Nordic Journal of Language Teaching and Learning

Design and construction of the Tracking Written Learner Language (TRAWL) Corpus: A longitudinal and multilingual young learner corpus

Hildegunn Dirdal, University of Oslo
Ingrid Kristine Hasund, University of Agder
Eli-Marie D. Drange, University of Agder
Eva Thue Vold, University of Oslo
Elin Maria Berg, University of Agder

Abstract

This article describes the design and construction of the Tracking Written Learner Language (TRAWL) Corpus. The corpus combines several features that are all rare for learner corpora: it is longitudinal, following individual pupils over several years; it has data from young learners from school years 5 to 13 (ages 10–18); it is multilingual, containing learners' texts in several L3s (French, German and Spanish), L2 English and L1 Norwegian; and it includes teacher comments on a number of the texts. In addition, some of the texts exist in both a first and a second revised version, all tied to a rich set of meta-data. Not only does such a corpus offer new possibilities for research on language acquisition in general; it can also be used to provide valuable insights for teachers, teacher training and policymaking within the national context of Norway. In this article, we describe the design of the TRAWL Corpus and outline its uses and benefits for the research community. We also describe the compilation process in the hope that it may inspire and enable others to build similar corpora for their own national contexts.

Keywords: Longitudinal written learner corpus, young learners, Norwegian L1, English L2, French L3, German L3, Spanish L3.

Nordic Journal of Language Teaching and Learning, Vol. 10, No. 2, 2022. ISSN: 2703-8629
https://doi.org/10.46364/njltl.v10i2.1005

1. Introduction

Over the last decades, learner corpora, defined as “electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria” (Granger et al., 2015, p. 1), have become increasingly important in SLA research. They provide researchers with large amounts of learner language data and offer a plethora of possibilities for analysis using computerized tools. Along with the rapid technological development, the field has seen a steady growth in corpus construction, resulting in more and larger corpora, with increasingly advanced tools for automated analysis.

Despite these important advances, there are still issues that need to be addressed in the field of corpus compilation: firstly, few learner corpora are truly longitudinal, following the same individuals over time (Meunier, 2015, pp. 381–382; Gilquin, 2015, p. 14), despite the fact that development over time is central to the study of language acquisition and despite findings that individual trajectories are not necessarily the same as those found when comparing groups at different proficiency levels. Only 17 of the 198 learner corpora listed by the Centre for English Corpus Linguistics (2022) are described as longitudinal.

Secondly, there are few corpora with data from young learners and beginners, as researchers have found it easier to collect data from university students (Gilquin, 2015, p. 28; Osborne, 2015, p. 352). The scarcity of corpora with data from beginners is a problem in the study of language development. Corpora with the combination of longitudinal and young-learner data are particularly rare. Most of the longitudinal corpora listed by the Centre for English Corpus Linguistics have data from university students/young adults.

A third drawback is the lack of L1 corpora that are comparable to existing learner corpora. Cross-linguistic influence has been found in all areas of additional language acquisition (see e.g. Jarvis & Pavlenko, 2008), and most corpus compilers recognize the importance of recording metadata about the learners’ first languages (L1s). However, few corpora contain L1 texts from equivalent populations, not to speak of L1 texts from the same learners. There is also a lack of learner corpora that can allow comparisons across second and third languages, so-called “multilingual mono-L1 corpora” (Gilquin, 2015, p. 29), i.e. corpora with texts in several second/third languages produced by learners with the same L1. Such corpora are important in order to study differences in transfer effects depending on the target language and on the effects of knowing more than one language before learning an additional one (Osborne, 2015, p. 352).

Finally, learner corpora seldom include feedback from teachers on the texts. Obviously, this requires that the collected texts are from an educational context, whereas many learner corpora have texts written in reaction to prompts given by the corpus compilers. Writing instruction, including feedback, is important for learners' writing development (David & Doquet, 2016), and the inclusion of feedback with learner texts would therefore provide valuable data.

The TRAWL Corpus described in this article is designed to address all of these needs. It is a longitudinal corpus with data from young learners in primary and secondary school (ages 10–18) in Norway, and contains texts written in L1 Norwegian, L2 English and L3 French, German and Spanish – the most commonly taught L3s in Norwegian schools. The corpus also contains teacher feedback on some of the texts. The only other corpus with similar features that we are aware of is the LEONIDE Corpus from northern Italy (Glaznieks et al., 2022), a longitudinal corpus with L1 German and Italian, L2 Italian and German and L3 English data from lower secondary school.[1] Apart from the specific languages it contains and their status as L1, L2 or L3, the TRAWL Corpus differs from LEONIDE in the types of texts that are collected and the inclusion of teacher feedback. LEONIDE contains texts written in reaction to prompts given by the researchers to elicit one argumentative and one narrative text in each school year. In the TRAWL project, we collect the texts that the students write as part of regular schoolwork. This gives the corpus a less stringent design, but a larger range of text types that are authentic for the school situation and that can be used to study the kind of writing that actually goes on in Norwegian schools.

Corpora such as LEONIDE and the TRAWL Corpus are important for the study of language acquisition in general, but are also important in their local contexts and can be used to address questions of interest to teacher education and education policy. We therefore believe that the development of similar corpora would be of benefit in other countries and hope that the description of the design and compilation of the TRAWL Corpus can be an inspiration.

In the remainder of the article, we describe the participants that have contributed to the TRAWL Corpus (Section 2), the type of material that we have collected from them (Section 3), the processing of the texts, including anonymization and annotation (Section 4), the online search interface (Section 5) and the first version of the online corpus (Section 6). Finally, we describe the potential

[1] The Swiss Learner Corpus (SWIKO) is also a multilingual corpus with data from lower secondary school, but is not longitudinal (Research Centre on Multilingualism, 2022).

of the corpus and give examples from the work that has been done on TRAWL Corpus material so far (Section 7).

2. Participants

The TRAWL research group has collected data from Norwegian primary and secondary schools in different parts of the country since 2015[2]. In some cases, researchers or research assistants have visited school classes to inform students about the project; in other cases, they have informed teachers, who have recruited students from their classes themselves. Participation is based on informed consent, and students and teachers have received detailed information letters. In cases where students are below 16 years of age, parents are also asked to sign a consent form. Ethical approval has been granted from the Norwegian Centre for Research Data (NSD, now part of Sikt).

As Gilquin (2015, p. 18) notes, voluntary participation may lead to a set of participants that is not completely representative of the population we want to study, as it might favour self-confident, motivated and good students. The fact that we did not ask the students to do anything extra, but only to give us access to what they had written or would write for class anyway might have reduced this potential weakness. The students were informed that we were interested in the entire range of proficiency levels and that their texts would be anonymized upon collection. Sometimes entire classes decided to participate, and, judging from the grades given on the assignments, a large array of abilities are represented.

Students in Norwegian schools have English classes from year 1, but do not write much in the beginning. Both in the previous curriculum and the one implemented in 2020, the educational aims for the end of year 2 say that the students should "experiment with writing … simple sentences" (as well as words and phrases), whereas those for year 4 say that they should be able to write short/simple texts (Ministry of Education and Research, 2013; 2019). We therefore decided to collect English texts from years 5–13. In year 8 (the first year of lower secondary school), a large majority of students choose a second foreign language (79% at the start of our data collection and 74% in the school year of 2020–2022), primarily French, German or Spanish (The Norwegian National Centre for Foreign Languages in Education, 2016; 2022). At this age, writing development happens more quickly and we thus collect French, German and Spanish texts from years 8–

[2] See the editorial of this special issue for more information about the story of the project and for information about other corpora developed in the Norwegian context.

13. Collection has started at different levels for different students, and we have collected data from them for as long as possible, with the shortest period being one year and the longest currently being four years. The majority of students have contributed data over the course of lower secondary school or over the course of the first two years of upper secondary school.

In most cases, the texts in the different languages have been collected from different sets of students, with a few exceptions: Our L1 Norwegian material has been collected from a subset of the students from whom we have collected English texts, and in a few cases we also have L3 texts from the same students. Some of the TRAWL project members are currently involved in collecting a new set of texts consisting of L1, L2 and L3 data from the same students through the project MULTIWRITE, funded by the Norwegian Research Council (NFR). This data set will be incorporated into the TRAWL Corpus at a later stage.

Our use of the terms L1, L2 and L3 above to refer to Norwegian, English and French/German/Spanish, respectively, is based on the order and level at which these languages are taught in Norwegian schools. Norwegian is the language of schooling, English the first additional language introduced in school and French, German and Spanish second additional languages. However, the students may have learnt other languages elsewhere. All the students who agreed to contribute to the corpus were asked to fill in a questionnaire on language knowledge and use. Around 15% of the participants listed other L1s than Norwegian. Searches in all sub-corpora can be filtered by the L1s listed by students (see also Section 5). To avoid the possibility that individuals may be recognized, rare L1s have been collapsed into the category "other".

It is not only the L1 that can affect the learning of additional languages, but any other language that a learner knows (Jarvis & Pavlenko, 2008, pp. 21–22). In the questionnaire, the students were therefore asked to supply information about other languages they know – whether they can speak, understand, read or write these languages. The students contributing English, French, German and Spanish texts were also asked to specify how often they used the relevant language outside of school by indicating the number of hours spent on different activities. Such information is often missing in learner corpora (Gilquin, 2015, pp. 30–31). Table 1 lists the metadata that are recorded about the students, and the questionnaire can be found in the appendix.

Table 1. Student metadata in the TRAWL Corpus

- Education program (for upper secondary school)
- Study level (for L3s in upper secondary school)[3]
- Gender
- L1
- Parents'/guardians' L1s
- What languages, in addition to the L1, that they can 1) read, 2) write, 3) speak, 4) understand
- Whether they have attended schools with other teaching languages than Norwegian
- If so, which language(s) and in what school year
- When they started having English/French/German/Spanish classes
- Whether they have lived in an English/French/German/Spanish-speaking country, and for how long
- How many hours per week (apart from school and homework) they spend on the following activities:
  - reading English/French/German/Spanish on the Internet
  - playing computer games using English/French/German/Spanish
  - chatting/writing emails/text messages in English/French/German/Spanish
  - talking with someone in English/French/German/Spanish
  - watching series/films with English/French/German/Spanish speech and Norwegian subtitles
  - watching series/films with English/French/German/Spanish speech and without Norwegian subtitles
  - listening to audiobooks/radio programmes/podcasts etc., with English/French/German/Spanish speech
  - other (the students are asked to specify)

[3] Students who have not studied a second additional language in lower secondary school, may start in upper secondary. It is also possible to choose a new additional language at this stage. Language classes in upper secondary school are therefore given at two levels: "Level 1" for those who start in upper secondary and "Level 2" for those who have already studied the language in lower secondary school.

We also collected information about the students' birth country and whether and when they had lived in other countries than Norway. To ensure anonymity, we have decided not to include this information in the online corpus. All the other student metadata are connected to each text and can be accessed in the online corpus, and some of the categories are available for filtering searches (see Section 5).

The students contributing texts to the corpus have each received a unique student code starting with P (for "pupil") and continuing with a five-digit number, e.g. P01000 or P60454. This ensures that individual students can be studied longitudinally. The student codes are searchable in the online corpus, which means that it is possible to find the texts produced by the same student in different school years and in different languages.

The teachers of the students who have contributed to the corpus have consented to the collection of texts from their students. Some of the teachers have also consented to the inclusion of their feedback in the corpus. These teachers have been assigned unique teacher codes starting with a T (for "teacher") and continuing with a four-digit number, e.g. T0002. The teacher codes are also searchable, which makes it possible for example to compare feedback from the same teachers over different school years or to students at different levels of achievement.

At the point of publishing this article, more than 1200 students have contributed material to the corpus. Section 5 below gives details about the number of students and texts included in the first online version.

3. Texts

The participating students have allowed us to collect all the texts that they write for the relevant language classes (although they have the option to withhold particular texts). Students typically write more and longer texts for L1 than L2 classes and the shortest for L3 classes, and they write longer texts in the later school years. In addition, individual teachers may have different practices concerning the use of written hand-ins, and students may be absent on some writing occasions. This means that the volume of material and the density of data collection points may vary between sub-corpora, between school years and between individual students.

We decided to collect texts written for regular schoolwork instead of responses to pre-set tasks. The main drawback is that comparisons across levels may be more difficult to carry out and may require a careful selection of the texts to include in each study, depending on the focus of the investigation. However, there are several benefits to collecting regular schoolwork. Some are practical: It is easier to recruit participants if they do not have to do additional work, and by collecting schoolwork, we do not have to arrange writing sessions with the schools and teachers. Because such sessions would take away time from other activities, it would be difficult to arrange them very often, and the collection of schoolwork thus allows us to get denser data collection points. But there are also other benefits. The types of data included in our corpus give a picture not only of the students' language abilities, but also of the kind of writing that actually goes on in Norwegian schools and the feedback that teachers give on this writing. Tasks and text types are not uniform throughout the school years, and the changes are part of the context for the language development that the students go through. Looking at the actual texts that they write is therefore important when explaining the development that takes place. In the Norwegian context, the results of studies on such texts are also more immediately useful for teachers, teacher educators and policy planners.

For each text, we included copies of the task instructions when these were available, and as close a description of the task as we could manage when the original instructions were missing. Each task has received a unique four-letter task code, e.g. ARWO, VIAC, LANG. Unique task codes give the possibility of finding texts written in response to the same tasks. Some tasks are used more widely than in one class: sometimes by several classes in the same school and sometimes by several schools – for example national exams that have later been used as mock exams. Some task instructions include several different tasks and sometimes students have a choice between a range of tasks. This means that the texts with the same task code may not always be on exactly the same topics or contain the same text types. We have still chosen to give them the same codes, so that they can be linked to the original task descriptions.

Because we collect authentic texts written at/for school, a range of text types is represented in the corpus. A subset of the English texts have been annotated for (and are searchable by) text type/genre, based on a typology described in Hasund (2022). Currently, this subset consists of all the English texts written by one class throughout lower secondary school (years 8–10). Hasund and Hasselgård (2022), who studied this set of texts, found that the six main genres/text types of argumentative, expository, descriptive, dialogic, reflective and narrative texts were all represented, with a predominance of narrative texts in the first year and argumentative and expository texts in the last year.

Each text is connected to metadata about date, format, task type, version, whether it comes with teacher comments and, if so, the teacher code for the teacher that commented on the text, as specified in Table 2. This information can be accessed for each text in the online corpus, and all of the categories except for the date of writing can be used to filter searchers.

Table 2. Text metadata in the TRAWL Corpus

Date (when the text was handed in)
Text format (handwritten or electronic)
Task type (classroom writing, homework or test)
Version (only version, first version or second version)
Teacher comments/feedback (yes or no)
Teacher code (if there are comments/feedback)

The filename of each text includes the pupil code, the school year, the task code and the version of the text. For example, the filename “P0100_Y10_ARWO_V0” shows that the text is written by student P0100 in year 10, that it is a response to the task ARWO, and that it is the only version of the text.

4. Processing of the texts

Each of the texts that the students contribute to the corpus is processed to create three or four versions (depending on whether teacher comments are included):

- a PDF file of the anonymized original student text
- a PDF file of the anonymized original with teacher comments (if included)
- an annotated XML file of the anonymized student text
- an annotated XML file of the anonymized student text with corrected spelling

The first step in the process is the anonymization of the texts and the feedback by replacing all names and numbers that may lead to the identification of students or teachers with designated codes (NAME_PERSON1_F, NAME_PLACE1, NUMBER_PHONE, etc.) and by removing author information saved with the document, using an ordinary Word editor. For electronic texts, this happens before we create PDF files. If teachers have not consented to the use of their feedback, any feedback included in the file is deleted; if they have consented to such use, we save one PDF file with and one without feedback. For handwritten texts, names or numbers that may identify the student (or the teacher who has given feedback) are covered to become illegible before the texts are scanned to create PDF files, whereas the designated codes described above are used in the anonymization of transcriptions of the texts, reproducing exactly the students' own spelling and grammar. Unclear or ambiguous words or letters are indicated by the use of square brackets in the transcriptions.

The anonymized transcriptions/electronic texts are then annotated for headings (at different levels), sentence divisions, paragraph divisions, italics, boldface, source lists, quotes, mentioned items, pictures/drawings/figures, tables and lists following TEI conventions (TEI Consortium, 2009). This annotation happens in several stages. In the first stages, we use Word macros originally developed for the BAWE Corpus and later adapted to the VESPA Corpus (Paquot et al., 2013). The macros are written in Visual Basic and have a graphical user interface that guides the annotator through the various steps of the process (Ebeling & Heuboeck, 2007, pp. 250–252). In the final stage, we use a Perl script also originally from the above-mentioned corpus projects, but adapted to the needs of the TRAWL Corpus by Jarle Ebeling. The Perl script implements TEI XML structure, converts the pseudo-tags from the macros to XML TEI tags, marks and numbers sentences and paragraphs, marks footnotes and endnotes, normalizes hyphens, dashes, quotes, etc., imports comments created in separate files during the use of the macros, and imports the student and text metadata from Excel files into the headers of the XML files (see Paquot et al., 2015, p. 23).

Finally, an XML version with corrected spelling is created so that searches may result in hits even if students have misspelled words. These are created by opening the XML files in the Oxygen XML Editor and correcting misspelt words. We have chosen not to correct word choice mistakes or grammar errors because this will involve more interpretation and introduce a level of analysis into the data.

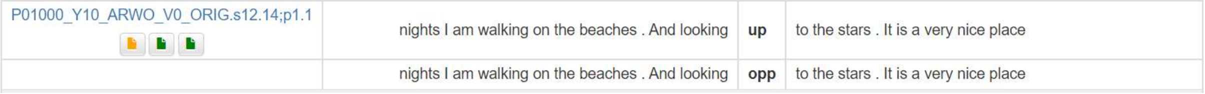
The processes of anonymization and annotation are detailed in the VESPA manual (Paquot et al., 2015) and the TRAWL Corpus manual (Dirdal, 2022a), which can both be accessed via the TRAWL Corpus website (see section 5).

5. Online search interface and POS tagging

The online version of the TRAWL Corpus is searchable through the Glossa search interface developed by the Text Laboratory at the University of Oslo (Nøklestad et al., 2017) and can be accessed via the TRAWL Corpus website at https://tekstlab.uio.no/trawl from 1 February 2023. The Text Laboratory has adapted the Glossa interface to the corpus, and in the process of entering the data into the online corpus, they conduct parts-of-speech (POS) tagging using the TreeTagger[4] (Schmid, 1994; 1995) on the English, French, German and Spanish texts and the Oslo-Bergen Tagger[5] (Johannessen et al., 2012) on the Norwegian texts. The corrected versions of the texts are used for more reliable results, and the original and corrected versions of the texts are aligned at word level. It should be noted that the different taggers follow different principles in certain cases. Researchers should therefore familiarize themselves with the tagging before comparing search results from the different languages of the corpus.

The Glossa search interface has three search options: simple searches, extended searches and CQP queries (Corpus Query Protocol). Extended searches give the possibility to specify part-of-speech, do lemma searches, search for the start/middle/end of words, search for sentence-initial words or search in the original rather than the corrected versions. Many of the metadata categories can be used to filter searches, as described in Sections 2 and 3. The full set of metadata for each text can be accessed by clicking on the corpus identifier at the left of the KWIC (Key Word In Context) list (see Figure 1).

The TRAWL Corpus KWIC lists show both the original and corrected versions of the student texts, as illustrated in Figure 1. As the example in the figure shows, the corrected version ensures that a search for *up* results in a hit even for a text where the word has been misspelled, and in the lower row, we see that this student used the spelling *opp*.



P01000_Y10_ARWO_V0_ORIG.s12.14;p1.1	nights I am walking on the beaches . And looking	**up**	to the stars . It is a very nice place
	nights I am walking on the beaches . And looking	**opp**	to the stars . It is a very nice place

Figure 1. KWIC list with linked original and corrected version

[4] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[5] tekstlab.uio.no/obt-ny/English/index.html

The identifier at the left of the KWIC list shows the student number, the school year, the task code and the version (V0 means an only version), as well as the sentence and paragraph where the keyword occurs. Below the identifier are icons that can be clicked to access the task description, the PDF showing the formatting of the original text and the PDF with teacher comments (if comments have been submitted to the corpus).

6. The first version of the TRAWL Corpus

The first version of the TRAWL Corpus (released 1 February 2023) consists of 1664 texts written by 216 students, 988 of them with teacher feedback. The texts are organized into five sub-corpora, one for each of the languages represented. Links to all the sub-corpora can be found on the main page of the corpus at https://tekstlab.uio.no/trawl.

For the first version of the corpus, we have given priority to the school years for which we have the most texts so far, which are years 8–10 for Norwegian, years 8–11 for English (year 11 is the last year for which English is an obligatory subject), and years 11–12 for French, German and Spanish. The students represented in the Norwegian sub-corpus are a subset of the ones who have contributed English texts. Figure 2 shows the composition of the corpus.

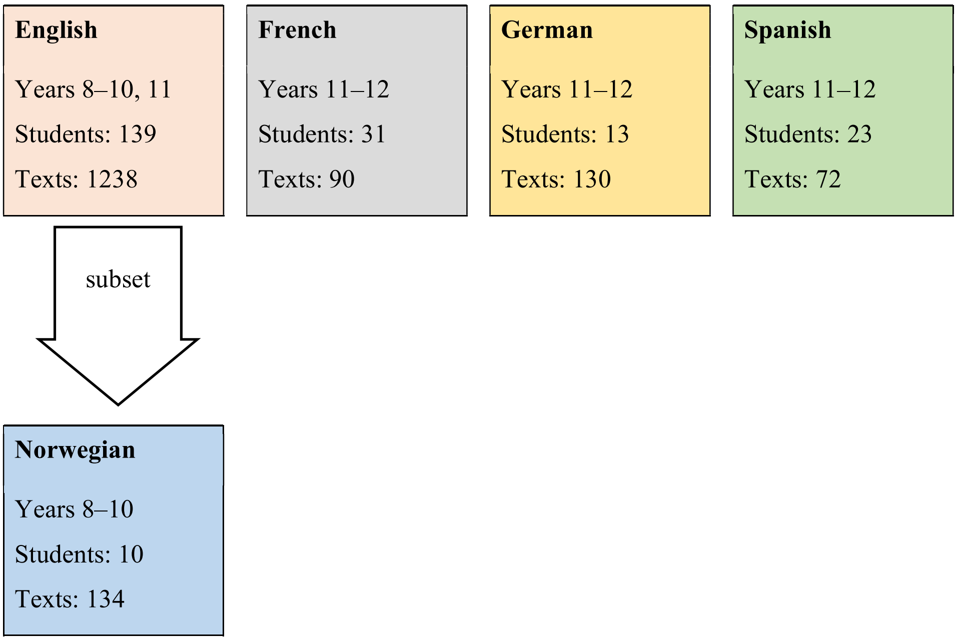


Figure 2. The composition of the TRAWL Corpus (first version, 2023)

The first version of the corpus is still relatively small. The Norwegian and German sub-corpora contain data from only 10 and 13 students, respectively. On the other hand, they contain on average 13 and 10 texts per student (4–6 per year), which makes it possible to study individual development in detail.

The largest sub-corpus is the English one with 1238 texts written by altogether 139 students. Most of the material (1120 texts) is from lower secondary school, i.e. years 8–10, and consists of longitudinal data from the same students, with around 4–6 texts per year. It has been possible to collect texts from a few of these students even in year 11, but otherwise the texts from that year come from a separate set of students. All the other sub-corpora in version 1 of the corpus have data from the same students for the years included.

Only a few of the texts in the Norwegian and the L3 sub-corpora come with teacher feedback so far (Norwegian: 19 texts, French: 33 texts, German: 7 texts). The English sub-corpus contains 929 texts with feedback from a total of 13 teachers, and 129 texts exist in two versions.

An updated version of the corpus is planned for the summer of 2023, and a further update in early 2024. More texts will then be added to the corpus.

7. Use and potential

The nature of the TRAWL Corpus makes it a valuable source of material both for studies related to language acquisition and the nature of learner language and for studies related to school writing, feedback and assessment. Even before the online release of the corpus, the material collected for it has been used in several research studies. These studies illustrate both the wide range of topics that can be investigated and the way that material can be selected to form the basis for different kinds of investigations.

Since the corpus contains a large number of texts written as ordinary schoolwork, researchers must take care to select students and texts to suit their research questions and control important variables. For example, Hasund and Hasselgård (2022), who looked at features of writer/reader visibility, selected only argumentative and expository writing from the subset of texts annotated for genre (see Section 3). Dirdal (2022b), who investigated complexity development, selected only certain text types in order to ensure that the results were not skewed by the varying proportions of genres over the school years[6]. Berg (2020) and Vold (2021) selected texts on which there was teacher feedback since they were interested in the effect of grammar correction and uptake of feedback, respectively.

One of the strengths of the TRAWL Corpus is that it contains truly longitudinal data, and many of the studies conducted therefore look at the development of the same individuals over time, whether it is for the whole period of data collection, or a shorter part of that period. The volume of material from individual students means that it is even possible to perform case studies where the development of particular students can be studied in detail (e.g. Berg, 2020; Dirdal, 2021, 2022b).

In future versions of the TRAWL Corpus, data from a larger range of school years will be added. Although the main bulk of the longitudinal data that has been collected spans different two- or three-year periods, it is possible to use the TRAWL Corpus data pseudo-longitudinally, and thereby get a longer time frame. Nacey (2019) did this when she studied metaphor use in a selected number of English texts from different students in years 5 to 13. A pseudo-longitudinal design can also be used to compare the young learners in the TRAWL Corpus with older and more advanced learners from the L1 Norwegian components of the International Corpus of Learner English (ICLE; Granger et al., 2020), the Varieties of English for Specific Purposes dAtabase (VESPA; Paquot et al., 2013) or the Elenor Corpus (Español Lengua Extranjera en Noruega; Department of Literature, Area Studies and European Languages, 2022), as is done in Evang's (2019) study of phraseological development and Hasund and Hasselgård's (2022) study of writer/reader visibility features.[7] It is of course also possible to conduct cross-sectional studies on the TRAWL Corpus material when development is not the main focus, as is done in Auensen (2019), an investigation into the correlation between lexical richness and the assessment of mock exams from year 9.

Some studies have combined the TRAWL Corpus material with other sources of data in a mixed methods design. This is the case with Høegh-Omdal (2018), where an analysis of argumentative texts from year 10 is combined with interview and survey data from teachers to investigate whether year 10 students are prepared for the argumentative texts they will have to write in upper secondary school. Kolsvik (2019) combined a corpus study of L2 English texts from the TRAWL Corpus and

[6] Since only a few of the texts in the corpus are currently annotated for genre, genre or text type selections must at present be done by investigating individual texts or tasks. Searches can then be filtered to include the chosen texts/tasks/students.

[7] Both studies use a combination of pseudo-longitudinal and truly longitudinal design.

ICLE with student and teacher surveys in order to study the use of American features in the vocabulary, spelling and pronunciation of Norwegian students. Because of their involvement in the corpus compilation, some authors have had the chance to collect further data from the same students that have contributed to the corpus. Dasic (2019) conducted interviews with students about their attitudes to language learning through gaming and investigated the correlation between their attitudes and their grades and the lexical richness of their texts, Garshol (2019) followed up an analysis of collected texts with an intervention study that aimed to improve the accuracy of English agreement marking, and Woldsnes and Vold (2018) tested French students' explicit knowledge of agreement rules and investigated to what extent they utilized such knowledge in their free writing. Although the users of the online version do not have the possibility to collect further data from the students that have contributed to the corpus, there is a wealth of meta-data in the corpus that may be combined with the textual data to form the basis for interesting studies. For example, the meta-data about extramural language use can be utilized in combination with analyses of student writing to explore how it correlates with performance and development. This kind of meta-data is scarce in learner corpora and will hopefully form the basis for new and original studies.

Another aspect of the corpus that is yet to be exploited is the fact that it contains L2 and L3 data in different languages from students with the same L1. The data may inform studies of the effect not only of the L1, but also the L2 on an L3. Comparisons may be made between L2 and L3 learning and use across different L3s, and these comparisons may concern a range of issues, such as language development, tasks that are given and feedback practices.

The research that can be conducted on the TRAWL Corpus material can feed into teaching and teacher education, but the corpus can also be used more directly in these contexts. The TRAWL Corpus material is from young learners, which makes it possible for teacher students to study the language production and development of students at the level at which they will be teaching, rather than the language of university students like themselves. Several of the studies mentioned above are masters' theses, many of them by teacher students. A large majority of the researchers in the TRAWL Corpus group are involved in teacher training in English and/or foreign languages, and TRAWL Corpus data are being used in teaching and examination work in teacher training.

Now that the corpus is being released online, we look forward to seeing its rich data being explored further, both for research and in teaching.

References

Auensen, M. (2019). *The correlation between lexical richness and Norwegian lower secondary school EFL teachers' assessment of written compositions* [Master's thesis, University of Agder]. Agder University Research Archive (AURA). https://uia.brage.unit.no/uia-xmlui/handle/11250/2617582

Berg, E. M. (2020). *Written corrective feedback and the development of L2 learner language. A longitudinal study of lower secondary EFL writing in Norway* [Master's thesis, University of Agder]. Agder University Research Archive (AURA). https://uia.brage.unit.no/uia-xmlui/handle/11250/2726433

Centre for English Corpus Linguistics. (2022). *Learner corpora around the world*. Université catholique de Louvain. https://uclovain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Department of Literature, Area Studies and European Languages. (2022*). The Elenor Corpus: Español Lengua Extranjera en Noruega*. University of Oslo. https://www.hf.uio.no/ilos/tjenester/kunnskap/sprak/nettsprak/spansk/lesesal/elenor/elenor.html

Dasic, J. (2019). *Acquiring English vocabulary through virtual worlds: Exploring the connection between Norwegian EFL lower secondary pupils' gaming habits, essay grades and written lexical richness in light of their attitudes* [Master's thesis, University of Agder]. Agder University Research Archive (AURA). https://uia.brage.unit.no/uia-xmlui/handle/11250/2617514

Dirdal, H. (2021). L2 development of *-ing* clauses: A longitudinal study of Norwegian learners. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond concordance lines: Applications of corpora in language education* (pp. 75–96). John Benjamins. https://doi.org/10.1075/scl.102.04dir

Dirdal, H. (2022a). Instruks for assistenter i TRAWL-prosjektet/Instructions to assistants in the TRAWL-project.

Dirdal, H. (2022b). Development of L2 writing complexity: Clause types, L1 influence and individual differences. In A. Leńko-Szymańska & S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research* (pp. 81–114). John Benjamins.

Ebeling, S. O., & Heuboeck, A. (2007). Encoding document information in a corpus of student writing: The British Academic Written English corpus. *Corpora*, *2*(2), 241–256.

Evang, K. H. S. Ø. (2019). *From 'car motor' to 'fishing boat': Tracking intermediate learners' phraseological development by use of association measures* [Master's thesis, University of Oslo]. DUO Research Archive. https://www.duo.uio.no/handle/10852/73232

Garshol, L. (2019). *I just doesn't know: Agreement errors in English texts by Norwegian L2 learners: Causes and remedies* [Doctoral dissertation, University of Agder]. Agder University Research Archive (AURA). https://uia.brage.unit.no/uia-xmlui/handle/11250/2589044

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (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, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research – past, present and future. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1–5). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Hasund, I.K. (2022). Genres in young learner L2 English writing: A genre typology for the TRAWL (Tracking Written Learner Language) corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 242-271. https://doi.org/10.46364/njltl.v10i2.939

Hasund, I.K., & Hasselgård, H. (2022). Writer/reader visibility in young learner writing: A study of the TRAWL corpus of lower secondary school texts. *Journal of Writing Research*, *13*(3), 447–472. https://doi.org/10.17239/jowr-2022.13.03.04

Høegh-Omdal, L. (2018). *English argumentative writing in Norwegian lower secondary school: Are year 10 lower secondary students sufficiently prepared for L2 argumentative writing in upper secondary?* [Master's thesis, University of Agder]. Agder University Research Archive (AURA). https://uia.brage.unit.no/uia-xmlui/handle/11250/2564581

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Johannessen, J. B., Hagen, K., Lynum A., and Nøklestad, A. (2012). OBT+stat: A combined rule-based and statistical tagger. In G. Andersen (Ed.), *Exploring newspaper language. Corpus compilation and research based on the Norwegian Newspaper Corpus* (pp. 51–65). John Benjamins.

Kolsvik, S. G. (2019). *Moving toward(s) Americanization: A study of the use of and attitudes toward American spelling, vocabulary and pronunciation among Norwegian students and teachers* [Master's thesis, University of Louvain and University of Oslo]. DIAL.mem. https://dial.uclouvain.be/memoire/ucl/en/object/thesis%3A18891

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.017

Ministry of Education and Research. (2013). *English subject curriculum (ENG1-03).* www.udir.no/kl06/ENG1-03?lplang=http://data.udir.no/kl06/eng

Ministry of Education and Research. (2019). *English subject curriculum (ENG1-04)*. https://www.udir.no/lk20/eng01-04?lang=eng

Nacey, S. (2019). Development of L2 metaphorical production. In A. M. Piquer-Píriz & R. Alejo-González (Eds.), *Metaphor in foreign language instruction* (pp. 173–198). De Gruyter Mouton. https://doi.org/10.1515/9783110630367-009

Norwegian National Centre for Foreign Languages in Education. (2016). *Elevenes valg av fremmedspråk på ungdomstrinnet for skoleåret 15/16 og utviklingen de siste ti årene*. Notat 1/2016. https://www.hiof.no/fss/sprakvalg/fagvalgstatistikk/20160113_notat_ungdomstrinn_15-16.pdf

Norwegian National Centre for Foreign Languages in Education. (2022). *Elevane sitt val av framandspråk på ungdomsskulen 2021–2022*. Notat 01/2022. https://www.hiof.no/fss/sprakvalg/fagvalgstatistikk/elevane-sine-val-av-framandsprak-i-ungdomsskulen-21-22.pdf

Nøklestad, A., Hagen, K., Johannessen, J.B., Kosek, M., & Priestley, J. (2017). A modernised version of the Glossa corpus search system. In J. Tiedemann & N. Tahmasebi (Eds.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 251–254). Assoc. for Computational Linguistics. https://www.duo.uio.no/handle/10852/59550

Osborne, J. (2015). Transfer and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 333–356). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.015

Paquot, M., Hasselgård, H. & Ebeling, S. O. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead*. Corpora and Language in Use – Proceedings 1. Presses universitaires de Louvain. https://doi.org/10.1017/cbo9781139649414.010

Paquot, M., Ebeling, S. O., Heuboeck, A., & Valentin, L. (2015). *The VESPA tagging manual*. Version 2.3. Centre for English Corpus Linguistics, Université catholique de Louvain. https://cdn.uclouvain.be/groups/cms-editors-cecl/VESPA_Manual_version2.3_0.pdf

Research Centre on Multilingualism. 2022. *Swiss Learner Corpus SWIKO*. University of Fribourg. https://centre-plurilinguisme.ch/en/research/swiss-learner-corpus-swiko

Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees.* Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.

Schmid, H. (1995). *Improvements in part-of-speech tagging with an application to German*. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

TEI Consortium (Eds.). (2009). *TEI P5: Guidelines for electronic text encoding and interchange*. Version 1.5.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Vold, E. T. (2021). Assessing writing in French-as-a-foreign-language: Teacher practices and learner uptake. *Languages*, *6*(4), 210. https://doi.org/10.3390/languages6040210

Woldsnes, A.-K. & Vold, E. T. (2018). L'interlangue écrite de lycéens norvégiens: l'usage de connaissances explicites de grammaire dans la production libre. *Synergies Pays Scandinaves*, 13, 105–117. https://gerflint.fr/Base/Paysscandinaves13/woldsnes_thue.pdf

Appendix 1. Questionnaire used to collect meta-data about the students

SPØRRESKJEMA

Etternavn:	Fornavn:	Landet du ble født i:
Ditt morsmål:	Morsmålet til forelder/foresatt 1:	Morsmålet til forelder/foresatt 2:

Kjønn: □ jente/kvinne □ gutt/mann □ annen kjønnsidentitet

Har du bodd i andre land enn Norge? **Hvilke** og **når**?

____

____(Fortsett på baksiden om nødvendig.)

Hvilke språk kan du i tillegg til morsmålet ditt? (Dette kan være språk du snakker hjemme eller med slektninger, språk du har lært på skolen eller språk du har lært på andre måter. Kryss av for om du kan **lese**, **skrive**, **snakke** eller **forstå** språket når du hører det.)

Språk	Lese	Skrive	Snakke	Forstå

Har du gått på skoler med et annet undervisningsspråk enn norsk (dvs. det språket læreren brukte i de fleste fagene)? Oppgi i så fall klassetrinn og språk nedenfor:

Klassetrinn	Undervisningsspråk

I hvilken klasse begynte du med engelskundervisning? ________________________________

Har du bodd i et engelsktalende land? Kryss av for perioden.

□ Nei, eller mindre enn 2 uker
□ 2-4 uker
□ 1-3 måneder

□ 4-6 måneder
□ 7-12 måneder
□ Mer enn ett år

Kryss av for omtrent hvor mange **timer i uka** (utenom skole og lekser) du bruker til å…	Over 10 timer	5-10 timer	1-4 timer	Opp mot 1 time	Ingen
Lese engelsk på internett					
Lese engelske bøker/aviser/blader					
Spille dataspill hvor du bruker engelsk					
Chatte/skrive e-post/SMS på engelsk					
Samtale muntlig med noen på engelsk					
Se serier/filmer med engelsk tale og norsk teksting					
Se serier/filmer med engelsk tale uten norsk teksting					
Høre på lydbøker/radioprogrammer/podcast e.l. med engelsk tale					
Annet (spesifiser)					

Benytt gjerne baksiden av arket dersom du har noe mer du ønsker å opplyse om.