Tracking Written Learner Language (TRAWL): A longitudinal corpus of Norwegian pupils' written texts in second/foreign languages

Hildegunn Dirdal1 , Eli-Marie Danbolt Drange2 , Anne-Line Graedler3 , Tale M. Guldal2 , Ingrid Kristine Hasund2 , Susan Lee Nacey3 , Sylvi Rørvik3 University of Oslo1 , University of Agder2 , Inland Norway University of Applied Sciences3 hildegunn.dirdal@ilos.uio.no, eli.m.drange@uia.no, anneline.graedler@inn.no, tale.guldal@uia.no, kristine.hasund@uia.no, susan.nacey@inn.no, sylvi.rorvik@inn.no

TRAWL is a research project where the primary objective is to explore and describe how Norwegian pupils develop writing skills in second and foreign languages throughout their education journey. Texts are being collected longitudinally from pupils at Norwegian schools and compiled into a searchable corpus, which will remain accessible as a resource for researchers, teachers and teacher educators after the end of the project. Our poster will present the TRAWL corpus and describe its design, the transcription and annotation of the texts, the research aims of the project group and other potential applications for the corpus.

Design:

Learner corpora containing data from the early stages of SLA are scarce (Tono et al. 2012: 8), as are longitudinal learner corpora with data collected from the same learners over time. Longitudinal data are essential since the time perspective is a crucial aspect of the language learning process and development in general (Granger 2012: 11). Also, more data from a Norwegian context are needed. The CORYL corpus contains error-tagged texts by Norwegian pupils in years 7, 10 and 11 in the National tests of English writing from 2004– 2005. It is, however, a small corpus, containing cross-sectional data only. English L2 learner texts have also been collected for Norwegian components of three international corpora initiated at the Centre for English Corpus Linguistics (CECL) in Belgium: the written corpora ICLE and VESPA, and the spoken corpus LINDSEI. These corpora are cross-sectional and contain texts written at university level.

The TRAWL corpus supplements these corpora with longitudinal data from younger learners: in years 5–13 for English and years 8–13 for French, German and Spanish, the most common foreign languages that pupils can select to study in addition to English. Data are collected from all school years from the start and will continue for at least three years to allow for both pseudo-longitudinal and truly longitudinal studies. Some pupils are also asked to contribute texts written in Norwegian to enable comparisons of L1 and L2 writing development. The collected texts have been written as part of the pupils' regular class work. 183 The corpus contains metadata describing pupils (age, gender, language background, etc.) and texts (format, task prompts, task conditions etc.), as well as teachers' written assessment.

Transcription and annotation:

At lower levels, most texts are written by hand. The first stage of data processing is therefore transcription of these texts, without any changes to spelling, grammar or punctuation. Since spelling variation makes automatic searches difficult, corrected versions will be linked up with the primary transcriptions. It will also be possible to view pdfversions of the (anonymized) original texts.

The texts are annotated using macros and Perl scripts originally created for the British Academic Written English (BAWE) corpus and adjusted for VESPA by Alois Heuboeck and further for TRAWL by Jarle Ebeling (see Ebeling and Heuboeck 2007, Paquot et al. 2015). The annotation follows the TEI

conventions and includes sentence, paragraph and various other text divisions, formatting, lists, tables, figures and quotes/mentioned items, as well as the metadata described above.

Research goals:

In addition to the primary objective of the TRAWL research project mentioned above, the secondary objectives are 1) to map grammatical, lexical and text coherence features that characterize learner language at various stages and age levels, 2) to research factors that may affect learner L2 development. The first stage of data collection will enable analysis of cross-sectional data by comparing texts from different levels. At the second stage the data will be genuinely longitudinal, and will contribute with unique empirical evidence for second language (L2) proficiency development.

Other applications:

The TRAWL corpus will remain accessible for researchers, teachers and teacher educators after the end of the project. In addition to providing data for further studies, including master's theses, the corpus will be useful in teacher training. In courses that use corpus data, students currently study language by learners at the same level as themselves. With TRAWL, they can investigate learner language from younger pupils as well. The corpus will also function as a useful source of examples in the development of teaching materials.

References:

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.

Granger, S. (2012). How to Use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.). Research Methods in Second Language Acquisition: A Practical Guide. London: Wiley-Blackwell, 7–29.

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.

Tono, Y., Kawaguchi, Y. & Minegishi, M. (2012). Developmental and Crosslinguistic Perspectives in Learner Corpus Research. Amsterdam: John Benjamins.