university students with different L1s and was created for a PhD thesis (Bratánková 2015). LOCCLI (Siyanova-Chanturia and Spina 2020) is a longitudinal learner corpus consisting of 350 essays (ca. 97,000 tokens) written by 175 Chinese university students of Italian (levels A1-B1) in 2016 at the Università per Stranieri di Perugia. Data for the LOCCLI corpus was collected at the beginning and at the end of a 6- to 8-month language course. The topics of the essays were predetermined but the learner could choose two out of three. Though small, the corpus is homogeneous with respect to the writers' origin. Similarly, CORITE is also a small learner corpus (ca. 100,000 tokens) of a homogeneous learner group consisting of 45 Spanish FL learners of Italian. It is a collection of informal letters written by university students in Spain. It provides longitudinal data over a period of 7-8 months with 5-25 texts for each learner depending

1 LIPS is available at http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/653-corpus-lips (05.08.2021).
2 COLI is accessible via https://www.unistrapg.it/cqpwebnew/ (registration required, 04.08.2021).
3 VALICO is available at www.valico.org (04.08.2021), the Corpus Italiano Scritto L2 at http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/662-corpus-italiano-scritto-l2 (04.08.2021).
4 All three corpora are accessible via https://www.unistrapg.it/cqpwebnew/ (registration required, 05.10.2022).
5 CORITE and its Spanish counterpart CORSPI are available at https://corespiycorite.altervista.org/ (registration via CATMA required, 06.08.2021).

on their productivity. It provides balanced data of learners of different proficiency levels from A1 to B2 and can be used for cross-sectional and longitudinal analyses on relatively dense data (ca. 80% of the learners wrote 1-2 texts per month) (Bailini and Frigerio 2018). The recently published pseudo-longitudinal CELI corpus (Spina et al. 2022) consists of Italian L2 learners who passed the language certification exams (Certificati di Lingua Italiana – CELI) of the Università per Stranieri di Perugia. The roughly 3,000 texts (ca. 600,000 tokens) are evenly distributed across the CEFR levels B1-C2 and represent a variety of text genres, including articles, blogs, emails, essays, letters, reports and stories.

Finally, the LADDER corpus (Brocca 2021) represents learner language produced in computer-mediated writing contexts and investigates learners' interlanguage at different levels of formality and social distance. It contains a total of about 50,000 tokens of WhatsApp messages and emails.

2.2 L2 corpora for the study of German learner language

As has been the case for Italian learner corpus linguistics, the first German L2/FL corpora also concerned spoken language data. The oldest collections of spoken L2 German can be found in the ESF (European Science Foundation Second Language) database, which lists two corpora containing speech samples of Turkish and Italian immigrants in Germany (Perdue 1993). An example of a more specific corpus is LeaP (Gut 2012), which is a phonologically annotated corpus of German and English FL learners with additional L1 speech samples for comparison. Finally, there is the GeWiss project, which collects academic language of L1 and L2 speakers of German, English, Polish and Italian; it provides audio recordings and transcripts of academic talks and discussions, as well as oral exams supplied by the multilingual GeWiss Corpus (Fandrych, Meißner, and Slavcheva 2012).

The first written learner corpus of advanced L2/FL German to be made freely available for research was FALKO (Lüdeling et al. 2008; Reznicek et al. 2012). It consists of summaries and essays written by FL learners of German, and of reference subcorpora for both text types written by German L1 speakers. While FALKO is relatively small (almost 150,000 tokens in the L2 subcorpus and an additional 70,000 tokens in the L1 subcorpus), it is fully error-annotated within a multilayer corpus architecture. In recent years, FALKO has become a corpus "family" combining several corpora in an ANNIS search interface, including WHiG, an essay corpus of about 130,000 tokens of writers with L1 English from the UK (Krummes and Ensslin 2014); Kobalt-DaF, which combines FL essays of Swedish, Chinese and Belarusian L1 speakers into an L2 corpus of approximately 35,000 tokens (plus an additional 13,000 tokens from L1 texts), and which is more deeply error-annotated than the FALKO essay corpus; KanDeL (a longitudinal corpus with about 75,000 tokens of a homogeneous group of novice learners with American English as L1 (Vyatkina 2016)); the Georgetown longitudinal corpus (almost 80,000 tokens) that provides texts collected from university students at Georgetown University Washington between 2001 and 2004; and CLEG13 (used in Maden-Weinberger (2015)), a collection of more than 700 argumentative essays (ca. 285,000 tokens) written by German FL learners at Lancaster University. In total, the FALKO Family provides around 640,000 tokens of L2/FL German learner data.[6] Further German L2/FL written corpora

6 The FALKO search interface is accessible at https://korpling.german.hu-berlin.de/falko-suche/ (28.04.2023).

worth mentioning are ALeSKo (Zinsmeister and Breckle 2012), which contains 43 essays (ca. 15,000 tokens) written by Chinese L1 learners of German combined with an L1 essay subcorpus for comparison annotated with information about topological fields, DiSKo (Deutsch im Studium: Lernerkorpus (Wisniewski, Muntschik, and Portmann 2022)), for which texts have been collected from standardized university admission language tests of students with various L1s, and deL1L2IM (Höhn 2015), which, to our knowledge, is the only social media learner corpus of German texts.

2.3 Multilingual corpora containing German and/or Italian learner language

In addition to Italian- or German-only written learner corpora, one can also find corpus collections with mixed target languages where Italian L2 texts or German L2 texts are substantially represented. The MERLIN corpus (Boyd et al. 2014), for example, as well as providing German and a Czech subcorpus, also contains Italian texts collected from adult foreign language learners with different L1 backgrounds and residing in different places in Europe. The Italian part consists of 816 texts (ca. 92,000 words) of A1-B2 proficiency levels, whereas the German part consists of 1,035 texts (ca. 126,000 words) of A1-C2 proficiency levels (Boyd et al. 2014, p. 1283). LEONIDE (Glaznieks et al. 2022), on the other hand, consists of Italian, English and German learner texts written by lower secondary school students in South Tyrol. It is longitudinal as it contains texts of two different text types written by the same students over a period of three years. The Italian part (844 texts, ca. 93,000 tokens) was produced by 78 L1 (408 texts, ca. 51,100 tokens) and 84 L2 writers (436 texts, ca. 41,600 tokens) of Italian, the German part (833 texts, ca. 74,000 tokens) by 63 L1 (351 texts, ca. 41,200 tokens) and 98 L2 writers (483 texts, ca. 32,600 tokens). The corpus can be used to analyze and compare learners' progression of L2 Italian and German to L1 writings of students of the same age. The SWIKO corpus (used in Karges, Studer, and Wiedenkeller (2019)) is similar to LEONIDE but consists of texts produced by young Swiss learners with German or French as their L1, and provides texts in German, French, and English written by the same writer. The TRAWL corpus (Dirdal et al. 2022) combines English, French, German and Spanish texts of L1 Norwegian learners,[7] and the multilingual learner corpus MLC (Tagnin 2006) consists of English, Spanish, German and Italian learner texts produced by Brazilian learners in Brazil. The texts of these multilingual corpora were collected in teaching institutions and schools. SWIKO and MLC are yet to be made available online.

Finally, the Uppsala WordReference Corpus (Berdičevskis 2020) represents learner language in English, Spanish, French and Italian collected from the WordReference Language online Forums[8] and can thus be used as another example of Italian L2 writing in social media with comparable data from other languages. Table 1 summarizes relevant written L2 corpora that contain Italian and German learner texts.

3. Kolipsi – A family of resources for learner corpus research in German and Italian

The KCF complements the existing learner corpora for Italian and German by providing homogeneous L2 corpora with respect to the designed tasks and the L1 languages of the learners. It provides two text types (letter and email) and genres (argumentative and narrative) for each learner. Compared to most available corpora, where informants

7 Access to the corpus can be requested at https://tekstlab.uio.no/trawl/ (28.04.2023).
8 The forums are accessible at https://forum.wordreference.com/ (28.04.2023).

Table 1
Overview of written learner corpora for L2 German and Italian.

text language	*corpus*	*L1*	*size*	*study context*
L2 Italian				
	VALICO	various	382k	university
	Corpus Italiano Scritto L2	en, other	23k	university
	CAIL2	various	237k	university
	LOCCLI	zh	97k	university
	CELI	various	608k	official lang. testing
	CORITE	spanish	103k	official lang. testing
	LADDER	de	50k	university
L2 German				
	FALKO	various	145k	university
	WHiG	en	130k	university
	Kobalt-DaF	ru, sv, zh	33k	school, university
	KanDeL	en	74k	university
	CLEG13	en	320k	university
	ALeSKo	zh	14k	university
	DiSKo	various	397k	university
	deL1L2IM	de, ru-be	50k	various
Multiple L2				
cz, de, it	MERLIN	various	340K	official lang. testing
de, en, it	LEONIDE	de, it, other	240k	lower sec. school
de, en, es, fr	TRAWL	no	undefined	primary and lower sec. school
en, es, fr, it	Uppsala WordReference Corpus	various	170M	Informal language learning online

were mainly university students, the texts in the KCF were collected in upper secondary schools in the multilingual Italian province of South Tyrol on two separate occasions. The first data collection took place in 2007 (Kolipsi-1) and was repeated in 2014 with students of the same age (Kolipsi-2), thus facilitating comparisons of the two resulting corpora. As reference data, the KCF provides two kinds of corpora with texts written by L1 writers of German and Italian, respectively. For the first reference corpus (Kolipsi-1_L1), students of the same age repeated the same tasks of the first edition of the L2 data collection in their L1. The second reference corpus (Kolipsi-Matura) is composed of final school examinations written by a sample of participants of Kolipsi-1 in 2009, enabling parallel investigations into a writer's L1 and L2 writing competences.

In the following sections we present the details of the diverse corpora in the KCF. Sections 3.1-3.4 outline the origin of the texts and the purpose of their collection.

Section 3.5 details the resources available in the KCF and their composition in terms of languages and author backgrounds. Sections 3.6-3.7 describe the corpus creation process from the initial transcription of the handwritten student essays to the manual and automatic annotation and conversion of the data. The available metadata, including text evaluations and sociodemographic data on the students represented in the corpora, is described in Section 3.8. Finally, we conclude this resource description with notes on corpus access and availability (Section 3.9), and a discussion on the relevance and possible uses of the resources (Section 4).

3.1 Background and Scopes of the Kolipsi Corpus Family

The Kolipsi corpora are a key outcome of the two related research projects Kolipsi I (Abel, Vettori, and Wisniewksi 2012) and Kolipsi II (Vettori and Abel 2017), which sought to analyze and describe the L2 competences of South Tyrolean students from upper secondary schools. Making use of the Common European Framework of Reference for Languages (CEFR (Council of Europe 2001)), the projects collected and labelled German and Italian L2 texts from upper secondary school students and related them to sociopsychological metadata, thereby identifying linguistic and extra-linguistic factors that influence the students' L2 competence. The first Kolipsi project ran between 2007 and 2009; the study was repeated between 2014 and 2017.

3.2 Data Collection

A statistically representative number of students from almost all upper secondary schools in South Tyrol participated in both project editions. They took a written test and filled out a questionnaire that collected information about, among other things, their language biography and habits, as well as their motivation towards L2 learning, attitudes towards speakers of the other language group, etc. Parents and L2 teachers were also involved in the studies and provided additional information to contextualize the linguistic data collected by means of a questionnaire. The sample consisted of students attending the penultimate year of upper secondary school (average age of 17). Due to the large number of participants to be tested, a cluster sampling was chosen, whereby clusters corresponded to the classes present in each school, stratified proportionally by school type (grammar school, i.e. *liceo/Gymnasium* vs. technical school, i.e. *istituto tecnico/Fachoberschule*) and location (Bolzano vs. other municipalities). During the 2007-2008 school year we collected a total of 1,275 language tests (279 written in German L2 and 996 in Italian L2) and 1,470 questionnaires. During the 2014-2015 school year we collected 1,580 language tests (421 written in German L2 and 1,159 in Italian L2) and 1,692 questionnaires. For purposes of comparison, additional data was collected in 2009 from a subset of 152 of the participants of the first edition, gathering texts from the official graduation exams (*maturità/Matura*) written in the students' first language. Furthermore, a separate data collection of L1 reference data, which took place outside the two projects in 2010, asked around 450 comparably aged L1 students from schools in Germany and Italy to perform the same tasks as those used in the Kolipsi projects.

3.3 Methodology: Writing tasks and CEFR level annotation

In order to assess South Tyrolean students' second language skills in the real world, we relied on Bachman and Palmer's model (Bachman and Palmer 1996) and on the very similar action-oriented approach of the CEFR (Council of Europe 2001), where good

language proficiency refers to the students' ability to actively participate in community life in the respective L2.[9] The language skills of the sample were tested by means of a series of standardized instruments with direct reference to the CEFR scales (Council of Europe 2001) and particular focus on the central B1 and B2 levels, considering that the target level at the end of upper secondary school is B2. Among these standardized instruments were two types of texts that the students had to write, one narrative and one argumentative.[10] In the first task, students were provided with a picture story without an ending (for a total of four illustrations), which they were asked to put into words and invent an ending to in an email to a friend (ca. 130 words). Unlike the narrative task, the format of the argumentative writing task was not identical for the two project editions. In this case, in Kolipsi I, students had to write a letter to a friend (ca. 150 words) to organize a holiday together, proposing two different destinations and describing the advantages and disadvantages of either option also on the basis of their own personal experience. In Kolipsi II, students had to write an email in response to a letter published in a teen magazine about problems that can arise in chat communication, expressing their opinion, reporting personal experience and giving advice. The Herder Institute of the University of Leipzig (Germany) provided test 1 and 2 for the first project,[11] while test 2 of the second project edition came from TELC.[12] The written instructions for both tasks were available in identical form in German and Italian: the formulations were clear, concise and comprehensible also for students with lower levels of competence.[13] The pen and paper writing time available for each text was 30 minutes and no mono- or bilingual dictionaries were allowed.

The main expectation of the narrative task was a description of the events depicted and the provision of a possible outcome for the story. Given the interactive nature of the text, students were also expected to comment on and explain the events that occurred to the main character. The spectrum of skills required to complete the task was that of the CEFR B1 level. In the second case, the writer was expected to express opinions, compare, argue, convince and suggest, so the spectrum of skills required was that of a CEFR B2 level. The texts were assessed with the help of grids specifically developed from the CEFR descriptors and adapted for the target group (Alderson 1991; North 2000), reflecting the levels of competence, from A1 to C1/2, as closely as possible (Council of Europe 2001, Table 3). The results were evaluated by a team of specially trained raters (three raters for German L2, six raters for Italian L2) and submitted to the Many-Facet Rasch Measurement Model (MFRM, (Linacre 1989)), which allows for the independent parameterization of subjects, items, and other facets relevant to the research for a fair evaluation.[14]

9 Cf. the definition of "proficiency assessment" provided by the CEFR (Council of Europe 2001, p. 183).

10 In Kolipsi I, the students took an additional written test that was part of the official bilingualism exam of the Province of Bolzano. In Kolipsi II, they took an additional listening comprehension and lexical competence test (Dialang); only Italian-speaking students took a C-test (Test-DaF). In both Kolipsi I and II, a subsample of students also took an oral test.

11 The Herder Institute is engaged in scientific research on the development, analysis and optimization of language tests.

12 TELC (The European Language Certificates) is an international provider of language certificates based on international standards in more than 30 countries.

13 All task instructions are provided at https://www.porta.eurac.edu/lci/kolipsi-family/ (28.04.2023).

14 All measures taken for quality management, such as inter-/intra-rater agreement measurements and the measurement of rater effects, are described in detail in the report of the Kolipsi I project (Abel, Vettori, and Wisniewksi 2012, p. 38-55).

3.4 Main outcomes of the Kolipsi projects

For historical-political reasons, the school system in the multilingual province of South Tyrol is divided to guarantee instruction in German, Italian and, in some valleys, Ladin.[15] German or Italian are, however, compulsory as an L2 from the first grade of primary school up to the last year of secondary school in schools where Italian or German are the languages of instruction. Second language teaching occupies approximately 2,000 hours or more if the school applies other measures to promote L2 acquisition. This separation between language groups, which is also evident in many other areas of daily life (Pallaver 2017) and reinforced by the uneven distribution of the German and Italian language groups throughout the territory,[16] does not encourage encounters between the groups, let alone exposure to and practice in the L2. This leads to unsatisfactory L2 competences, which are not only perceived in society and proclaimed by the media (Laner 2007; Marchiodi 2011) but were also confirmed by the two Kolipsi projects in 2007-2009 (Abel, Vettori, and Wisniewksi 2012) and in 2014-2017 (Vettori and Abel 2017).

As far as German as L2 is concerned, the level of competence revealed by the second edition of the project was inferior to that of the first edition. In Kolipsi I, 46.7% of the students showed B1-level competences and 28.1% A2 level competences. Only 13% of the sample showed B2-level competences. In Kolipsi II, the sample showed the following competence levels: 34.3% B1, 37.6% A2 and 12.9% B2. A deterioration of competences was also observed in the Italian L2 tasks, with the second edition of the project ranking lower than the first. In Kolipsi I, in fact, 44.2% of students showed B1-level competences, 40.3% of the sample showed B2-level competences and only 4.1% A2-level skills. In Kolipsi II, on the other hand, most of the sample exhibited B1-level competences (51%), 23.3% A2-level competences and only 19.5% demonstrated B2-level competences. The data thus reveal that for many students active and daily participation in L2 is difficult due to a language competence still at elementary level (A2), and that it is only a minority of students who is able to "produce clear, detailed text(s) on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options" (Council of Europe 2001, p. 24). Figure 1 shows the CEFR distribution for Kolipsi-1 and Kolipsi-2.

With regard to the sociopsychological aspects of the study, a number of factors emerging from Kolipsi I, which more or less directly influenced L2 skills, were also found and further explored in Kolipsi II. Although school is an important location for L2 learning and practice in South Tyrol, it is what happens outside the school walls that most influences both the attitudes and the habits of the students and consequently their achievements. Kolipsi I data showed that having contact and friendships with members of the other language group (out-group) improves students' attitudes towards that group and increases their motivation to learn the L2. The regression analysis we carried out in Kolipsi II, which calculates the estimated relationship between a dependent variable – in this case, L2 skills – and one or more explanatory variables, also highlighted the extent to which the private use of L2 is decisive in acquiring good language skills. In both German and Italian learners, practicing L2 outside of school "predicts" better

15 For a better understanding of the current educational situation in South Tyrol, see Baur, Mezzalira, and Pichler (2008).

16 The territorial distribution of the groups is extremely uneven: more than 80% of the Italian-speaking population is concentrated in major towns (>15,000 inhabitants), while the German-speaking group resides mostly in rural municipalities and only a quarter of all German speakers are present in the major towns.

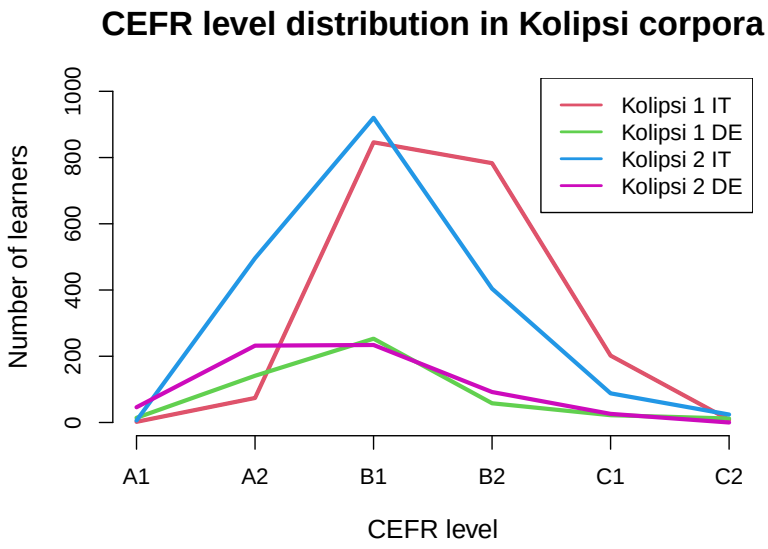

Figure 1
CEFR level distribution in Kolipsi-1 and Kolipsi-2.

language skills. In the case of Italian-speaking students, a good command of the South Tyrolean dialect – a language not taught in school but acquired through contact with native speakers – also predicts better language skills; as for German-speaking students, it is having frequent contact with the "closest" person in the out-group and being comfortable speaking the L2 that predict better skills. Moreover, for both student groups, attending high school (*liceo/Gymnasium*), as opposed to a technical institute (*istituto tecnico/Fachoberschule*), is a predictor of better skills.[17]

3.5 Resources in the Kolipsi Corpus Family

The KCF combines eight different corpus resources, with a total of 4,059 L2 learner texts in Italian, 1,223 L2 learner texts in German and 443 reference texts of L1 writers of the same age. In addition, the KCF provides 152 texts from the final school exams written by students who participated in the first edition of the project and whose L2 texts are available in the Kolipsi-1 corpus. Table 2 lists all KCF resources with their respective year of creation and size in terms of number of participating students and number of produced texts.

As described in Section 3.3, each corpus contains two different text types: an argumentative text, written as a letter (opinion text), and a narrative text based on a picture story and written as an email (picture story). Since each student was asked to write both text types, all corpora contain about the same amount of opinion texts and picture stories with only small deviations (+/- 2,5%) owing to a few students not

17 For detailed results and statistical analysis see Abel, Vettori, and Wisniewksi (2012) and Vettori and Abel (2017).

Table 2
Resources included in the Kolipsi Corpus Family.

corpus	*year*	*#students*	*#texts*	*#tokens*
L2 data				
Kolipsi-1_L2_IT	2007	1,000	1,990	387k
Kolipsi-1_L2_DE	2007	267	523	87k
Kolipsi-2_L2_IT	2014	1,035	2,063	400k
Kolipsi-2_L2_DE	2014	357	700	106k
Reference data				
Kolipsi-1_L1_IT	2010	43	80	11k
Kolipsi-1_L1_DE	2010	183	363	80k
Kolipsi-Matura_L1_IT	2009	53	53	41k
Kolipsi-Matura_L1_DE	2009	99	99	64k

submitting both texts. The two Kolipsi-Matura subcorpora are an exception with respect to available text types, as they provide only one type, which was individually chosen by each student for the final school exam.[18]

The average text length between the two text types and the two L2 corpus editions was comparable with a mean ranging between 185 and 195 (median 186-190), while standard deviation ranged between 50 and 64 depending on the resource.[19] However, students who attended schools with German as the main language of instruction wrote significantly longer texts than those instructed in Italian schools (Mann-Whitney U test for non-normally distributed data, p-value: 2.2e-16). This difference can also be observed in Figure 2, which illustrates that all corpora collected in German schools, i.e. the German L1 corpus and the Italian L2 corpora, show distributions with higher numbers of tokens per text. Owing to the different task requirements, texts written for the final school exams were three to four-times longer than the L2 texts, with average text lengths between 660 (Kolipsi-Matura DE) and 822 tokens (Kolipsi-Matura IT).

All L2 corpora were collected in South Tyrolean schools and contain extensive metadata on the authors of the texts (see Section 3.8). The majority of the students were between 17 and 18 years old when the texts were produced and the gender balance is skewed towards female representation (56%). Furthermore, about two-thirds of the students in the sample resided in rural and thus more German-centered areas, whereas roughly one-third lived in urban, more multilingual districts. Most of the German L2 writers were Italian native speakers (87%) and vice versa (85%). A smaller number of students identified as bi- or plurilingual (combining Italian with German or Italian or German with other languages, in total about 11%) or as having other language backgrounds, such as Ladin (approx. 3% of the total number of L2 texts), Albanian, Arabic or Spanish (approx. 3%). German-Italian bilingual students contributed to the L2 corpora with texts written in the language which was not the main language of schooling. The L1 reference data was sampled to contain only L1 speakers, the vast

18 The official examination sheets are provided at https://www.porta.eurac.edu/lci/kolipsi-family/ (28.04.2023).

19 There was no significant difference between the text length of the two text types (opinion text vs. picture story), judging from a non-parametric Mann-Whitney U test at an alpha-level of 0.05.

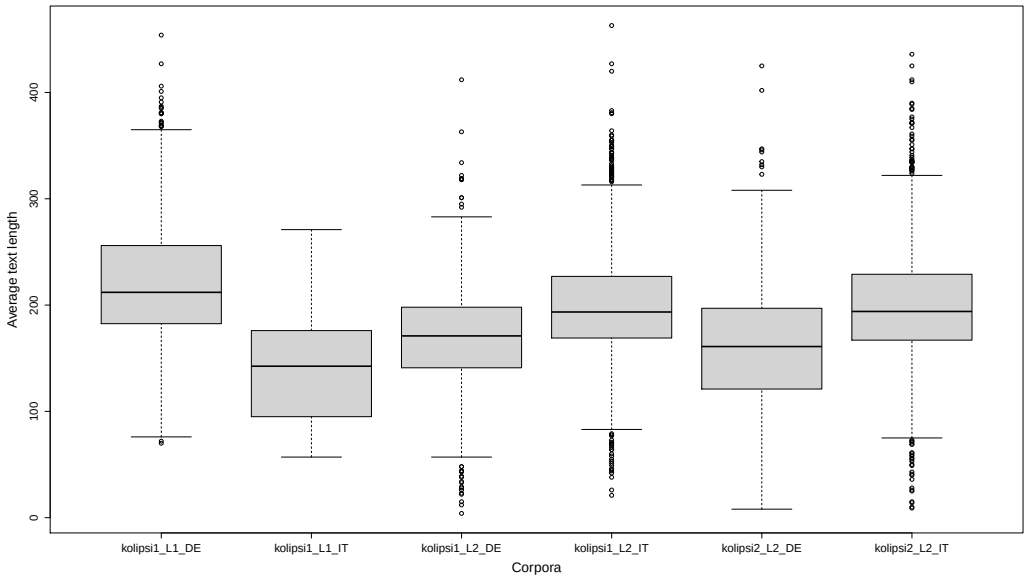

Figure 2
Average text length in KCF L1 and L2 corpora.

majority of whom grew up with only one L1. Only three out of the L1 speakers grew up with another language alongside their L1. All three of them, however, were raised in Germany.

3.6 Transcription and manual annotations

All KCF corpora are based on handwritten essays, which were scanned and then manually transcribed using the XMLmind editor.[20] We equipped the editor with a style sheet that allowed us to make manual annotations while transcribing. In addition, the editor includes a user-friendly styled visualization of the XML structure, which was helpful for annotators who were not familiar with XML. Transcribers underwent training before the transcription task, with a second training and discussion phase after their first few annotations. Reconciliation meetings were held throughout the transcription and annotation process. A varying number of transcribers employed for each corpus ranged from 1 (Kolipsi-1_L1 corpora and Kolipsi-2_L2_DE) to 5 (Kolipsi-1_L2_IT). Since there was no overlap between transcribers in the initial transcription phase, we opted for a subsequent evaluation of the transcription quality.

For each corpus in the Kolipsi family, we performed an evaluation of the transcription quality on a random subset of approximately 10% of the transcriptions.[21] For this, the transcriptions of the random subsets were checked by a second annotator and corrected whenever the transcription deviated from the original or did not adhere to

20 A full description of the editor is available at http://www.xmlmind.com/xmleditor/ (28.04.2023).
21 We regard as a transcription one XML file containing all texts written by a student (i.e., two texts, with a few exceptions) encoded according to our custom-made Kolipsi XML schema.

the transcription guidelines used during the transcription process.[22] We then compared the original and corrected transcript versions, and extracted deviations between these (as well as between XML annotations) using the SequenceMatcher of the Python *difflib* library. For the transcription accuracy calculation we used Word Error Rate as a measure of estimating errors while taking into account potentially changing text lengths between original and corrected transcripts, e.g., for cases where the original transcripts missed out words that should have been transcribed. The Word Error Rate measure, which is typically used in evaluations of aligned versions of texts, such as speech-to-text or machine translation systems, was calculated by converting deviations found in the transcripts into token-level counts of substitutions, insertions and deletions looking at both capitalization and punctuation. All token substitutions, deletions and insertions were then summed up and divided by the total number of tokens in the final, correct version of the transcript. We used this measure to estimate the overall transcription quality and to investigate error rates for different languages, transcribers and text types.

Our analysis showed that the overall transcription quality of the corpora was rather high, with a word accuracy ranging between 98.4% and 99.9%. The most common transcription errors were typos in the handwritten learner text that were not carried over to the XMLmind editor. Occasionally, it so happened that transcriptions missed out entire lines. Table 3 presents the number of texts and transcribers in each evaluation sample and shows the results of the transcription error analysis. When comparing the subcorpora, we noticed that Italian language transcriptions were less accurate than the German transcriptions and that the word error rates of some of the transcribers were visibly higher.[23] We thus verified all transcriptions reporting word error rates above 0.01 before publishing the final version of the corpus.[24]

Table 3
Evaluation of the transcription quality for each subcorpus in the KCF.

corpus	word error rate	transcripts (transcribers)	missing words	superfluous words	substituted words
Kolipsi-1_L1_DE	0.001	40 (1)	8	2	13
Kolipsi-1_L1_IT	0.016	40 (1)	151	0	24
Kolipsi-1_L2_DE	0.008	59 (3)	46	2	94
Kolipsi-1_L2_IT	0.003	85 (5)	61	5	41
Kolipsi-2_L2_DE	0.006	88 (1)	19	1	50
Kolipsi-2_L2_IT	0.010	240 (2)	132	45	341

The annotations used in the transcription process were developed to preserve most of the features of the handwritten original text, including (1) the structure of the text,

22 The transcription guidelines are available at https://www.porta.eurac.edu/lci/kolipsi-family/ (28.04.2023).
23 Indeed, both variables, the text language and the transcribers significantly affected the word error rate in a linear model ($p < 0.01$, F-Test).
24 The high number of missing words in the Kolipsi L1-DE corpus evaluation sample is mainly due to one essay transcription, which left out two full paragraphs.

(2) orthography, (3) the choice of linguistic means, (4) handwriting legibility, (5) self-correction and (6) the use of stylistic means.[25]

1. Annotated structural features are **paragraphs**, **footnotes**, **greetings** and **closings** in letter-like texts. This type of annotation enables studies on differences in structuring a text in meaningful units, which is a major challenge in text production.
2. With regard to orthography, misspelled words were annotated as **orthographic errors**. This annotation tag requires an orthographic correct spelling of the word to be added to the tag (i.e., a target hypothesis), which can then be used for any further processing, such as part-of-speech and lemma tagging. **Hyphens**, which indicate word division at the end of a line, were also annotated because hyphenation at line breaks is regulated in both German and Italian orthography. Therefore, this annotation can be used to investigate the correctness of hyphenations and to distinguish them from other types of hyphen usages in, for instance, compounds.
3. Annotations concerning the choice of linguistic means are **foreign words**, i.e., words that do not belong to the target language, and **variant groups**, i.e., indecisive word use of two or more variants (e.g., the use of both *child* and *kid* instead of one or the other).
4. Sometimes, handwritten texts are difficult to transcribe due to illegible handwriting. For this reason, we conceived two annotation tags to help the annotators: if words or part of words where not readable, annotators could use an **unreadable** tag; alternatively, if they could not decide between two or more potential readings of a word, they could use the **ambiguous** tag.
5. Two tags were created to distinguish between different kinds of self-correction, namely **deletion** and **insertion** of letters and words. Generally, deletions and insertions were annotated within a **correction** annotation.
6. Several annotations were created to annotate the different stylistic features that students used to express pragmatic or discursive meaning: firstly, fully capitalized words were annotated with a **capitalization** tag, which added a standard spelling to each word to be able to distinguish between lower- and upper-case letters. Moreover, **emoticons** (i.e., combinations of punctuation signs, letters and numbers to graphically represent facial expressions), all kinds of **emphases** (e.g., bold or underlined words), the use of **images** (i.e., drawings within the text) and **symbols** (i.e., icons with a symbolic meaning, e.g., arrows or hearts) were annotated separately. Uncommon abbreviations and word **reductions**, which cannot be found in standard dictionaries (e.g., *Ita* for *Italian*), were annotated adding a non-abbreviated word form, which again can be used for further processing.

25 The annotation schema was also used and adapted for other learner corpora, e.g., MERLIN (Boyd et al. 2014), KoKo (Abel et al. 2014) and LEONIDE (Glaznieks et al. 2022).

Table 4 shows a list of annotations and their occurrence in the KCF.[26]

Table 4
Frequency of manual annotations used in the KCF.

	frequency in			
annotation	*Kolipsi-1_L2*	*Kolipsi-2_L2*	*Kolipsi-1_L1*	*Kolipsi-Mat*
1. structure of the text				
closing	1,820 (0)	2,096 (0)	384 (0)	0 (0)
greeting	2,476 (0)	2,637 (0)	429 (0)	0 (0)
paragraph	13,924 (12,977)	18,406 (15,189)	2,803 (2,682)	1,865 (1,355)
2. orthography				
hyphen	292 (0)	170 (0)	296 (0)	232 (0)
orthographic error	13,241 (13,233)	13,871 (13,849)	1,454 (1,452)	439 (431)
3. choice of linguistic means				
foreign word	612 (609)	613 (612)	10 (10)	0 (0)
variant group	10 (0)	27 (0)	0 (0)	0 (0)
4. legibility of handwriting				
ambiguous	1,080 (1,003)	81 (57)	23 (21)	3
unreadable	6,003 (598)	6,080 (315)	570 (25)	358 (59)
5. self-correction				
correction	11,642 (0)	19,477 (0)	1,743 (0)	1,479 (0)
deletion	8,982 (0)	17,842 (0)	1,390 (0)	1,109 (0)
insertion	3,293 (0)	5,609 (0)	357 (0)	556 (0)
6. use of stylistic means				
overcapitalization	318 (318)	122 (122)	1,442 (1,422)	39 (11)
emphasis	186 (186)	67 (67)	18 (18)	26 (16)
image	2 (0)	16 (0)	0 (0)	0 (0)
reduction	98 (98)	23 (23)	62 (62)	79 (79)
symbol	25 (25)	19 (18)	12 (12)	0 (0)

3.7 Conversion and automatic annotation of linguistic features

We used the Salt and Pepper framework to convert the XML files containing the transcriptions and manual annotations into various other, further enriched, file formats.[27] For this purpose, we created several reusable Pepper modules that would perform individual conversion and annotation tasks, such as transferring the custom XML format used in the transcription and annotation phase of the project to an intermediate conversion format (Salt), from which it can then be further transferred to other relevant file formats, or adding automatic linguistic annotation using an NLP pipeline relying on well-known NLP frameworks. All texts were automatically annotated for sentence and token boundaries and part-of-speech (POS) tags (UD-POS tagging)[28] with the Stanford

26 The numbers in brackets refer to the searchable corpora https://commul.eurac.edu/annis/kolipsi (28.04.2023).
27 Both Salt and Pepper can be downloaded at https://corpus-tools.org/home/ (28.04.2023).
28 The full list of universal POS tags is available at https://universaldependencies.org/u/pos/ (28.04.2023).

CoreNLP pipelines for Italian and German (Manning et al. 2014),[29] and for lemma and language-specific POS tags with the Italian and German versions of TreeTagger (Schmid 1999).[30]

3.8 Metadata

Metadata in the KCF refer either to the students and describe demographic, linguistic or educative features related to each person who provided texts for the corpus collection, or to the texts they wrote. Person-related metadata is identical for all texts produced by the same student, while text-related metadata may change from text to text. The chosen metadata items follow the unified metadata scheme for learner corpora (Paquot et al. 2023).[31] Tables 5 - 9 summarize all metadata annotations provided in the KCF.[32]

3.8.1 Person-related metadata

The person-related metadata included in the KCF was collected through a questionnaire that asked students for their sociodemographic information and sociopsychological attitudes towards the second language. Additionally, the corpus provides information about the students' school, class and teacher.

For each student, the KCF provides basic demographic information (Table 5). Each student ("author") is identified by an anonymous student code (**author_id**) and a unique identifier of their class (**author_class_id**), allowing us to monitor for class effects in the hierarchically-structured data while maintaining anonymity. The **author_gender** item distinguishes female (f) from male (m) authors. Additionally, **author_environment** specifies whether the students' local living environment is urban or rural.[33] The socioeconomic background of the students was also surveyed (**author_socioeconomic_status**).[34]

All metadata describing the language background of the students (Table 6) focused on the official languages of the province – German (DE), Italian (IT) and Ladin (LAD) – and offered also a DE-IT bilingual option. Other languages were grouped into one single category (OTHER) to maintain the anonymity of this smaller group. Metadata regarding the learner's L1 (**author_L1**) is based on what authors perceive as their first language(s). To know more about the family languages, the students were also asked to indicate the first language(s) of their mother (**author_mother_L1**) and father (**author_father_L1**). Furthermore, the students were asked to indicate their sense of belonging to categories overlapping with the recognized language groups in

29 CoreNLP is available at https://stanfordnlp.github.io/CoreNLP/ (28.04.2023).

30 The TreeTagger is available at https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (28.04.2023).

31 The full scheme can be downloaded at https://doi.org/10.14428/DVN/4CDX3P.

32 For L1 corpora a restricted set of metadata was collected, ensuring comparability of the data without focusing on the sociopsychological aspects of the L2 study.

33 Urban environments include those towns with city charters and a population of over 10,000 inhabitants, i.e., Bolzano, Bressanone, Brunico, Laives, and Merano. All other towns are considered as rural environments.

34 For indicating the socioeconomic status of three levels for the Kolipsi I study, an index was built taking into consideration four survey items: 1) the number of cars the family owns, 2) the number of holidays taken in the last 12 months, 3) the availability of an own room at home, 4) the number of computers at home. Instead, for the Kolipsi II study an index was built taking into consideration slightly different aspects, namely 1) a group of items referring to the home conditions that support learning (e.g., internet connection, PC), 2) the material prosperity of the family (e.g., number of cars, TVs) and 3) the number of books at home (Campodifiori et al. 2010).

Table 5
Person-related metadata annotations: basic demographics.

annotation	*explanation*	*values*
author_id	unique identifier of the student	combination of letters/numbers
author_class_id	unique identifier of the class	number
author_gender	gender of the student	f (female) m (male)
author_environment	characteristics of the student's residence	urban rural
author_socioeconomic_status	socioeconomic status of the student's family	low medium high

South Tyrol (**author_language_group_affiliation**), that is, the German-Italian bilingual group, which de facto exists but is not officially recognized, and the "OTHER" group.[35] In general, a broad correspondence was observed between the indicated language group affiliation and the first language for those students who identified as belonging to the German or the Italian language group. Finally, another category of language background-related metadata took into consideration the language environment in which the students live (**author_language_environment**). For this purpose, the places of residence of the students are grouped according to the proportion of the three official language groups living there, i.e., more than 70% of the inhabitants belong to the German language group (mainly German (> 70%)), more than 70% of the inhabitants belong to the Italian language group (mainly Italian (> 70%)), between 30-70% of the inhabitants belong to the German or Italian language group (mixed (30-70%)), or the majority of the inhabitants belong to the Ladin language group (mainly Ladin).[36]

Additional person-related metadata concerned the school the student attended (Table 7). One metadata field related to the language of schooling (**school_language**), so either German (DE) or Italian (IT).[37] Another referred to the school type (**school_type**), distinguishing between grammar schools and technical schools, both of which award a diploma for university admission.[38] In Kolipsi-1, the metadata also indicated the places where the schools are located (**school_location**), which mainly correspond to some of the larger towns in South Tyrol (Bozen-Bolzano, Brixen-Bressanone, Meran-Merano,

35 The declaration of linguistic affiliation is the basis of the "ethnic proportional" representation system, the legal regime that in South Tyrol governs admission to public employment and the enjoyment of certain rights, in particular the allocation of social housing, so as to guarantee a proportional allocation to the three officially recognized language groups: German, Italian and Ladin.

36 These data come from a pre-2018 census carried out by the provincial statistics office ASTAT every 10 years (since then, data are collected by ASTAT on an annual basis). In this context, the language group affiliation declaration is also recorded and linked to other variables, such as the place of residence. The data are publicly available at https://astat.provinz.bz.it/ (28.04.2023).

37 The Kolipsi projects disregarded the few high schools in the Ladin valleys that have their own regulation for language use.

38 *Maturità* in Italian, *Matura* in German.

Table 6
Person-related metadata annotations: the students' language background.

annotation	*explanation*	*values*
author_L1	L1 of the student	DE IT LAD DE-IT OTHER
author_mother_L1	L1 of the student's mother	
author_father_L1	L1 of the student's father	
author_language_group_affiliation	sense of belonging to one of the recognized language groups	
author_language_environment	language that is mainly used at the student's residence	mainly German mainly Italian mainly Ladin mixed

Sterzing-Vipiteno). An anonymous identifier for the school (**school_id**) and the teacher (**teacher_id**) was given whenever there was more than one class per school or more than one class per teacher in the corpus so as to watch for group effects in the hierarchical data.

Table 7
Person-related metadata annotations: the students' schools.

annotation	*explanation*	*values*
school_language	language of instruction	DE, IT
school_type	type of upper secondary school	grammar_school technical_school
school_location	place of the school	Bozen-Bolzano Brixen-Bressanone Meran-Merano Sterzing-Vipiteno Italy (Kolipsi_L1) Germany (Kolipsi_L1)
school_id	unique identifier of each school	combination of letters/numbers
teacher_id	unique identifier of each school	number

To analyze the students' L2 proficiency, L2 corpora were given CEFR level indications for the students (as described in Section 3.3), as well as information on additional language tests and proficiency indicators (Table 8). A CEFR level indication for the students based on their performance in both writing tasks was assigned manually, with

potential values spanning all levels, from A1 up to C2 (**author_proficiency_level**).[39] Furthermore, the students were asked to indicate their final school grade in the L2 for the previous school year (**author_L2_school_grade**). This value ranges from 4 (the worst grade) to 10 (the best grade). For Kolipsi-1, an additional language test based on the local bilinguality exams (**author_bilinguality exam**) was given,[40] while for Kolipsi-2 the additional language test was based on the Dialang test (**author_dialang test**).[41] Additionally, Italian speakers in Kolipsi-2 were asked to indicate their dialect competences in the L2 German (**author_L2_dialect_competence**).

Table 8
Person-related metadata annotations: competence-related metadata.

annotation	*explanation*	*values*
author_proficiency_level	CEFR level assignment of the student	A1-C2
author_L2_school_grade	school grade in L2 German or L2 Italian	4-10
author_bilinguality_exam	result of the exam	failed passed unclear
author_dialang_test	result of the exam	A1-C2
author_L2_dialect_competence	estimation of dialect competence	A1-C2 NONE

3.8.2 Text-related metadata

Text-related metadata (Table 9) contain a unique identifier for the text (**text_id**), information on the text language (**text_language**) and an indication of the writing task (**task_type**), i.e., whether the text is a sample of the narrative writing task (picture story) or the argumentative writing task (opinion text).[42] All texts of the matura subcorpus were assigned to the task type matura. All L2 texts were manually annotated for sociolinguistic appropriateness (**cefr_appropriateness**), coherence (**cefr_coherence**), grammar (**cefr_grammar**) and lexis (holistically as **cefr_lexis** in Kolipsi 1, and split in lexical accuracy, **cefr_lex_accuracy**, and lexical diversity, **cefr_lex_diversity**, in Kolipsi-2). Kolipsi-2 also contains a CEFR level score for orthography (**cefr_orthography**).

3.9 Corpus access, availability and licensing

Following an open science strategy, we have made all corpora in the Kolipsi family as well as related tools and resources (e.g., Salt and Pepper modules used for data conversion and linguistic annotation, XSLT files with the annotation scheme, etc.) available for academic personal use under an ACA-BY-NC-NORED license.[43] All possible steps were performed to provide the corpora as FAIR (Findable, Accessible, Interoperable and

39 The rating grids can be accessed at https://www.porta.eurac.edu/lci/kolipsi-family/ (28.04.2023).
40 A detailed description of the local bilinguality exams is available at https://www.provinz.bz.it/bildung-sprache/zweisprachigkeit/die-zweisprachigkeitspruefung.asp (28.04.2023).
41 The test is accessible at https://dialangweb.lancaster.ac.uk/ (28.04.2023).
42 The writing tasks can be accessed at https://www.porta.eurac.edu/lci/kolipsi-family/ (28.04.2023).
43 Non-commercial, academic use with attribution to original authors. No redistribution allowed.

Table 9
Text-related metadata annotations.

annotation	*explanation*	*values*
text_id	unique identifier of each text	combination of letters/numbers
text_language	language of the text	DE, IT
task_type	type of task prompt	opinion text picture story matura
cefr_appropriateness	CEFR level assignment of sociolinguistic appropriateness	A1-C2
cefr_coherence	CEFR level assignment of text coherence	A1-C2
cefr_grammar	CEFR level assignment of grammar skills	A1-C2
cefr_lexis	CEFR level assignment of lexical skills	A1-C2
cefr_lex_accuracy	CEFR level assignment of lexical accuracy	A1-C2
cefr_lex_diversity	CEFR level assignment of lexical diversity	A1-C2
cefr_orthography	CEFR level assignment of orthography	A1-C2

Re-usable (Wilkinson et al. 2016)) resources, although we appreciate that perfect interoperability and re-usability can only be achieved with standardized and acknowledged domain-specific formats for data and metadata representation. Both are still missing in the field of learner corpus research. We tried to account for this shortcoming by harmonizing all of the resources within the Kolipsi family in terms of used annotations, metadata and provided data formats, adapting search interfaces and annotation and metadata vocabulary also to other learner corpora hosted at our institute. To cater for both a linguistic and a computational audience we offer a) a browser-based search interface to consult the corpus and extract frequency lists of searched items for aggregated corpora or subsets of individual Kolipsi corpora using the corpus query software ANNIS[44] and b) the option to download the full corpus with annotations and metadata in different, community-relevant file formats from the Eurac Research Clarin Centre (ERCC) repository.[45] The corpus search interface and corpus downloads are available via the Learner Corpus Portal PORTA, where we also provide additional documentation and list corpus-derived research outputs.[46]

44 The search interface is accessible at https://commul.eurac.edu/annis/kolipsi (28.04.2023).
45 Kolipsi-1 can be downloaded at http://hdl.handle.net/20.500.12124/64 and Kolipsi-2 at http://hdl.handle.net/20.500.12124/66.
46 PORTA is accessible at https://www.porta.eurac.edu/ (28.04.2023).

4. Conclusion and future work

The aim of this contribution was to introduce the Kolipsi Corpus Family (KCF), a collection of eight closely related corpora of L2 learner texts and L1 reference data in German and Italian. The KCF is a freely available resource, which can be directly queried via an ANNIS search interface or downloaded from the repository of the Eurac Research CLARIN Centre (ERCC).

The L2 Kolipsi-1 and Kolipsi-2 learner corpora constitute the core of the KCF. They can be used for detailed linguistic analysis of upper secondary L2 writing in German and Italian. As such, they close the gap between corpora of academic writings of university students in Italian and German L2 (represented e.g., in CELI and FALKO) and writings of lower secondary school students such as those provided by LEONIDE (Glaznieks et al. 2022). The resulting network of learner corpora facilitates comparisons between students of different ages and competence levels (e.g., Glaznieks, Frey, and Abel (2023)). In addition, the L1 reference data within the KCF enables researchers to compare students' L2 competences with their L1 competences with respect to the structure and coherence of the texts, for example. Reference data of students of the same age is necessary to understand what can be expected from a comparable group of L1 writers. Reference data of the same students, however, allows researchers to distinguish between language-specific and language-independent knowledge, challenges and problems. We hope that the results of the linguistic analyses performed on KCF resources find their way into language teaching materials seeing as the motivation of the learner corpus designers was to improve language learning. A didactic use of KCF corpora is exemplified in Schmiderer et al. (2021), which describes how extensive analyses of correct and erroneous multiword expressions lead to the creation of collocation-focused exercises for learners of Italian to enhance their communicative competences.

As well as supporting linguistic investigations and the creation of derivative educational materials, KCF corpora can also be used for pedagogical purposes by language teachers for data driven learning (DDL). Forti and Spina (2019) accurately describe how to draw conclusions from learner corpus analyses and use learner corpus data for DDL activities in the classroom to react to learner specific problems in a timely fashion. The KCF provides interested secondary school teachers with empirical data of learner varieties produced at the end of upper secondary school education. Teachers are now able to catch typical problems and evaluate whether they can be tackled with purposeful activities early on. A necessary in-between step would be to develop and improve corpus literacy for teachers, and to provide a user-friendly installation of corpus resources so that they, together with the students, might be encouraged to use them.

In providing KCF, we move one step closer to the integration of empirical data into teaching activities. We hope that the (meta)data will be used by linguists and language teachers to accumulate more knowledge on learner competences and errors at the end of secondary school education, and on the factors that facilitate or impede language learning. We are particularly hopeful that any insights gained from KCF empirical data will help create a supporting learning environment for all learners of Italian and German, especially for students in South Tyrol.

Acknowledgments
The authors thank Greta Hayley Franzini for proofreading the article.

References

Abel, Andrea, Aivars Glaznieks, Lionel Nicolas, and Egon W. Stemle. 2014. KoKo: an L1 Learner Corpus for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2414–2421, Reykjavik, Iceland, 26-31 May.

Abel, Andrea, Chiara Vettori, and Katrin Wisniewksi. 2012. *KOLIPSI. Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale*. Accademia Europea di Bolzano, Bolzano.

Alderson, J. Charles. 1991. Bands and scores. In J.-C. Alderson and B. North, editors, *Language testing in the 1990s*. British Council/Macmillan, London, pages 71–86.

Ambroso, Serena and Elisabetta Bonvino. 2008. Livelli diversi di competenza nella gestione dell'italiano L2. ipotesi dall'analisi di un corpus. *Testi e linguaggi*, 1:1–22.

Andorno, Cecilia Maria and Giuliano Bernini. 2003. Premesse teoriche e metodologiche. In A. Giacalone Ramat, editor, *Verso l'italiano. Percorsi e strategie di acquisizione*. Carocci, Roma, pages 27–36.

Bachman, Lyle F. and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press, Oxford.

Bailini, Sonia and Aldo Frigerio. 2018. CORESPI e CORITE, due nuovi strumenti per l'analisi dell'interlingua di lingue affini. *CHIMERA: Romance Corpora and Linguistic Studies*, 5(2):123–129.

Baur, Siegfried, Giorgio Mezzalira, and Walter Pichler. 2008. *La lingua degli altri. Aspetti della politica linguistica e scolastica in Alto Adige-Südtirol dal 1945 ad oggi*. Franco Angeli, Milano.

Berdičevskis, Aleksandrs. 2020. Foreigner-directed speech is simpler than native-directed: Evidence from social media. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 163–172, Online, November.

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland, 26-31 May.

Bratánková, Leontýna. 2015. *Le collocazioni Verbo+Nome in apprendenti di italiano L2*. Ph.D. thesis, Univerzita Karlova, Prague.

Brocca, Nicola. 2021. LADDER: Un corpus di scritture digitali per l'insegnamento della pragmatica in L2. Un esempio di analisi di disdette in WhatsApp. *Italiano LinguaDue*, 13(1):241–259.

Cacchione, Annamaria and Margarita Borreguero Zuloaga. 2018. I corpora di L2 come spunti per la riflessione didattica. L'interazione nativo-non nativo nel percorso formativo degli insegnanti. *RiCOGNIZIONI. Rivista di Lingue e Letterature straniere e Culture moderne*, 4(8):15–30, Feb.

Campodifiori, Emiliano, Elisabetta Figura, Monica Papini, and Roberto Ricci. 2010. *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*. Invalsi.

Corino, Elisa and Carla Marello. 2009. Elicitare scritti a partire da storie disegnate: il corpus di apprendenti VALICO. pages 113–138.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Dirdal, Hildegunn, Ingrid Kristine Hasund, Eli-Marie D. Drange, Eva Thue Vold, and Elin Maria Berg. 2022. Design and construction of the Tracking Written Learner Language (TRAWL) Corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning*, 10(2):115–135.

Fandrych, Christian, Cordula Meißner, and Adriana Slavcheva. 2012. The GeWiss corpus: Comparing spoken academic German. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins Publishing, Amsterdam, pages 319–337.

Forti, Luciana and Stefania Spina. 2019. Corpora for linguists vs. corpora for learners: Bridging the gap in Italian L2 learning and teaching. *Educazione Linguistica. Language Education (EL.LE)*, 8(2):349–362.

Giacalone Ramat, Anna, editor. 2003. *Verso l'italiano. Percorsi e strategie di acquisizione*. Carocci, Roma.

Giacalone Ramat, Anna, Maria Chini, and Cecilia Andorno. 2013. Italiano come L2. In *Linguistica italiana all'alba del terzo millennio (1997-2010)*. Societá di linguistica italiana, pages 149–205.

Glaznieks, Aivars, Jennifer-Carmen Frey, and Andrea Abel. 2023. Weil-Sätze bei Lernenden des Deutschen. Vergleich zwischen immersiv und nicht immersiv Deutschlernenden in Südtirol.

In M. Beißwenger, E. Gredel, L. Lemnitzer, and R. Schneider, editors, *Korpusgestütze Sprachanalyse: Linguistische Grundlagen, Anwendungen und Analysen*. Narr, Tübingen.
Glaznieks, Aivars, Jennifer-Carmen Frey, Maria Stopfner, Lorenzo Zanasi, and Lionel Nicolas. 2022. LEONIDE: A longitudinal trilingual corpus of young learners of italian, german and english. *International Journal of Learner Corpus Research*, 8(1):97–120.
Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier. 2015. Introduction: Learner corpus research–past, present and future. In S. Granger, G. Gilquin, and F. Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, Cambridge, pages 1–6.
Gut, Ulrike. 2012. The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins Publishing, Amsterdam, pages 3–23.
Höhn, Sviatlana. 2015. Corpus of long-term instant messaging based dialogues between advanced learners of German as a foreign language and German native speakers: deL1L2IM. Technical report.
Karges, Katharina, Thomas Studer, and Eva Wiedenkeller. 2019. On the way to a new multilingual learner corpus of foreign language learning in school: Observations about task variations. In A. Abel, A. Glaznieks, V. Lyding, and L. Nicolas, editors, *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, pages 137–165, Lovain-la-Neuve. Presses universitaires de Louvain.
Krummes, Cédric and Astrid Ensslin. 2014. What's hard in German? WHiG: a British learner corpus of German. *Corpora*, 9(2):191–205.
Laner, Josef. 2007. Im Vinschgau ist Italienisch eine „Fremdsprache". *Der Vinschger*, 26.
Linacre, John Michael. 1989. *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.
Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67–73.
Maden-Weinberger, Ursula. 2015. "Hätte, wäre, wenn…": A pseudo-longitudinal study of subjunctives in the Corpus of Learner German (CLEG). *International Journal of Learner Corpus Research*, 1(1):25–57.
Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, Baltimore, Maryland, USA, June.
Marchiodi, Mirco. 2011. Le imprese altoatesine: lingue da migliorare. *Alto Adige*.
North, Brian. 2000. *The development of a common framework scale of language proficiency*. Peter Lang, Oxford.
Palermo, Massimo. 2009. L'ADIL2 come strumento per la ricerca. In M. Palermo, editor, *Percorsi e strategie di apprendimento dell'italiano lingua seconda: Sondaggi su ADIL2*.
Pallaver, Günther. 2017. Il sistema politico in provincia di Bolzano: la complessa ripartizione del potere e le sfere di influenza etnica. In H. Atz, M. Haller, and G. Pallaver, editors, *Differenziazione etnica e stratificazione sociale in Alto Adige. Una ricerca empirica*. FrancoAngeli, Milano, pages 57–74.
Pallotti, Gabriele, Stefania Ferrari, Elena Nuzzo, and Camilla Bettoni. 2010. Una procedura sistematica per osservare la variabilità nell'interlingua. *Studi Italiani di Linguistica Teorica e Applicata*, 39(2):215–241.
Paquot, Magali, Alexander König, Egon Stemle, and Jennifer-Carmen Frey. 2023. *Core Metadata Schema for Learner Corpora*. https://doi.org/10.14428/DVN/4CDX3P.
Perdue, Clive. 1993. *Adult language acquisition: Cross-linguistic perspectives*. Cambridge University Press.
Reznicek, Mark, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Technical report, Humboldt-Universtät zu Berlin.
Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. *Natural language processing using very large corpora*, pages 13–25.
Schmiderer, Katrin, Lorenzo Zanasi, Christine Konecny, and Erica Autelli. 2021. *Facciamo bella figura!: 8 task fraseodidattici per studenti di italiano L2/LS. Con una prefazione e con la consulenza scientifica di Barbara Hinger*. innsbruck university press, Innsbruck.

Siyanova-Chanturia, Anna and Stefania Spina. 2020. Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, 70(2):420–463.
Spina, Stefania, Irene Fioravanti, Luciana Forti, Valentino Santucci, Angela Scerra, and Fabio Zanda. 2022. Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, 14(1):116–138.
Tagnin, Stella EO. 2006. A multilingual learner corpus in Brazil. *Language and Computers*, 56(2):195–202.
Turco, Giuseppina and Miriam Voghera. 2010. From text to lexicon: the annotation of pre-target structures in an Italian learner corpus. In *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*, pages 141–174. Firenze University Press.
Vettori, Chiara and Andrea Abel. 2017. *KOLIPSI II. Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale*. Eurac Research, Bolzano.
Vyatkina, Nina. 2016. The Kansas Developmental Learner corpus (KANDEL): A developmental corpus of learner German. *International Journal of Learner Corpus Research*, 2(1):101–119.
Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.
Wisniewski, Katrin, Elisabeth Muntschik, and Annette Portmann. 2022. Schreiben in der Studiersprache Deutsch: Das Lernerkorpus DISKO. In K. Wisniewski, W. Lenhard, J. Möhring, and L. Spiegel, editors, *Sprache und Studienerfolg bei Bildungsausländer/-innen*. Waxmann, Münster, pages 283–304.
Zinsmeister, Heike and Margit Breckle. 2012. The ALeSKo learner corpus. design – annotation – quantitative analyses. In T. Schmidt and K. Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins Publishing, Amsterdam, pages 71–96.