CORPUS REPORTS

Leonide

A longitudinal trilingual corpus of young learners of Italian, German and English

Aivars Glaznieks, Jennifer-Carmen Frey, Maria Stopfner, Lorenzo Zanasi and Lionel Nicolas
Eurac Research

This article presents the longitudinal trilingual corpus of young learners of Italian, German and English called LEONIDE. The corpus consists of L1, L2 and L3 learner texts. L1 texts were written in two languages of schooling (i.e. Italian and German), L2 texts in two languages learned as second languages (i.e. German and Italian), and L3 texts in an additional foreign language (i.e. English). All texts were collected from a group of lower secondary school pupils from the multilingual Italian province of South Tyrol whose development in all three languages was observed over a period of three years. Each text comes with rich metadata as well as manual and automatic annotations.

Keywords: longitudinal learner corpus, young learners, multilingual corpus, L2 German, L2 Italian, L3 English

1. Introduction

Longitudinal corpora consisting of young learners' writings are rare and difficult to access. Examples for such corpora are the publicly accessible *International Corpus of Crosslinguistic Interlanguage* (ICCI; Tono & Díez-Bedmar, 2014) of English as a foreign language or the yet unavailable *Tracking Written Learner Language Corpus (*TRAWL; Dirdal et al., 2017) of L1 Norwegian young learners of English, French, German and Spanish, and the SWIKO corpus of French and German-speaking Swiss pupils learning German, French and English (used in Karges et al., 2019). ICCI and TRAWL are compiled of texts written by primary and secondary school pupils, while SWIKO considers only upper secondary school pupils. To the best of our knowledge, TRAWL and SWIKO are the only written multilingual corpora of young learners.

https://doi.org/10.1075/ijlcr.21004.gla
International Journal of Learner Corpus Research 8:1 (2022), pp. 97–120. ISSN 2215-1478 | E-ISSN 2215-1486

This paper introduces the recently compiled learner corpus LEONIDE, which fills a gap in the map of existing learner corpora and combines some of the less available characteristics in learner corpus research (LCR), while being freely available to the research community as it can be downloaded in several formats or queried online. It is a trilingual learner corpus of German, Italian and English texts, and provides a balanced spectrum of Italian, German, Italian-German bilingual and other mono- or plurilingual language backgrounds. It contains longitudinal data in the form of originally handwritten essays from three consecutive years of lower-secondary schooling. It thus represents young learners between 11 and 14 years of age and mainly reflects lower levels of language proficiency in L2 (Italian/German) and L3 (English), with reference data from the same pupils in their main language of instruction (Italian/German, usually the learners' L1).

In this corpus report, we will refer to languages mainly from a teaching perspective. The main languages of instruction of the schools represented in our data can be either Italian or German, which are the two official languages of the province of South Tyrol (henceforth L_{inst}). Language instruction in the school language is designed for native speakers of the language. The term L_{inst} therefore also refers to the language used in native language instruction (i.e. Italian in schools with Italian as L_{inst}, German in schools with German as L_{inst}). The term *L1* is reserved for the pupils' first language(s), as indicated by themselves. The L1 often, but not always, coincides with the L_{inst}. *L2* will be used for Italian and German when it is not used as L_{inst} but taught at school as a second language. English is not an official language in South Tyrol and, compared to both L2s, takes up fewer teaching hours and plays a minor role in the local community (see Section 2.2). Hence, we will refer to English as *L3*. This classification makes it possible to take into account both the local school system and the distribution of the different linguistic backgrounds of the pupils.

In comparison to existing learner corpora of young learners, in particular to the most similar learner corpora mentioned above, LEONIDE shows some important differences and unique features. For instance, while ICCI focuses on English as a foreign language and samples texts obtained from various regions and countries and different school grades (cross-sectional data), LEONIDE includes texts in three target languages written by the same pupils and is longitudinal in a strict sense as it follows the same classes and pupils over a period of three years. Like in ICCI, the participating pupils in LEONIDE vary with respect to their L1 and the L_{inst}, but in LEONIDE they live in the same multilingual region.

The TRAWL corpus is in many aspects similar to LEONIDE: It is also multilingual with respect to the target languages, it contains longitudinal data, and, for a subset of writers, allows also for comparisons between L1, L2 and L3 development; however, TRAWL covers pupils with the same L1 (Norwegian), whereas LEONIDE represents monolingual German and Italian, German-Italian bilinguals and pupils with other mono- and plurilingual linguistic backgrounds. In addition, TRAWL assembles texts written in class during regular schoolwork, whereas LEONIDE used specifically designed prompts to facilitate comparisons across participants and time.

Finally, SWIKO has also many overlaps with LEONIDE. It contains texts in three languages (German, French, English) which are partly used as L_{inst} (German, French), as L2 (German, French) and as L3 (English). Although the corpus design is similar, the texts for SWIKO were collected in two predominantly monolingual regions, in German- and French-speaking Switzerland, whereas the texts for LEONIDE were collected in the multilingual province of South Tyrol. Unlike LEONIDE, SWIKO does not provide longitudinal data and cannot therefore be used for developmental studies.

In the coming sections, we will briefly present the language situation in the province of South Tyrol and introduce the project for which the data was collected (Section 2). In Section 3 we describe the corpus design and in Section 4 the corpus data. Section 5 then gives indications on corpus availability and access. Finally, in Section 6, we hint towards potential applications of the corpus in LCR and a language teaching environment.

2. Origin of the corpus data

2.1 Linguistic situation of South Tyrol

All data was collected in the Autonomous Province of South Tyrol, Italy's northernmost province. Like in many border regions, the linguistic profile of South Tyrol is heterogenous. In general, German is the language of everyday life for most people in South Tyrol, coexisting with Italian as the language of the nation state and Ladin, a minority language closely related to Friulian and Romansh spoken in the Dolomite valleys. According to the last census (Astat, 2012), around 70% of the population belong to the German, 25% to the Italian and 5% to the Ladin language communities. Furthermore, South Tyrol is characterised by an inverse sociolinguistic situation whereby the number of Italian speakers increases in urban areas and in the southern parts of the province, while German and Ladin speakers dominate in rural areas and in the Ladin Dolomite valleys (Val Gar-

dena and Val Badia), respectively. In addition, the use of different varieties can be noticed among German speakers. Whereas the medium of instruction at school is the standard variety of German, the language exclusively used at home and in the private sphere is the Austro-Bavarian dialect. To make this already complex puzzle even more diverse, since the early 1990s, a slowly increasing number of immigrants have been settling in South Tyrol (Voltmer, 2007), contributing to the territory's language diversity with further languages, such as Albanian, Arabic, Urdu, etc.

The educational system of South Tyrol, while adhering to the comprehensive and inclusive education system of Italy, has adapted to this multilingual setting by establishing three parallel schooling systems for the three official language groups. Consequently, schools in South Tyrol can be divided into schools with German as L_{inst}, schools with Italian as L_{inst}, and schools in the Ladin valleys, where instruction is equally split between German and Italian, while Ladin is taught as an additional subject and can be used for explanations (e.g. Alber, 2012). Starting years and the number of hours of language teaching differ depending on the school system: In schools with German as L_{inst}, L2 Italian is taught from first grade and L3 English from the fourth grade of primary school. In schools with Italian as L_{inst}, L2 German and L3 English are taught from the first grade onwards, but with different numbers of hours per year (see Table 1). While pupils in the German school system finish their primary school education with a minimum of 646 hours of L2 instruction in Italian and 136 of L3 instruction in English, pupils from schools with L_{inst} Italian enter lower secondary school education with a minimum of 969 hours of L2 German and 357 hours of L3 English. In lower secondary schools, the minimum number of hours in L2/L3 instruction for both school systems is similar with 408 hours for L2 Italian and 426 hours for L2 German, while L3 English is represented with 204 and 255 hours of language instruction, respectively.

Table 1. Minimum number of hours of L2/L3 language teaching in South Tyrol (Deutsches Schulamt, 2009; Dipartimento Istruzione e Formazione italiana, 2008)

	German schools	Italian schools
Primary school		
L2 German	–	799
L2 Italian	646	–
L3 English	136	357
Lower secondary school		
L2 German	–	426
L2 Italian	408	–
L3 English	204	255

2.2 The project *One School, Many Languages*

The research project *One School, Many Languages* was founded with the aim of providing insight into the current situation of multilingualism in South Tyrolean schools, studying how the educational and linguistic landscape is evolving, and how linguistic repertoires and competences can be assessed, valued, and promoted. Since its beginning in 2012, the project has developed a series of work packages dealing with different aspects of multilingualism, ranging from teaching material and teacher training to in-class workshops and interactive tools for parents and families (see Engel & Stopfner, 2019). One of the work packages aimed to capture a holistic view of the linguistic repertoire and the development of plurilingual competences of individual learners in a multilingual setting such as South Tyrol (see Busch, 2012; Ehlich, 2005; Lüdi, 2006). In a longitudinal linguistic study, a variety of instruments for data collection were used to triangulate plurilingual competences within different communicative settings, ranging from written and oral assessments of the languages of schooling (see Bettoni & di Biase, 2015; Grießhaber, 2006, 2010; Grotjahn, 2014; Keßler, 2006; Pienemann, 1998) to ethnographic and systematic observations in and outside of the classroom (see Gogolin et al. 2011) and semi-structured interviews and questionnaires (see Flick, 2011; Gogolin, 2004; Reich, 2010).[1]

1. For more information about the project, visit the project website: http://sms-project.eurac.edu/.

3. Corpus design

LEONIDE is one of the products of the longitudinal study in the *One School, Many Languages* project described above. Between 2015 and 2018, the project followed eight classes and over 40 language teachers from the first to the third (and last) year of lower secondary school. The schools were chosen together with the cooperating school boards based on the sociolinguistic environment, i.e. combining schools with German as L_{inst} with schools with Italian as L_{inst}, and, for each group, schools situated in a predominantly Italian-speaking environment with schools situated in a predominantly German-speaking environment. Together with the school headmasters, the project researchers chose one class per school, ensuring that there would be at least one pupil from a linguistic background other than those typical of South Tyrol. The study then followed the approximately 170 pupils for three years, observing their language competences in the three languages that are taught in lower secondary schools in South Tyrol, namely German, Italian and English.

The multilingual LEONIDE corpus contains written texts collected from pupils as part of written language assessments conducted in-class once a year for each language and consisting of two genre-specific writing tasks (an opinion text and a picture story). The corpus thus contains a balanced number of texts in three languages and two text genres, collected from pupils of different L1 backgrounds but taught in either German or Italian as L_{inst}. The corpus allows one to trace pupils' progress over the span of three years and across various languages (L_{inst}, L2, L3), and to make cross-sectional comparisons with pupils with other language backgrounds and school instruction.

Participant metadata provided in the corpus was collected through an additional questionnaire and includes anonymous identifiers for each pupil and school class, as well as age, gender and language background.

3.1 Written language assessment tasks

Two genre-specific writing tasks were used to elicit narrative and argumentative texts. The time limit for each writing task was set to 20 minutes.[2] For the narrative

2. To test the comprehensibility and comparability of the pictorial and written writing prompts in all three languages, pre-tests were conducted in two classes that did not participate in the study. The pre-test classes were furthermore used to check the general comprehensibility of the instructions, the temporal and psychological reasonableness of the writing tasks and the test fairness with regard to gender-specific, cultural and/or ethnic discrimination (Ingenkamp & Lissmann, 2008; Pospeschill, 2010).

writing task, we used a set of picture stories as story-telling input. Writing narrative texts is a lower secondary school teaching objective for years 1 to 3. As pupils had to complete this narrative task every year with no restrictions on word count or page length, special attention was paid to appropriate stimuli for each year. In general, pictures and passages used as input for the story-telling tasks did not contain any written language.

In year 1, pupils were asked to write a funny story based on the common picture stories of the *Father and Son* series by E.O. Plauen (*Der Schmöker, Der gelöschte Vater, Die gute Gelegenheit*) under the assumption that they would be familiar with these stories and hence comfortable with completing this task even in this unusual setting of scientific language assessment. In year 2, pupils had to write a scary story based on passages from the Mariko and Jillian Tamaki's coming-of-age graphic novel *This one summer* (2014). In contrast to years 1 and 2, the stimuli of year 3 were taken from three different sources: For German, we chose a passage from Shaun Tan's graphic novel *The Arrival* (2006) for its allusion to the 3rd Reich, one of the main study topics in year 3. For Italian and English, we chose graphic novels and comics that supposedly relate to adolescent life: For Italian, a passage from Vera Brosgol's graphic novel *Anya's Ghost* (2011) about teenage love; for English, a scene from Marjane Satrapi's *Persepolis* (2000, 2004), where the main protagonist gets into a conflict with the authorities. Over the three years, the pupils were asked to write something for each picture in the chosen passages, i.e. a total of 12 pictures for German, 8 pictures for Italian and 12 pictures for English. In addition, in the second year, the pupils were also asked to invent an ending for the story. Appendix 1 shows an example of a picture story task used in the first year.

The second writing task aimed at the production of an opinion text, a text type that is not part of the curriculum and, hence, not taught and practised until year 3 of lower secondary school. However, by the end of lower secondary school, pupils are expected to be able to write argumentative texts expressing their opinion in a sophisticated way, assuming and evaluating different perspectives and giving reasons for their own point of view (see Deutsches Schulamt, 2009). Considering the importance of this text genre also with respect to the pupil's educational path in upper secondary school, the opinion-oriented texts are meant to give insight into the development of academic language proficiency (Cummins, 1984) before and after the onset of explicit formal education in argumentative writing. Owing to the difficulty of the writing tasks and the pupils' unfamiliarity with the genre, cooperation partners and teachers insisted on introductory texts for the L2 and L3 task that should illustrate possible arguments and typical writing style. For these tasks, we used low complexity topics and longer introductory

texts. An example of an introductory text used in the second year for L3 English can be found in Appendix 2.[3]

Unlike the picture story task, the argumentative task was the same for the first and third year, so as to be able to discern differences in individual development with respect to the same topic. In so doing, we accepted the risk that pupils in their third year might remember and be influenced by the task they performed in their first year.

3.2 Participants

The total number of pupils whose texts were integrated in the corpus is 163 (76 female, 87 male).[4] Most pupils were between the ages of 11 and 12 old when the study started. Only a small percentage (9/163 = 5%) was slightly older. As we collected data from minors, it was necessary to obtain an informed consent from their parents. In addition, all participating pupils were informed about the study and the tasks by their teachers and the members of the research team prior to data collection. Of the 163 pupils, 82 were attending schools with German as L_{inst} and 81 pupils were attending schools with Italian as L_{inst}.

Regarding the pupils' first languages, LEONIDE provides a rather heterogeneous learner group (see Table 2).

Table 2. Writers and their L1s in LEONIDE

monolingual	German	41 pupils
	Italian	46 pupils
	other	40 pupils
plurilingual		36 pupils

Forty out of the 41 pupils who indicated German as their only L1 attended schools with German as L_{inst}, and 43 of the 46 pupils who indicated Italian as their only L1 attended schools with Italian as L_{inst}. Forty pupils were raised with a language other than German or Italian (17 of which attended German schools, 23 Italian schools) as L1. In addition, 36 pupils came from a multilingual household

3. All task sheets for the opinion text task and more examples of picture stories and detailed reference to the passages used in the tasks are provided on the corpus website www.porta.eurac.edu.

4. There were 26 pupils with declared special educational needs: (a) 10 pupils with learning difficulties (e.g. dyslexia), (b) 4 pupils with a physical or mental impairment, (c) 3 pupils with a combination of (A) and (B), and finally (d) 9 pupils with other special educational needs.

in which at least one of the three target languages was spoken (German, Italian or English).

The maximum number of texts any individual pupil could contribute to LEONIDE is 18. This is equivalent to two texts in all three languages across all three years of data collection. This number was reached by 94 pupils and makes it possible to study their parallel development in all three languages. The number of complete collections per language (= six texts, two tasks for each of the three years), however, is higher: 116 in Italian (59 by non-native speakers of Italian), 115 in German (59 by non-native speakers of German), 115 in English (110 by non-native speakers of English).

4. Corpus data

4.1 Corpus size

LEONIDE is subdivided into three sub-corpora according to the target language of the texts: LEONIDE_EN, LEONIDE_DE and LEONIDE_IT.[5] Table 3 shows the number of texts and tokens[6] for each sub-corpus.

Table 3. Corpus size of LEONIDE split by sub-corpus

			Text length	
Sub-corpus	Number of texts	Number of tokens (rounded)	Ø (median)	Variance (IQR)
LEONIDE_EN	835	69,700	77	58
LEONIDE_DE	833	73,900	77	79
LEONIDE_IT	844	93,300	96	72
Total	**2,512**	**236,900**	**83**	**69**

The size of LEONIDE amounts to ca. 236,900 tokens coming from 2,512 texts. On average, each text has 94 tokens; however, the range of tokens per text is quite wide: the shortest text consists of only 1 token, the longest of 517 tokens. The median for the number of tokens per text is 83 with an inter-quartile range of 69 tokens.

5. The suffixes refer to the official ISO 639–1 codes of the respective languages English (EN), German (DE) and Italian (IT).

6. We refer to both words and punctuation signs by the term *token*.

Tokens are almost equally distributed over the three languages and over the task type (1,246 narrations on a picture story vs. 1,266 opinion texts, see Figure 1). There are slightly more texts from the second year (860 texts) compared to the first year (831) and the third year (817). The main reason for this distribution is the fluctuation of pupils' presence on the days the texts were written.

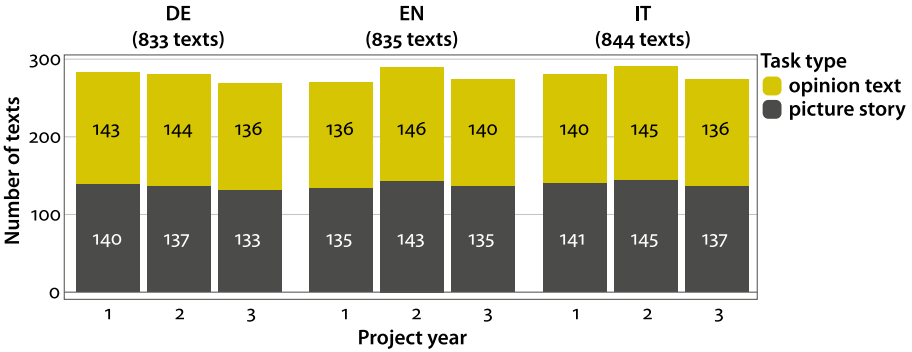


Figure 1. Distribution of texts in LEONIDE by text language, year of production and task type

Each sub-corpus contains texts written by monolingual or plurilingual L1 writers of the respective languages and by writers with other language backgrounds. Moreover, it assembles texts from all participating schools, regardless of whether the L_{inst} was German or Italian. The corpora can, however, easily be filtered in order to include, e.g., only L2 writers or texts written in schools with Italian as L_{inst}, or a combination of both, through the respective metadata information. Figure 2 shows the distribution of texts split by task type and L_{inst}. In addition, the figure shows the portion of native speakers of the target language of each sub-corpus. In the German and Italian sub-corpus, a considerable number of texts are written by native speakers of the respective language (40–45%). This data can be used as a reference corpus for the non-native texts.

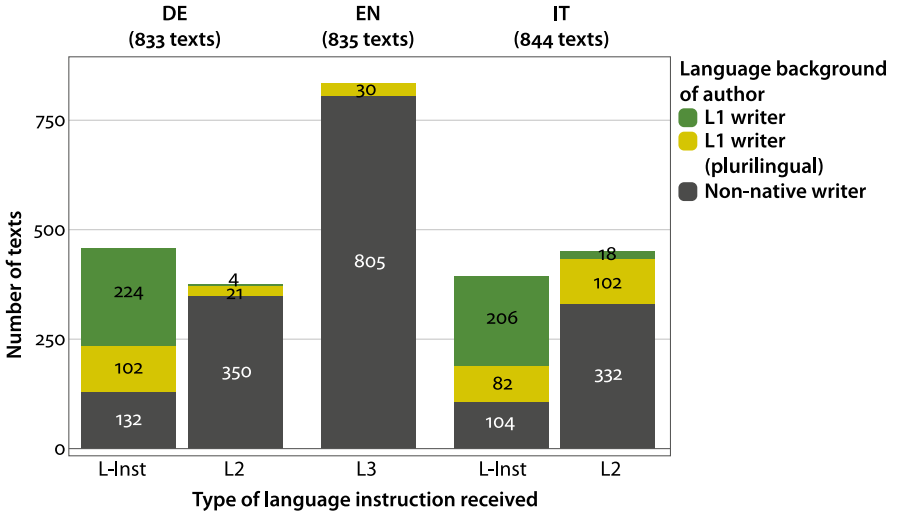


Figure 2. Distribution of texts by L_{inst} and language background of the writer

4.2 Transcription

The originally handwritten texts were scanned and transcribed using the transcription tool Transc&Anno (Okinina et al. 2018). Transc&Anno[7] is a browser-based tool that allows users to upload scans as picture or PDF files and to transcribe and annotate them collaboratively using a split screen visualization that shows both the scan and a simple text editor with annotation options. The transcription and subsequent annotation (see Section 4.3) of the handwritten texts was performed by two trained transcribers/annotators according to explicit transcription and annotation guidelines.[8]

As there was no overlap between the texts transcribed in this initial digitization phase, we evaluated the quality of the transcriptions retrospectively on an evaluation sample of 10% of the corpus, checking the correctness of the transcripts word-by-word and creating a revised transcript (gold standard) for all the transcripts in the evaluation sample. For the creation of the evaluation sample, we extracted a stratified random sample with fixed sample size for each stratum, considering the transcribers ($N=2$), text languages ($N=3$) and the different years in which the texts were written ($N=3$), to observe the effects of transcribers, text languages or text proficiency on the quality of the transcripts. For each stratum, a

7. The tool is available online at https://kommul.eurac.edu/transcanno/.

8. https://www.porta.eurac.edu/wp-content/uploads/2020/06/LEONIDE_Guidelines_06_2020-2.pdf

total of 15 texts was randomly sampled, amounting to 270 texts in the evaluation sample, which equates to 10.7% of the total number of texts in the corpus (2,512).

All texts in the evaluation sample were then reviewed for completeness and correctness by a third linguistically trained person, who made sure that the transcription guidelines were applied correctly, that all learner spellings were transcribed as they appeared in the handwritten originals (e.g., retaining all errors) and no text was missing or added erroneously to the transcripts.

After obtaining a corrected version of the transcripts we compared both versions and noted all deletions, insertions or substitutions that were made for the corrected version. We then calculated the word error rate for the evaluation sample in order to get an estimate of the overall quality of the transcriptions in the corpus. The word error rate measure, usually used for speech recognition or machine translation systems, allows to account for changing text lengths between a text and its target version (in our case the original transcript vs. the corrected transcript) that can occur due to erroneously omitted or inserted words or lines. It is calculated by summing all substitutions, deletions and insertions of tokens and dividing it by the total number of tokens in the final, correct version of the text. A comparison of the original texts with the revised texts in our evaluation sample showed a total of 129 modifications (89 amended deviations from the original spelling of the pupils, 35 insertions and 5 deletions of words or punctuation marks) in the revised texts, for a total of 25,258 tokens amounting to a word error rate of 0.51%. With a word accuracy of over 99% we decided to accept this error margin without further correcting the transcriptions for the remaining corpus.

4.3 Manual annotations

The manual annotations were carried out in several steps. The annotations are based on the schema used in the learner corpus projects MERLIN (Boyd et al. 2014) and KoKo (Abel et al. 2014). They refer to different aspects of the learner texts: (1) the structure of the text, (2) orthographic errors, (3) the choice of linguistic means, (4) legibility of handwriting, (5) self-corrections, (6) the use of stylistic means, and (7) anonymization (see Table 4).

1. Annotations about the structural characteristics of the text are **lines, paragraphs,** and **pages**. The ability to structure a text in meaningful units is a major challenge in text production. This type of annotation helps users of the corpus study differences in text structuring among participants.
2. All misspellings were annotated with an **orthographic error** tag. In doing so, annotators also provided the inherent target hypotheses (i.e. the orthographic correct spelling of the word the learner probably intended to write).

The added word token in standard spelling can be used for further automatic processing, and thus reduce potential erroneous automatic annotations of, for instance, part of speech and lemma.

3. Annotations were provided for the use of **foreign words**, i.e. words and expressions that according to common dictionaries do not belong to the target language.[9] The annotation also allows to specify the foreign language used and thus analyse the language(s) that pupils fall back on when they do not know a word in the target language. The second annotation helped to investigate uncertainty in word choice by indicating the indecisive use of two or more **variants** for one word (e.g. the use of both *child* and *kid* instead of one or the other).
4. Two annotation tags reflect the legibility of the handwriting: Annotators could use an **ambiguous** tag if unable to decide between two potential readings of characters or word spellings. The two potential readings were then added to the tag. In those cases where character, a part of the word or the entire word was not readable, an asterisk was inserted as a placeholder token and annotated as **unreadable**. These two annotations consider the fact that handwritten texts, especially if available as scans only, are often difficult to read and transcribe.
5. There are three annotations that specify self-corrections of the pupils in their texts. We distinguish between **word correction** within a word (e.g. correction of a letter or a group of letters), **word deletions** (i.e. the complete word is deleted) and **word insertions** (i.e. a complete word is inserted). Whereas word corrections relate to spelling corrections and corrections of inflexional affixes, word deletions and insertions often reveal vocabulary and syntactical challenges.
6. A series of annotations are added to maintain pragmatic and discursive meaning of certain elements. Annotated stylistic means include the **capitalisation** of entire words (i.e. the exclusive use of capital letters without distinguishing between lower and upper case letters), **emoticons** (i.e. combinations of interpunction signs, letters and numbers to graphically display facial expressions), all kinds of **emphases** (e.g. bold text, underlined words), **images** (e.g. drawings as part of the texts), all forms of non-lexicalised uncommon abbreviations or **reductions** of words (e.g. "Ita" for "Italian"), and **symbols** (i.e. icons with a symbolic or iconic meaning, e.g. arrows, hearts, etc.). In the capitalisation and reduction annotation, the common, non-capitalised and non-abbreviated word form is also, respectively, indicated. Again, the added target

9. We used the online editions of the Oxford dictionary for English, Duden for German and Treccani for Italian language texts.

word can be used for further automatic annotations and contributes to better search results for corpus queries.

7. In order to guarantee anonymity, all names of individuals, pets, places and schools were **anonymised** using placeholder tokens (e.g. "Forename", "Schoolname").[10]

Table 4. Manual annotations in LEONIDE

Annotation	Explanation	Frequency
(1) structure of the text		
lines		28,414
pages		2,694
paragraphs		5,485
(2) orthographic errors		
orthographic error	misspelled word (no morpho-syntactic errors)	14,509
(3) choice of linguistic means		
foreign word	words that do not belong to the expected target language	3,140
variants	indecisive use of variants for one word	33
(4) legibility of handwriting		
ambiguous	two potential readings of a word	1,164
unreadable	unreadable word or part of a word	1,647
(5) pupils' self-corrections		
word correction	correction of a letter or a group of letters within a word	5,710
word deletion	deletion of a word	5,333
word insertion	insertion of a word	1,760
(6) use of stylistic means		
capitalisation	upper case use throughout words	1,719
emoticon	combination of punctuation marks, letters, and numbers	31
emphasis	bold text, underlined words	199
image	e.g., drawings	40
reduction	non-lexicalised word reduction	45
symbol	icons with a symbolic or iconic meaning	446
(7) anonymisation		
anonymisation	names of individuals, pets, places, schools	160

10. The anonymization guidelines are available online at www.porta.eurac.edu.

All manual annotations of the texts were done after the transcription using a custom set of annotations within Transc&Anno. The tool makes it possible to add annotations by simply highlighting words in the text editor and selecting the type of annotation from a list. Detailed annotation guidelines provided within the interface, as well as through additional material and preliminary training guide the annotator through the annotation process.

We manually checked all annotations for unreadable words, ambiguous words, foreign words, reductions, and symbols after the annotation process was completed, correcting any errors encountered. For the orthographic error annotations and the assigned target hypothesis, which were too numerous to check in their entirety, we measured the quality of the annotations using the same evaluation sample as that for the evaluation of transcription quality (see Section 4.2).

Out of the 1,540 orthographic error annotations in the sample, 1,493 (96.9%) were correct annotations, 41 (2.6%) were erroneous and 6 (0.4%) were evaluated as unclear cases. In an additional analysis we categorised the erroneous annotations and found that most (22 = 53.7%) were indeed orthographic errors but showed a mistake in the target hypothesis. The remaining erroneous annotations (19 = 46.3%) were due to:

- transcription errors (5)
- annotations on text revisions that were marked as deleted by the pupil and thus were not part of the final corpus (5)
- annotations of grammatical errors that were mistakenly annotated as orthographic errors (4)
- correct orthographic variants (4)
- annotations on anonymised tokens that also did not find their way into the final corpus (1).

4.4 Automatic annotations

Sentence splitting, tokenisation, lemmatization, and part-of-speech tagging was done automatically for all texts in the corpus. For the German and Italian texts, the Open NLP toolkit[11] was used for sentence splitting, while tokenisation, lemmatisation and part-of-speech tagging were performed with TreeTagger.[12] For English, all automatic processing (sentence splitting, tokenisation, lemmatisation,

11. opennlp-tools 1.9.1 (https://mvnrepository.com/artifact/org.apache.opennlp/opennlp-tools/1.9.1)

12. org.annolab.tt4j 1.2.1 (https://mvnrepository.com/artifact/org.annolab.tt4j/org.annolab.tt4j/1.2.1)

and part-of-speech tagging) was done using the Stanford Core NLP toolkit.[13] Apart from the language-specific part-of-speech tags assigned by the TreeTagger/ Stanford Core NLP toolkit, we used conversion tables to provide standard Universal Dependencies part-of-speech to better compare the different sub-corpora.[14]

4.5 Metadata

The metadata provided in the corpus was obtained via an additional questionnaire that the pupils had to fill out. In keeping with the Wilkinson et al. (2016) guidelines, we included as much corpus metadata as possible to enhance the reusability of the data. Based on Granger and Paquot's (2017) proposal for standardised core metadata for learner corpora, we considered five main components for the corpus. Besides administrative information (e.g., corpus name, authors, version, availability, and licence information) and information about corpus design (e.g., target languages, corpus size, study level, place of data collection), we provided text and learner-related metadata for each text and each learner. Text-related metadata defines the task type (picture story vs. opinion text), as well as the year of text production (1st, 2nd, 3rd), the language of the text (Italian, German, English), the L_{inst} (Italian, German) and the class identifier indicating all texts written in one class. Learner-related metadata currently consists of the pupils' L1(s), gender and age (in the first year of data collection), as well as information about their special needs.[15] Text and learner metadata are linked via a unique author identifier, which makes it possible to search for all texts of a given author, as well as for all texts that fit a certain author profile. Additional metadata items on the text level also reveal whether the author has completed all writing tasks, whether (s)he completed all writing tasks for one of the languages or one of the text types, or whether some texts for these pupils are missing (e.g. due to sick leaves or blank submissions). This allows users of the corpus to easily restrict the analysis sample to pupils for which all texts of one or more categories are present. The metadata can be downloaded in a tab-separated format or used as a filter in the browser-based corpus query interface.

---

13. stanford-corenlp 3.9.2 (https://stanfordnlp.github.io/CoreNLP/history.html)

14. https://universaldependencies.org/tagset-conversion/

15. This information was not obtained through the questionnaires but was provided by the head of the participating schools.

5. Corpus access and reusability

In order to provide the corpus as a FAIR (findable, accessible, interoperable and re-usable, see Wilkinson et al., 2016) resource, the data has been made available to the scientific community by offering both a search interface using the corpus query software ANNIS[16] and the option to download the full corpus with annotations and metadata in different, community-relevant file formats (XML, ANNIS, TXT, CSV) from a research data repository[17] for individual use under the ACA-BY-NC-NORED licence. The corpus search interface, as well as corpus downloads, are available via the Learner Corpus Portal PORTA,[18] where we also give additional documentation and list corpus-derived research outputs relevant to the LCR community.

6. Potential future use of LEONIDE in LCR and corpus-based language teaching

LEONIDE is a valuable corpus for researchers interested in German and Italian learned as L2, or English learned as a foreign language. It offers longitudinal data for all three languages and rich metadata about each writer. However, languages are not learned separately from each other. The first language as well as other languages learned influence the learning of additional languages. Researchers have observed traces of cross-linguistic influence (CLI) in language learning, whether positive or negative, from L1 or an additionally acquired language, at all linguistic levels (see de Bot & Jaensch, 2015 for an overview). LEONIDE is therefore also a valuable corpus for researchers interested in the parallel acquisition of several languages and the cross-linguistic influences this may cause over a certain period of time, particularly at the early stages of language learning. It provides insights into L3 English learning considering the learners' development in a L_{inst} (usually the learners' L1) and L2. L_{inst} and L2 in LEONIDE are typologically different languages – German being a Germanic language and Italian a Romance language – which differ from each other and from English, making it possible to study CLI beyond the lexical level. Typological differences refer, e.g., to word order (V2 in German, SVO in Italian and English), the verbal system (e.g., no continuous aspect in German) or pronoun-dropping (subject pronoun drop in Italian, no

16. https://commul.eurac.edu/annis/leonide.

17. http://hdl.handle.net/20.500.12124/25.

18. https://www.porta.eurac.edu/.

pro-drop in German and English). The specific design of LEONIDE supports comparative investigations of L_{inst} German and L2 Italian with L_{inst} Italian and L2 German to determine CLI in L3 English. The fact that in many cases the pupils' L1 coincides with L_{inst}, but for many others it does not, can be used for further investigations and comparisons: On the one hand, development in L1 writing can be accounted for in any analysis of L2 and L3 writing. On the other, if L_{inst} differs from the pupils' L1, research on parallel learning of two L2s is possible.

In this respect, LEONIDE is a unique learner corpus as it enables research on the parallel development of L1 and L2/L3 proficiency and their interplay in two combinations over a period of three years.

In addition, LEONIDE is a useful corpus from a linguistic-pedagogical perspective. Following Leech's (1997) subdivision, LEONIDE can be exploited both indirectly, that is, by researchers for the preparation of teaching materials, and directly by language teachers and pupils together. In both cases the longitudinal and multilingual characteristics of the LEONIDE data play a central role.

The longitudinal characteristics of the LEONIDE data allows researchers and users interested in language teaching to conduct analyses in the field of language sequencing, one of the main components indicated by Granger (2015: 487) as essential in the construction of materials and educational paths based on corpora. Sequencing identifies the order in which language traits should be presented to the learner. A longitudinal corpus is ideal for designing didactic paths that highlight the emergence of linguistic structures in different steps, through various phases of interlanguage. As LEONIDE represents texts of young learners at the initial stages of language learning, the corpus can reveal interesting learner aspects of beginners. Furthermore, at the interlinguistic level, researchers and teachers can point out common transfer phenomena between related languages or between different languages, which share a communicative space (e.g. German and Italian in South Tyrol).

In a more direct use, LEONIDE makes it possible to highlight changes of linguistic knowledge in class by comparing texts over a given time period. Thus, the pupils receive an input of language and content adapted to their sociolinguistic position as the texts were produced by pupils and relate to subjects typical of the school period. Finally, the trilingual character of the corpus allows teachers and pupils to focus on contrastive aspects in a multilingual didactic framework.

Acknowledgements

This article has been awarded an Open Data badge. All data are publicly accessible at: http://hdl.handle.net/20.500.12124/25. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

References

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. W. (2014). KoKo: An L1 learner corpus for German. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2414–2421).

Alber, E. (2012). South Tyrol's education system: Plurilingual answers for monolinguistic spheres? *L'Europe en Formation*, *363*(1), 399–415. https://doi.org/10.3917/eufor.363.0399

Astat. (2012). *Volkszählung 2011 – Censimento della popolazione 2011*. astatinfo 38. https://astat.provinz.bz.it/downloads/mit38_2012.pdf

Bettoni, C., & di Biase, B. (Eds.). (2015). *Grammatical development in second languages: Exploring the boundaries of Processability Theory*. The European Second Language Association.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus: learner language and the CEFR. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1281–1288).

Brosgol, V. (2011). *Anya's ghost*. First Second Books.

Busch, B. (2012). The linguistic repertoire revisited. *Applied Linguistics 33*(5), 503–523. https://doi.org/10.1093/applin/ams056

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 2–19). Mulitlingual Matters.

de Bot, K., & Jaensch, C. (2015). What is special about L3 processing? *Bilingualism: Language and Cognition*, *18*(2), 130–144. https://doi.org/10.1017/S1366728913000448

Deutsches Schulamt (2009). *Rahmenrichtlinien des Landes für die Festlegung der Curricula für die Grundschule und die Mittelschule an den autonomen deutschsprachigen Schulen in Südtirol*. Deutsches Schulamt.

Dipartimento Istruzione e Formazione italiana (2008). *Indicazioni provinciali per la definizione dei curriculi del primo ciclo d'istruzione della scuola in lingua italiana della Provincia Autonoma di Bolzano*. Provincia Autonoma di Bolzano.

Dirdal, H., Danbolt Drange, E.-M., Graedler, A.-L., Guldal, T. M., Hasund, I. K., Nacey, S. L., & Rørvik, S. (2017). Tracking Written Learner Language (TRAWL): A longitudinal corpus of Norwegian pupils' written texts in second/foreign languages. *Book of Abstracts of the 4th Learner Corpus Research Conference – LCR 2017 (Bolzano/Bozen, 5–7 October 2017)* (pp. 182–183).

Ehlich, K. (2005). Sprachaneignung und deren Feststellung bei Kindern mit und ohne Migrationshintergrund – Was man weiß, was man braucht, was man erwarten kann. In Bundesministerium für Bildung und Forschung (Ed.), *Bildungsreform Band 11: Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund* (pp. 11–63). BMBF.

Engel, D., & Stopfner, M. (2019). Communicative competence in the context of increasing diversity in South Tyrolean schools. In E. Vetter & U. Jessner (Eds.), *International research on multilingualism: Breaking with the monolingual perspective* (pp. 59–80). Springer Nature.

Flick, U. (2011). Das episodische Interview. In G. Oelerich & H.-U. Otto (Eds.), *Empirische Forschung und Soziale Arbeit: Ein Studienbuch* (pp. 273–280). VS Verlag für Sozialwissenschaften/Springer. https://doi.org/10.1007/978-3-531-92708-4_17

Gogolin, I. (2004). Lebensweltliche Mehrsprachigkeit. In K.-R. Bausch, F. G. Königs, & H.-J. Krumm (Eds.), *Mehrsprachigkeit im Fokus: Arbeitspapiere der 24. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* (pp. 55–61). Narr.

Gogolin, I., Lange, I., Hawighorst, B., Bainski, C., Heintze, A., Rutten, S., & Saalmann, W. (2011). *Durchgängige Sprachbildung: Qualitätsmerkmale für den Unterricht*. Waxmann.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 486–510). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022

Granger, S., & Paquot, M. (2017). *Core metadata for learner corpora. Draft 1.0*. (Unpublished manuscript). Université catholique de Louvain.

Grießhaber, W. (2006). Testen nichtdeutschsprachiger Kinder bei der Einschulung mit dem Verfahren der Profilanalyse – Konzeption und praktische Erfahrungen. In B. Ahrenholz, & E. Apeltauer (Eds.), *Zweitsprachenerwerb und curriculare Dimensionen. Empirische Untersuchungen zum Deutschlernen in Kindergarten und Grundschule* (pp. 73–90). Stauffenburg.

Grießhaber, W. (2010). Sprachkenntnisse einschätzen – Schreibfertigkeiten fördern. In C. Benholz, G. Kniffka, & E. Winters-Ohle (Eds.), *Fachliche und sprachliche Förderung von Schülern mit Migrationsgeschichte. Beiträge des Mercator-Symposions im Rahmen des 15. AILA-Weltkongresses „Mehrsprachigkeit: Herausforderungen und Chancen"* (pp. 115–135). Waxmann.

Grotjahn, R. (Ed.). (2014). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current Trends*. Lang. https://doi.org/10.3726/978-3-653-04578-9

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. Beltz.

Karges, K., Studer, T., & Wiedenkeller, E. (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, & L. Nicolas (Eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference* (pp. 137–165). Presses universitaires de Louvain.

Keßler, J.-U. (2006). *Englischerwerb im Anfangsunterricht diagnostizieren. Linguistische Profilanalysen am Übergang von der Primarstufe in die Sekundarstufe I*. Narr.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. Mc Enery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Addison Wesley Longman.

Lüdi, G. (2006). Multilingual repertoires and the consequences for linguistic theory. In K. Bührig, & J.D. ten Thije (Eds.), *Beyond misunderstanding: Linguistic analyses of intercultural communication* (pp. 11–42). John Benjamins. https://doi.org/10.1075/pbns.144.03lud

Okinina, N., Nicolas, L., & Lyding, V. (2018). Transc&Anno: A graphical tool for the transcription and on-the-fly annotation of handwritten documents. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 701–705).

Pienemann, M. (1998). *Language processing and second language development: Processability Theory*. John Benjamins. https://doi.org/10.1075/sibil.15

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. UTB. https://doi.org/10.36198/9783838534312

Reich, H. (2010). Sprachstanderhebung, ein- und mehrsprachig. In B. Ahrenholz, & I. Oomen-Welke (Eds.), *Deutsch als Zweitsprache* (pp. 420–429). Schneider Hohengehren.

Satrapi, M. (2000). *Persepolis: The story of a childhood* (Book 1). Pantheon Books

Satrapi, M. (2004). *Persepolis: The story of a return* (Book 2). Pantheon Books.

Tamaki, M., & Tamaki, J. (2014). *This one summer*. First Second Books.

Tan, S. (2006). *The arrival*. Hodder Children's Books.

Tono, Y., & Díez-Bedmar, M.B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics*, *19*(2), 163–177. https://doi.org/10.1075/ijcl.19.2.01ton

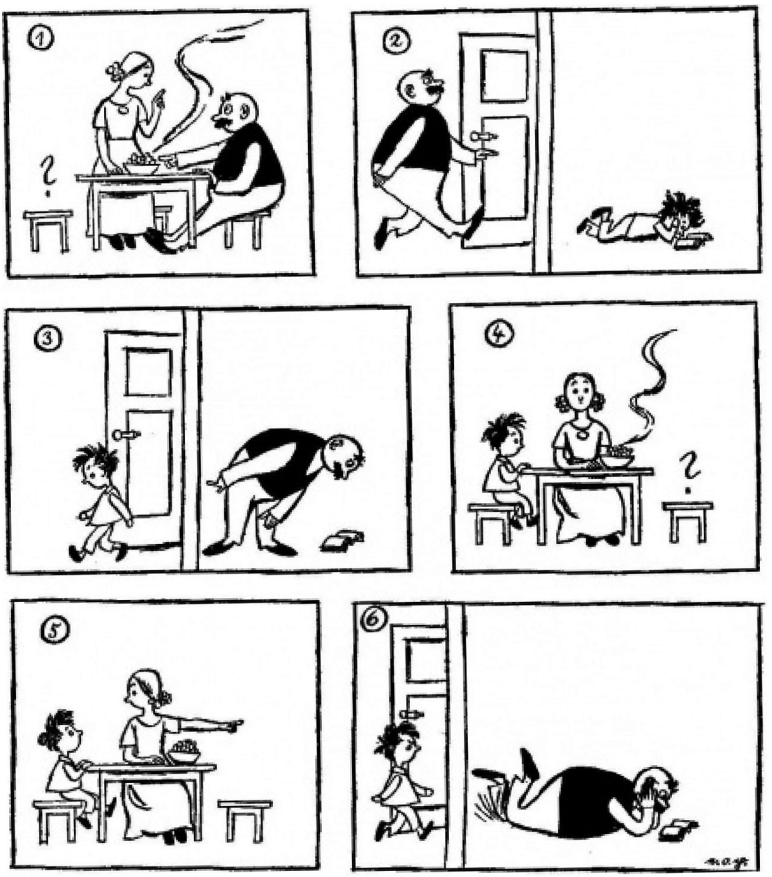
Voltmer, L. (2007). Languages in South Tyrol: Historical and legal aspects. In A. Abel, M. Stuflesser, & L. Voltmer (Eds.), *Aspects of multilingualism in European border regions: Insights and views from Alsace, Eastern Macedonia and Thrace, Lublin Voivodeship and South Tyrol* (pp. 201–219). Europäische Akademie Bozen.

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data*, *3*, Article Number 160018. https://doi.org/10.1038/sdata.2016.18

Appendix 1. Example of a picture story task (year 1, L3 English), E.O. Plauen: *Der Schmöker*

A picture story

What has happened here? Look at the pictures and write the story! Try to write something for every picture.

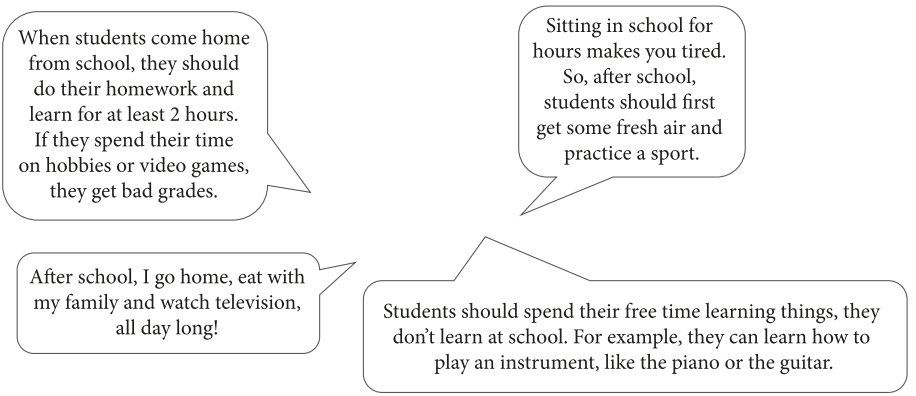


Appendix 2. Example of an opinion text task (year 2, L3 English)

What are your ideas?

Teenagers have to spend a big part of their day in school. But what do they do in the afternoon, when school lessons have ended? How can they use their free time in a good way?

Here are some ideas, how students at middle-school should spend their free time:



What do you think? How should students spend their free time after school?

- How much time should students spend on homework and learning? Why?
- How much time should students spend on homework and learning? Why?
- What do you do after school? How do you spend your free time and why?

We, the researchers of the EURAC, want your opinion on homework, hobbies and free time! You have 20 minutes to write your text.

Address for correspondence

Aivars Glaznieks
Eurac Research
Institute for Applied Linguistics
Viale Druso 1
39100 Bolzano
Italy
aivars.glaznieks@eurac.edu

Co-author information

Jennifer-Carmen Frey
Eurac Research
Institute for Applied Linguistics
JenniferCarmen.Frey@eurac.edu

Lorenzo Zanasi
Eurac Research
Institute for Applied Linguistics
lorenzo.zanasi@eurac.edu

Maria Stopfner
Eurac Research
Institute for Applied Linguistics
Maria.Stopfner@eurac.edu

Lionel Nicolas
Eurac Research
Institute for Applied Linguistics
lionel.nicolas@eurac.edu

Publication history

Date received: 9 March 2021
Date accepted: 14 June 2021