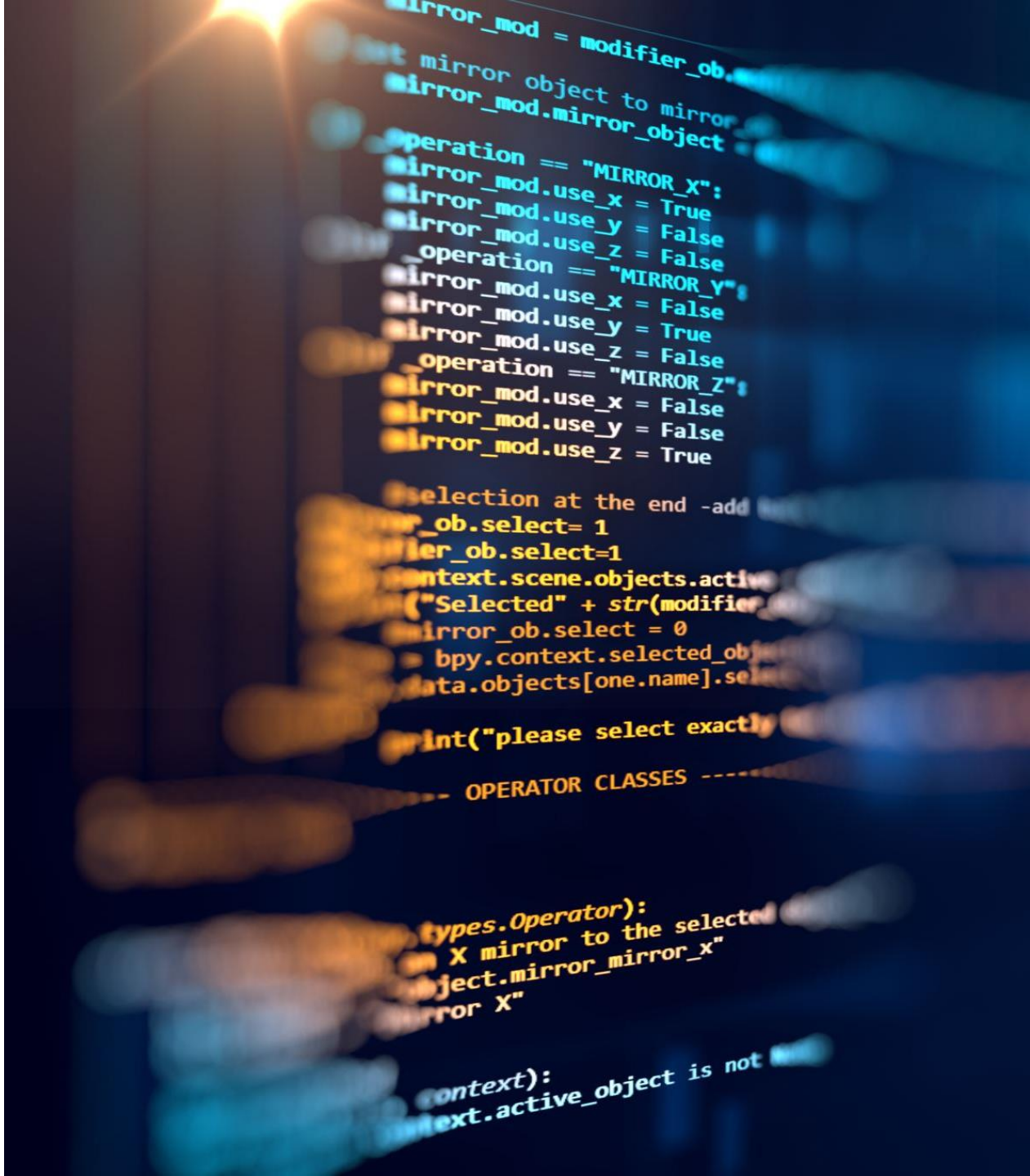
# Text selection, processing and metadata representation in learner corpora:

## Challenges of corpus compilation and consequences for corpus users

Hildegunn Dirdal, Stine H. Johansen (University of Oslo)
Philip Durrant (University of Exeter)

27.3.2024



# Aims of the paper

Based on our experience in compiling learner corpora with features that are less common:

- Young learners
- School writing
- Longitudinal

1) Discuss challenges we encountered in the creation of such corpora with respect to key aspects of corpus compilation: representativeness, metadata collection, and data processing
2) Explain what solutions were adopted and discuss consequences for subsequent corpus users

# Background: Alignment of data and research questions

- Data collection and processing must align with the research questions
- When we use already compiled corpus data, decisions of inclusion and processing have already been taken by the corpus compilers
- Their decisions are enforced upon corpus users, making it imperative that such decisions are made explicit (Ädel, 2020).
- Recent calls for more attention to such issues by both corpus compilers and corpus users (Egbert, Biber & Gray, 2022; Egbert, Larsson & Biber, 2020; Miller & Biber, 2015)

"In many cases, a corpus builder has constructed a corpus, and then numerous other researchers subsequently reuse that same corpus to investigate a wide range of different linguistic research questions. Thus, evaluating the adequacy of such corpora as the basis for investigating particular research questions becomes especially important in the subdiscipline of corpus linguistics."

(Egbert et al., 2022, p. 12)

# The Growth in Grammar (GiG) Corpus

(Durrant and Brenchley, 2018)

Created as part of the Growth in Grammar project

Aim: understand the writing development of children in primary and secondary schools in England; thereby inform teaching and curriculum design

Language: English

Texts written as part of regular schoolwork in various school subjects

Pseudo-longitudinal: texts from year 2, 4, 6, 9 and 11

# The Tracking Written Learner Language (TRAWL) Corpus

(Dirdal et al., 2022)

Aim: explore language learning and writing development in the languages most commonly taught in Norwegian schools

Languages: Norwegian, English, French, German and Spanish

Texts written as part of regular schoolwork in the language subjects

Longitudinal: texts from the same individuals over 1-4 years

Covers mainly years 8-12 (lower and parts of upper secondary school)

Data collection started in 2015, and texts are still added to the corpus

# Representativeness – background

A corpus is “a large and principled sample of texts designed to represent a target domain of language use” (Egbert, Biber & Gray, 2022, p. 7)

“[T]he ultimate goal of corpus analysis is a generalizable empirical description of language use in a target discourse domain. And the role of the corpus is to REPRESENT that targeted domain of language use” (p. 5)

- Domain considerations: “To what extent does the collection of texts in the corpus represent the range of texts and text types in the domain of interest?”
- Distribution considerations: “To what extent does the corpus accurately represent the quantitative distribution of the linguistic features of interest?” (a question of size) (p. 12)

# Representativeness – challenges

- To ensure representativeness, we need to think about both the range of texts produced in the educational contexts of interest and the range of pupils producing them
- Voluntary participation may bias our sample towards self-confident, motivated and good students (Gilquin 2015, p. 18)
- Constraints of access given by schools and teachers may bias our sample to text types that are easier to get hold of or easier for teachers to pass on
- Variation in texts produced across year groups and the classes of different teachers

# Representativeness – solutions and consequences

- Aimed to give as little extra work to students and teachers as possible
- Flexible in adapting to the wishes of schools and teachers for ways to get hold of texts (downloading from learning platforms, copying from notebooks, texts sent by teachers)

- More representative in terms of range than in terms of proportional representation
- More challenging to compare across individuals and across year groups than if we had given the students prompts
- Metadata crucial to be able to select sub-corpora for various purposes to control variables

# Metadata – background

- Data about the data
- Necessary in order to judge whether a corpus is appropriate for your purposes
- Learner corpora typically include information about both the learners (e.g. gender, age/school year, L1) and the texts (e.g. medium, genre, time constraints)
- A huge number of variables are involved in language acquisition and could potentially be included as metadata in learner corpora: “Learner corpora will … become more useful for advancing the SLA research agenda if they are accompanied with a wider variety of metadata” (Paquot 2022: 35)
- Granger and Paquot (2017) have called for more standardization of metadata in learner corpora and suggested that some categories be considered core/obligatory and others optional

# Metadata inclusion – challenges

- Practical considerations
  - How much information are participants willing/able to give us?
  - Are there other ways to get hold of metadata information?
  - Certain types of metadata require additional testing → how will this affect recruitment and retention of participants?
- Ethical considerations
  - With increased amounts of specific information, anonymity may be at risk
  - "Personal data is any information that can be linked to a person, either directly or indirectly, by putting different pieces of information together" (Sikt – Norwegian Agency for Shared Services in Education and Research, n.d.)

# Metadata inclusion – TRAWL

## Metadata about the students

- Educational program (for upper secondary school)
- Study level (for the second foreign language in upper secondary school)
- Gender
- Mother tongue
- Parents'/guardians' mother tongues
- What languages, in addition to the mother tongue, they can 1) read, 2) write, 3) speak, 4) understand
- Whether they have attended schools with other teaching languages than Norwegian
- If so, which languages and in which 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

# Metadata inclusion – TRAWL

## Metadata about the students

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)

# Metadata inclusion – solutions and consequences

- Recruitment and retention of students (and teachers) over time – chose to keep extra work at a minimum (no testing)
- To avoid possible identification of students:
  - we excluded collected information about birth country and whether + when they had lived in other countries than Norway
  - we collapsed very infrequent mother tongues into language groups

- No easily accessible proficiency information (can find indirect evidence via feedback and grades on texts)
- Not all research questions may be possible to answer
- Different sets of metadata may make it challenging to compare results from different corpora, e.g., when comparing L1 writing from one corpus with L2 writing from another

# Metadata categories – challenges (language background)

- Terminological problem: different interpretations of “first language/L1” and “second language/L2” and related terms (“native language”, “mother tongue”, “additional language”)
- Different terms used in questionnaires when collecting metadata for learner corpora
- Sometimes different terms are used in information about / studies based on corpora than those used in their metadata questionnaires
- Use of self-report data – the understanding of the researchers vs. the understanding of the participants
  - BAWE: different interpretations of “first language” by students with immigrant parents, but born and raised in the UK (Nesi, 2022)

# Solutions and consequences

- GiG: English as an Additional Language – "exposed to a language at home that is known or believed to be other than English" (Department for Education, 2020, p. 4)
- TRAWL: morsmål ('mother tongue') – chosen as the Norwegian term most familiar to students

- Both categories are dependent on the judgements of others than the corpus compliers
- We cannot be certain that the respondents interpret the terms in the same way
- Researchers have to keep in mind possible differences between understandings of the terms used to collect background information across corpora

- Alternative way forward: descriptions rather than terms when collecting information
  Which language(s) was/were the first you heard around you when you were little?
  Which language(s) do you feel that you know best? (MULTIWRITE project)

# Metadata categories – challenges (classification of texts)

- Terminological problem: inconsistencies in how labels are used (Durrant, 2022; Goulart, forthcoming)
- The fuzziness of text types (Biber & Egbert, 2023)
- School children are learning to write, and their writing develops over time
  - Children's writing may not be typical of "adult" text types
  - Tracing development over time may be a challenge
- Prompts may not be clear as to what text type/genre is expected, may demand a mix of types or be open for choice (Hasund 2022)
- Blending of text types may be a sign of mature writing (Bazerman et al. 2018; Berman & Verhoeven, 2002)

# Solutions and consequences

- GiG: literary and non-literary texts
- TRAWL: aims to categorize texts into six text types based on the task description – descriptive, expository, dialogic, argumentative, narrative and reflective (+ open)
- Inclusion of task descriptions / prompts

- Broad categories
- Lack of direct comparability with other corpora
- Need for detailed definitions and descriptions of the classification procedure
- Prompts can be used for further classification by other researchers

- Alternative way forward: characterization of texts based on multiple, graded, situational variables rather than classification based on purpose (Goulart et al. 2022)

# Text processing – background

- A distinctive feature of learner language (particularly that of young or inexperienced learners) is that it is erroneous (Gilquin & De Cock, 2011; Gilquin 2020)
- Erroneous language may cause problems for linguistic annotation:
  - Studies have shown that errors found in written learner corpora have affected the performance of POS taggers (Gilquin, 2020) and parsers (Huang et al. 2018)
  - Some studies propose to rewrite or normalize errors to represent a standard variant (e.g, Volodina et al. 2019)
  - However, any form of normalization or error annotation inevitably involves some kind of categorization or interpretation and thus "a necessary loss of information" (Lüdeling & Hirschmann 2015, p. 136)

# Text processing – challenges

- Young / less proficient learners produce more errors
- Automatic annotation tools were found to be inadequate
  - Complex structures diverged from the researchers' judgement in many cases
  - The grammatical categories made available by the tool were not always considered to be meaningful to teachers
- POS taggers for different languages do not always follow the same principles. This makes comparisons across languages (and corpora) challenging.

# Text processing – solutions and consequences

- Pure spelling mistakes were corrected in both corpora to improve POS tagging (a corrected version and the original version are both available in the TRAWL Corpus)
- Further normalization was not conducted
  - to avoid introducing interpretations and analysis of the data
- The POS taggers available for Norwegian and the other languages in TRAWL use somewhat different principles – no perfect solution was found

- Mistakes in the POS tagging (precision and recall)
- More effort required on behalf of corpus users to “tidy up” the data, but less in-built interpretations
- Potential problems in comparing searches across the languages of the TRAWL Corpus
- Necessity for documentation – also about the taggers used

# Summary

- The compilation of learner corpora in general and corpora of school writing by young learners in particular involve challenges that are not easily solved
- Different solutions have their own advantages and disadvantages
- Corpus researchers need to familiarize themselves with the choices made by the corpus compliers and what categories and labels signify before deciding whether the data can be used to answer their own research questions
- Corpus compilers must ensure that enough information is available for researchers to be able to do this

# References I


Ädel, A. (2020). Corpus compilation. In M. Paquot & S. Th. Gries (Eds.), A practical handbook of corpus linguistics (pp. 3–24). Springer.

Bazerman, C., Applebee, A., Berninger, V., Brandt, D., Graham, S., Jeffery, J. V., Matsuda, P. K., Murphy, S., Rowe, D., Schleppegrell, M. & Wilcox, K. (2018). *Lifespan Development of Writing Abilities*. Urbana, IL: NCTE Press

Berman, R. A., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy, 5*(1), 1–43. https://doi.org/10.1075/wll.5.1.02ber

Biber, D., & Egbert, J. (2023). What is register? Accounting for linguistic and situational variation within – and outside of – textual variables. *Register Studies*, *5*(1), 1–22. https://doi.org/10.1075/rs.00004.bib

Dirdal, H., Hasund, I. K., Drange, E.-M., Vold, E. T., & Berg, E. M. (2022). Design and construction of the Tracking Written Learner Language (TRAWL) corpus: A longitudinal and multilingual young learner corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), 115–135. https://doi.org/10.46364/njltl.v10i2.1005

Durrant, P., & Brenchley, M. (2018). *Growth in Grammar Corpus*. Available from <gigcorpus.com>. (registration required – contact Philip Durrant for access details: p.l.durrant@exeter.ac.uk).

Durrant, P. (2022). Studying children's writing development with a corpus. *Applied Corpus Linguistics*, *2*(3), Article 100026. https://doi.org/10.1016/j.acorp.2022.100026

Department for Education, (2020). https://www.gov.uk/government/publications/english-proficiency-pupils-with-english-as-additional-language

Egbert, J., Biber, D. & Gray, B. (2022). *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316584880

Egbert, J., Larsson, T. & Biber, D. (2020). *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108888790

Gilquin, G., & De Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, *16*(2), 141–172. https://doi.org/10.1075/ijcl.16.2.01gil

Gilquin, G. (2020). Learner corpora. In M. Paquot & S. Th. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 283 303). Springer.

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: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

# References II


Granger, S. & Paqout, M. (2017). Towards standardization of metadata for L2 corpora. https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools

Goulart, L., Biber, D. & Reppen, R. (2022). In this essay, I will …: Examining variation of communicative purpose student written genres. *Journal of English for Academic Purposes, 59*(3):101159: 1-43.

Goulart, L. (forthcoming). *Variation in university student writing*. John Benjamins.

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

Heuboeck, A., Holmes, J. & Nesi, H. (2010). The BAWE Corpus Manual, version 3. http://www.coventry.ac.uk/Global/08%20New%20Research%20Section/Current%20Projects/BAWEmanual%20v3.pdf

Huang, Y., Murakami, A., Alexopoulou, T. & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics 23*(1): 28-54.

Johannessen, J. B., Hagen, K., Lynum A., & Nøklestad, A. (2012). OBT+stat: A combined rulebased 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.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 135–158). Cambridge: Cambridge University Press.

Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics 20*(1), 30–53. https://doi.org/10.1075/ijcl.20.1.02mil

Nesi, H. (2022, September 22–24). Learner corpus research: Some problems, some questions, and some possible answers. [Plenary session]. 6th Learner Corpus Research Conference, Padua, Italy. http://www.maldura.unipd.it/lcr-2022/index.html

Paquot, M. (2022). Corpora and second language acquisition. In R.R. Jablonkai & E. Csomay (eds) *The Routledge Handbook of Corpora in English Language Teaching and Learning* (pp. 26-40). Routledge.

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.

Sikt – Norwegian Agency for Shared Services in Education and Research. (n.d.) https://sikt.no/

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C-J., Sundberg, G., & Wirén, M. (2019). The SweLL Language Learner Corpus. *Northern European Journal of Language Technology*, 6, 67–104. https://doi.org/10.3384/nejlt.2000-1533.196