Research Methods in Applied Linguistics 3 (2024) 100145

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Representativeness and metadata presentation in learner/child corpora: Lessons from the GiG and TRAWL corpora

Hildegunn Dirdal [a,*], Stine H. Johansen [a], Philip Durrant [b]

[a] Department of Literature, Area Studies and European Languages, University of Oslo, P.O. Box 1003, Blindern, N-0315 Oslo, Norway
[b] School of Education, Baring Court, University of Exeter St Luke's Campus, Heavitree Road, Exeter, EX1 2LU, UK

ARTICLE INFO

Keywords:
Representativeness
Metadata
School writing
Learner corpus
Child corpus

ABSTRACT

Representativeness is a key requirement in corpus linguistics, and the evaluation of the representativeness of an existing corpus depends on the provision of metadata. The present paper discusses challenges to both representativeness and metadata presentation based on our experiences in compiling corpora of school writing from young learners. Our discussion lends support to the calls for more transparent documentation and standardization, but also highlights some dangers that need to be kept in mind when attempting to standardize metadata.

Compiling a corpus is a time-consuming and costly process, especially for corpora with data from children and young learners – consent is needed at several levels (children, guardians, teachers, schools), texts are often handwritten and need to be transcribed, transcription and annotation may be difficult because of irregular language use, and shorter texts means that a greater number of texts are needed to create a substantial corpus. It is, therefore, natural that we should wish to reuse the data in such corpora to address questions beyond those for which they were originally designed. However, reusing corpora poses particular challenges: instead of tailoring data collection to align with their own research questions, users of existing corpora have to work with what has been provided by the compilers. Crucially, therefore, researchers need to ensure that the corpus is representative of both the situational and the linguistic variation in the language domain that they want to investigate (Egbert et al., 2022). To make informed decisions about which corpora to use and how to interpret data, users will need detailed information about the contents of the corpus and its creation (e.g. McEnery & Brookes, 2022; McEnery et al., 2006).

Such information about a corpus is often referred to as *metadata* and includes information about editing, analysis and administration in addition to classification of the corpus components (Burnard, 2005). Not only can metadata give a better understanding of the included material; it also provides information that can be used as variables in studies on this material (e.g. using metadata about language background to investigate how it influences writers'/speakers' text production). Within learner corpus research, calls for richer metadata are often motivated by a wish to study a wider range of the variables responsible for language learning and build bridges to the field of second language acquisition (e.g. Myles, 2021; Paquot, 2022). Recent initiatives involve attempts to reach agreement about standardization in terms of both what core metadata should always be included and how they are labelled and described (Granger & Paquot, 2017; Paquot et al., in press).

This article presents experiences from the creation of corpora containing school writing from young learners. We show how the creation of these corpora involve particular challenges related to representativeness and metadata collection, and discuss the

* Corresponding author at: Department of Literature, Area Studies and European Languages, University of Oslo, P.O. Box 1003, Blindern, N-0315 Oslo, Norway.
E-mail address: hildegunn.dirdal@ilos.uio.no (H. Dirdal).

https://doi.org/10.1016/j.rmal.2024.100145
Received 26 January 2024; Received in revised form 7 August 2024; Accepted 7 August 2024
Available online 23 August 2024

consequences of our solutions for subsequent users. Our discussion lends strong support to the calls for more transparent documentation, but also highlights some dangers involved in attempts to standardize metadata.

Background: representativeness and metadata

A key requirement in corpus linguistics is that corpora be representative of the language domain they are used to study: "the 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" (Egbert et al., 2022, p. 5; see also Biber, 1993; McEnery & Brookes, 2022).

Definitions of *corpus* often include the requirement that included texts should be authentic or naturally occurring (e.g. McEnery et al., 2006). This requirement distinguishes corpus data from various types of experimentally or clinically elicited data. Although 'texts' have been defined differently in different theories and research traditions (Titscher et al., 2000), most approaches involve the ideas both of having internal coherence and of being culturally recognized as units with situational characteristics and functional purposes (see e.g. Biber & Conrad, 2019; Egbert & Schnur, 2018).

To sample texts that adequately represent the target domain, Egbert et al. (2022) suggest three main steps: (1) describing the domain in terms of its boundaries and the text types it contains, (2) defining an operational domain by creating definitions for the set of texts that can be selected for the corpus, (3) using a sampling method to select actual texts from the operational domain. The representativeness of the corpus can be evaluated based on how well the sampling methods ensure inclusion of the range of texts in the operational domain and how well the operational domain matches the target discourse domain. Egbert et al. (2022) thus see representativeness as a continuous rather than dichotomous notion. Operational domains can be described in terms of internal categories, or strata, and Egbert et al. recommend a stratified sampling method. Such sampling may aim at equal-size strata or sizes that reflect their proportions in the discourse domain.

In addition to judging the extent to which the corpus represents the range of texts in the domain, researchers must also judge whether it is large enough to give an accurate representation of the features they want to study. Since the frequencies of linguistic features vary, the required size of a corpus will depend on the research question (Biber, 1993; Egbert et al., 2022).

Considerations about what metadata to include are connected to questions of representativeness in at least two ways. Firstly, information about corpus contents and how they were selected enables other researchers to judge its representativeness of the discourse domain they are interested in and how they might generalize findings. Secondly, metadata may allow researchers to create sub-corpora that better suit particular research questions (Gablasova et al., 2019), i.e. are representative of a smaller domain.

In addition, a wider set of metadata about individual texts, authors and contexts allows investigation of a wider range of factors that may influence text production or, in relation to child/learner corpora, factors that may influence language learning and writing development. As an attempt to offer guidelines about what metadata to include, improve study quality and ensure that learner corpus research follows the FAIR principles (data should be findable, accessible, interoperable and reusable), researchers have worked towards standardization of metadata for learner corpora (Granger & Paquot, 2017; Paquot et al., in press). This has resulted in a proposed schema (the Core Metadata Schema for Learner Corpora) with around 160 variables, some obligatory and others optional.[1] We will discuss challenges in relation to two central sets of metadata that are core/obligatory in the metadata schema, namely metadata about text types and first language.

The GiG and TRAWL corpora

A distinction is often made between child-language corpora and learner corpora. The former contain texts produced by children acquiring their first language (L1) / mother tongue and the latter texts produced by learners of additional languages (Gilquin, 2020). Not all corpora fit squarely into these categories, however, as is the case for the two corpora described here. The *Growth in Grammar (GiG) Corpus* (Durrant & Brenchley, 2018) contains texts written in English within an Anglophone school context. A majority of the students have English as their L1, but a sizeable proportion have English as an additional language. The *Tracking Written Learner Language (TRAWL) Corpus* (Dirdal et al., 2022) is compiled in a Norwegian school context. It contains texts written in Norwegian, English (the first foreign language taught in Norwegian schools) and French, German and Spanish (the most popular second foreign languages). Again, some of the children have other L1s than the majority language.

The GiG corpus was initiated in order to provide a collection of authentic school writing across multiple subject areas and age levels, and in these respects represents a type of corpus that is generally less common, and that is unique in the British context. It contains a total of 2898 texts written for English, science and humanities at five age levels in altogether 24 schools.[2] In the Norwegian context, such multi-disciplinary corpora of writing in the national language are represented by *The NORM Corpus* (Department of Linguistics & Scandinavian Studies, 2017), from primary school, and *SKRIV-korpuset* (Department of Linguistics & Scandinavian Studies, 2013), from upper secondary school. Like these corpora, TRAWL also contains authentic school writing from children of different ages. However, it includes texts written in other languages than the national language and only texts produced in language

[1] Version 2 of the schema can be found at https://dataverse.uclouvain.be/dataset.xhtml?persistentId=doi:10.14428/DVN/AAUEM2 (accessed 5 August 2024)

[2] Students and texts are divided over year groups as follows: year 2: 636 texts by 160 students, year 4: 49 texts by 10 students, year 6: 868 texts by 185 students, year 9: 804 texts by 457 students, year 11: 538 texts by 171 students. (The project did not originally target year 4, but a few texts were sent in from this year group and thus included in the corpus.)

classes. The TRAWL Corpus contains truly longitudinal data, with several texts collected from the same students over periods ranging from at least one and up to four school years. Around 8500 texts from more than 1200 students at 24 schools have been collected and are in various stages of processing. The first online version contains a total of 1663 texts by 216 students in years 8–12.[3]

Both GiG and TRAWL were intended for use by a broad research community and for multiple research questions, meaning that they are what Granger (2021, pp. 246–247) calls "all-purpose" rather than "purpose-built" corpora.

Challenges related to representativeness and metadata presentation

When compiling the corpora, we experienced challenges in relation to both representativeness and the collection, classification and reporting of metadata. Focusing on (pseudo-)longitudinal authentic texts from schoolchildren accentuated these challenges. As researchers follow the calls for more longitudinal data from young learners (e.g. Lu, 2022; Myles, 2021), we believe that it may be useful to share our experiences. However, we also believe that many of the problems we encountered are not unique to this particular type of corpora, and that our experience provides lessons for more general discussions of corpus design.

Domain representation

Creating corpora of authentic school writing led to challenges concerning both operationalization and sampling. This is due to the lack of a description of the full set of text types in the domain and the fact that we could only obtain texts from consenting participants. Both corpora targeted texts composed for ordinary classes by the general student population in the relevant years of primary and secondary school. Although the texts themselves are the primary data, an important part of acquiring a representative sample of child/learner texts is to ensure that they are written by a representative sample of students.

The collection of unpublished, private material and personal information is governed by strict rules of informed consent, e.g. the General Data Protection Regulation in Europe (Regulation, 2016/679). Both GiG and TRAWL required consent from three levels of participant: schools, teachers, and children (and, for children younger than 16, their parents/guardians), and at all three levels only a sub-set agreed to take part. It has been argued that students who volunteer to contribute to corpora may be among the more motivated and self-confident, and perhaps have a higher attainment level than average (Gilquin, 2015), which creates a challenge of representation. Similar factors may also affect teacher participation, especially if teacher feedback is collected along with the students' texts.

In corpora that aim to reflect text production by the general student population, it may be possible to evaluate the representativeness of the students who contribute by comparing their characteristics with data from national surveys. To what extent this can be done depends both on the existing national statistics and on the metadata that are collected about the students. National statistics would not include information about motivation or self-confidence, but may have information about attainment at national tests in at least some subjects. However, although the GiG project collected information about the attainment of the students, it was difficult to evaluate this against national statistics because each school had its own grading system. The TRAWL metadata do not include overall attainment for the students, but some texts are accompanied with feedback and grades. Average grades would have to be computed from these and compared with national statistics, something which has not been attempted yet. What we can say for both corpora is that a range of different attainment levels are represented. Even if these strata are not proportionally representative or of equal sizes, the corpora should be fairly representative in terms of range included. Other characteristics that may impact writing and attainment and that may be found in national statistics are language background and socio-economic status. The GiG project collected information about whether the children were classified as having English as an additional language and, as a proxy for socio-economic status, their eligibility for free school meals or a pupil premium. Comparisons with government statistics (Department for Education, 2015) showed that students with free school meals / pupil premiums were slightly over-represented (22% compared to 15% nationally), whereas students with English as an additional language were slightly under-represented (13.4% compared to 15%).

Although the representativeness of learner samples can be partially evaluated, it is more difficult to evaluate the representativeness of text types, as there are no existing lists of types or existing statistics to compare with. One solution might be not to sample, but to collect all the texts written by the consenting students. However, there are practical limitations, such as the type of access given to us by schools and teachers. In the TRAWL project, some schools gave access to learning platforms where electronic hand-ins could be downloaded; in other cases, teachers sent us texts or gave us permission to copy from notebooks. Not every piece of writing is available in learning platforms, and teachers may not save everything their students write. Corpus compilers have to respect the limitations to the time and effort the contributors are able to spend.

Although we were interested in a wide range of text types, some boundaries were set. These were not based primarily on text types, but on the type of language that was considered meaningful for the corpora. As mentioned above, both GiG and TRAWL are general-purpose written child/learner corpora, but we had in mind a focus on both vocabulary and grammar, and thus the ability to use words in context and combine them to make sentences. In the GiG project, pieces of connected prose were the units of writing aimed for. In the TRAWL project, the lower limit was set to at least one sentence. In both cases, we wanted texts originating from the students themselves, excluding translation exercises because of the way the source text influences and constrains the writing.

The decision to include texts consisting of only one sentence in TRAWL had to do with our aim of collecting as much as possible of

[3] Texts and students are divided over sub-corpora as follows: English: 1238 texts by 139 students (from years 8–12), Norwegian: 134 texts by 10 students (from years 8–10), French: 90 texts by 31 students (from years 11–12), German: 130 texts by 13 students (from years 11–12), Spanish: 72 texts by 23 students (from years 11–12). More texts will be added to all sub-corpora.

the writing of each learner. Young beginners may write very short texts even in response to tasks intended to elicit a piece of connected prose. Another common task type at school is to write answers to questions about, for instance, a book or a film, where answers may also consist of single sentences, even at more advanced levels. Placing the lower limit at a sentence allows for collection of such texts and, at the same time, ensures meaningful analysis of syntactic competence. We did not include inflection exercises or exercises where single words or phrases had to be provided. The principles used to select texts for TRAWL mean that studies based on this material can be generalized only to texts where students connect words into full sentences and are not translating from Norwegian. Detailed information about the principles according to which texts are included or excluded are thus important for subsequent users.

Even though the TRAWL project aimed at collecting all text of one sentence or more produced by the consenting students, the actual sample was affected by the limitations to access as described above. We expect the sample to be skewed towards longer texts submitted for assessment, as these are more commonly submitted through the learning platforms we were given access to and are more likely to have been kept by teachers who preferred to send us texts themselves. For GiG, the decision of what texts to submit was left to the teachers themselves. Teachers were told only that we wished to collect samples of their students' written work, with no specification of the type or format of that work. Most writing that we received was connected prose. However, a few submissions were excluded because they did not contain sentences. This was seen for example, in some labelled diagrams.

As we have seen, choosing to collect authentic school writing makes it difficult to judge how representative the sample is of the discourse domain. It also leads to variation in the texts contributed by individuals, as teachers have a lot of freedom in selecting tasks within the curriculum, students are often given choices, and not all students submit all the tasks that are set. Collecting texts based on prompts designed by the corpus compilers allows for more control of variables across individuals and levels. However, corpora of school writing allow us to answer questions about the kinds of texts through which students' development actually happens and the kinds of writing taking place in schools. In studies of language learning and writing development, it is important that findings can feed into the school context. Basing research on authentic school writing may be one way of bridging the gap between research and pedagogy that has been pointed out by e.g. Ellis (2010) and Spada (2015).

Metadata categories and standardization

We have argued that not having an inventory of types of school writing makes it difficult to evaluate the representativeness of the samples. Once texts are collected, though, we might expect to be able to categorize them and include this information as metadata. Such information is listed as obligatory categories in the draft for a standard metadata schema referred to above and would make it possible for corpus users to make principled selections of texts if they want to investigate or compare particular text types. However, there are challenges involved in such categorization.

One issue is terminological. Reviews of the literature (Durrant, 2022; Goulart, in press) have highlighted widespread inconsistencies in how labels for different text types are used. Diverse terms have been applied to similar texts and, perhaps more worryingly, the same label to widely different types. This means that traditional labels such as *exposition* and *essay* cannot be taken for granted as different researchers appear to mean different things by them, and often fail to provide clear definitions. Even the labels *register* and *genre* have been used in a variety of ways, sometimes for essentially the same concept and in other cases about contrasting concepts (Biber & Conrad, 2019; Biber & Egbert, 2023). This is highly problematic for learner corpus research. Text type has an effect on language use that often far outweighs the effects of other variables of interest, such as year of study (Durrant et al., 2021). However, uncertainty about how types have been defined and operationalized in much of the research literature makes it difficult to trace its influence and to determine the extent to which different studies are, in fact, studying comparable texts. Looking forward, the lack of shared understandings about what genre labels mean also creates problems for attempts to create a standardized metadata schema.

A more fundamental issue concerns the inherent fuzziness of text types. Recent studies have demonstrated the diversity that exists within culturally recognised text types (Biber & Egbert, 2023). Category membership is, thus, a graded phenomenon, with some texts being more prototypical of a type than others. Educational texts are no exception (Goulart et al., 2022). A central expectation of school is that the texts children write will develop as they mature (Coffin, 2006); a research report in primary school does not have all the features of a research report in the final year of secondary. Moreover, the skilled blending of text types (e.g. the incorporation of narrative elements into an argument) can be a sign of a more mature writing style (e.g. Bazerman et al., 2018; Berman & Verhoeven, 2002).

All this makes classification of school writing difficult. The solution adopted for GiG was to use two broad categories that could be applied across all year groups: literary and non-literary texts. Texts whose purpose was to be appreciated on their own terms were labelled literary, whereas texts that could be evaluated as successful or unsuccessful based on their relationship to the real world (whether propositional or directive) were labelled non-literary (Durrant et al., 2021). For TRAWL, work is also under way to classify texts. The classification of a subset of texts into fine-grained genres (or what the metadata schema calls *registers*) proved challenging in many cases (Hasund, 2022). It has therefore been decided to proceed only with classification into the broader categories descriptive, expository, dialogic, argumentative, narrative/poetic and reflective – categories representing what the proposed metadata schema calls *communicative purpose*. As the schoolchildren are still learning to create texts with such purposes, the classification is based on the prompts rather than the texts themselves. However, prompts may be fuzzy, mixed or unspecified, requiring decisions on the most dominant purpose and resulting in the addition of an 'open' category.

Acknowledging the limitations of our current solutions, we believe that the fuzziness of text types and the developmental nature of children's writing calls for an approach that is more dynamic and flexible. A promising way forward is suggested in the recent work of Goulart et al. (2022), building on Biber et al.'s (2020) research with Internet registers. Their approach replaces *classification* based on a single defining purpose with *characterization* based on multiple, graded, situational variables. This is beneficial because, by positioning

each text in a multidimensional space, it enables multiple types of situational variables to be accounted for (purpose, audience, topic, format, etc.); it captures the variety within, and the overlaps between, classical text types; it enables the emergence of novel groupings of texts with similar features; and it does not force texts to belong to a single type, thus enabling the types of hybridity that are common in school writing. By summarizing text type variation in a small number of dimensions, rather than a potentially unlimited number of discreet categories, it may also help teachers and researchers better understand underlying principles of text variation.

Inconsistency in the use of labels is not exclusive to text types, but also an issue for core categories of learner/child metadata. As an example, we will discuss *first language (L1)* and related labels. We have noticed that different terms are chosen in questionnaires used to collect metadata. For example, the submission form used for the *British Academic Written English corpus* (BAWE) asks for *first language* (Heuboeck et al., 2010), whereas the questionnaire[4] used for the *International Corpus of Learner English* (ICLE) asks for *native language*. The TRAWL questionnaire (Dirdal et al., 2022) uses the term *morsmål* ('mother tongue') based on considerations of what term is most understandable to Norwegian children. However, in information about and studies based on these corpora, other terms may be used than those in the questionnaires, often *first language*.

Some differences in the use of terms have to do with conventions in different languages or familiarity for different demographic groups, but there are also theoretical ambiguities involved, as well as problems to do with potentially different subjective interpretations by questionnaire respondents. The term *first language* is used about the language learnt first in life, but also about the language that someone is most proficient in Loewen and Reinders (2011). People will often be most proficient in the language they learnt first. However, this is not necessarily the case, and we do not know exactly how learners interpret the terms in the questionnaires. Nesi (2022) recounts how some BAWE contributors claimed not to have English as their first language even if they had grown up in the UK and received all their education there, whereas others with the same background did list English as their first language. The first language variable in BAWE, then, cannot be used to compare L1 and L2 English without careful consideration of the metadata collection process and the context in which it was conducted. Similarly, the GiG metadata includes information on whether students are counted by their school as having 'English as an Additional Language' (EAL). In the UK, students are classified as having EAL if they are "exposed to a language at home that is known or believed to be other than English" (Department for Education, 2020, p. 4). This covers a wide range of students, including both new arrivals who may have little knowledge of English and children born and raised in the UK but one of whose parents sometimes speaks to them in another language. The "believed to be" element of the definition also highlights the fact that classification of a student as EAL often depends on teacher perceptions.

If a corpus contains several types of metadata about language background, this can help in the evaluation of the participants' learner status. BAWE contains metadata about how much of the participants' secondary education was in the UK, and ICLE contains metadata about parents' mother tongues, language(s) spoken at home and medium of instruction. TRAWL also contains metadata on parents' mother tongues and whether participants have lived in countries other than Norway or gone to schools with a medium of instruction other than Norwegian. Together, these give a better picture of the participants' language expertise. Another solution to the problem of the ambiguity of the terms *first language, native language* and *mother tongue* is to ask participants more specifically about what language(s) they learnt first in life and what language(s) they are most proficient in. This is the solution chosen by the MULTIWRITE project,[5] where the following questions are asked (our translation from Norwegian):

> Which language(s) was/were the first you heard around you when you were little?
>
> (For example: If your parents spoke Norwegian with you when you were little, you write "Norwegian" here.)
>
> Which language(s) do you feel that you know best?
>
> (For most people this is the same language as in the previous question. An exception could for example be if you heard Japanese at home when you were little, but have used Norwegian in kindergarten, at school and in your daily life and feel that you know that language better now.)

As many as 38% of the respondents gave different answers to the two questions (although in a third of these cases there was overlap because more than one language was entered for either of the two questions or both). More specific questions such as these give metadata that can be interpreted in a more reliable way. In the fields of bilingualism and second language acquisition, there has been work on developing questionnaires that gauge the use, mastery and starting age for both/all an individuals' languages (e.g. Castilla-Earls et al., 2022; Kaushanskaya et al., 2020), and learner corpus research may draw more inspiration from this work in the future.

Conclusion

Our experiences highlight the need for detailed metadata, including thorough descriptions of definitions, delimitations and collection processes, to enable corpus users to make more informed decisions about the degree of representativeness of a corpus and judge what kinds of material it actually contains. When compiling corpora of school writing, we had to take a somewhat 'pragmatic'

[4] Accessible from https://cdn.uclouvain.be/public/Exports%20reddot/cecl/documents/LEARNER_PROFILE.txt

[5] "MULTIWRITE – Interactions Between First, Second and Third Languages" is a project comparing syntactic complexity in the writing of the same individual students in Norwegian, English and French/German/Spanish, as well as feedback practices in the different language subjects and collaboration between language teachers. Data from the project will be added to the TRAWL Corpus. The project is funded by the Norwegian Research Council (grant number 324962).

approach to operationalization and sampling, as the texts were not publicly available and had to be collected based on voluntary participation by students and teachers, who gave us access to texts in various ways. Despite these limitations, we believe it is important to compile corpora of authentic school writing because they allow us to study writing development the way it actually happens in schools and offer findings that can feed into the school context. Some evaluation of the student samples can be carried out based on comparisons with national statistics and by looking at the range of attainment levels included. The possibility of making such evaluations will depend on the types of metadata collected about students and texts, further supporting the inclusion of rich metadata.

Our review of how labels are used to classify texts and language background supports the need for standardization in this area. At the same time, it raises questions about the possibility of arriving at agreed definitions. Moreover, we need to ensure that the use of standard labels does not fool us into believing that classifications are more similar than they are. We thus agree with the creators of the Core Metadata Schema for Learner Corpora (Paquot et al., in press) that we need explicit information about the corpus compilers' interpretations of the terms and how they collected relevant information. For both genre and first language background, we may want to adopt a more complex, nuanced approach than is currently the norm. Genres may be summarized according to situational dimensions rather than set categories, and learners may be asked more specific questions letting us distinguish between competing interpretations of labels such as *first language* and *mother tongue*.

Author note

This work was supported by the Research Council of Norway [grant number 324962].

We would like to thank the special issue guest editor, Lee McCallum, and two anonymous reviewers for insightful comments on earlier versions of this paper.

CRediT authorship contribution statement

Hildegunn Dirdal: Writing – review & editing, Writing – original draft, Conceptualization. **Stine H. Johansen:** Writing – review & editing, Writing – original draft, Conceptualization. **Philip Durrant:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

None.

References

Bazerman, C., Applebee, A. N., Berninger, V. W., Brandt, D., Graham, S., & Jeffery, J. V. (2018). *The lifespan development of writing*. National Council of Teachers of English. https://wac.colostate.edu/books/ncte/lifespan-writing/.

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

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press. https://doi.org/10.1017/9781108686136

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

Biber, D., Egbert, J., & Keller, D. (2020). Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory, 16*(3), 581–616. https://doi.org/10.1515/cllt-2018-0086

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxbow Books for the Arts and Humanities Data Service. https://users.ox.ac.uk/~martinw/dlc/chapter3.htm.

Castilla-Earls, A., Ronderos, J., & Fitton, L. (2022). Can bilingual children self-report their bilingual experience and proficiency? The Houston questionnaire. *Journal of Speech, Language, and Hearing Research, 65*(10), 3835–3853. https://doi.org/10.1044/2022_JSLHR-21-00675

Coffin, C. (2006). *Historical discourse: The language of time, cause and evaluation*. Continuum.

Department for Education. (2015). *Schools, pupils and their characteristics: January 2015*. https://assets.publishing.service.gov.uk/media/5a7f759740f0b6230268f9f0/SFR16_2015_Main_Text.pdf.

Department for Education. (2020). *English proficiency of pupils with English as an additional language: Ad-hoc notice. February 2020*. https://assets.publishing.service.gov.uk/media/5e55205d86650c10e8754e54/English_proficiency_of_EAL_pupils.pdf.

Department of Linguistics and Scandinavian Studies. (2013). *SKRIV-korpuset*. University of Oslo. https://www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/skriv/.

Department of Linguistics and Scandinavian Studies. (2017). *The norm corpus*. University of Oslo. https://www.hf.uio.no/iln/english/about/organization/text-laboratory/projects/norm/index.html.

Dirdal, H., Hasund, I. K., Drange, E.-M. D., 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. (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. Article.

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

Durrant, P., Brenchley, M., & McCullum, L. (2021). *Understanding development and proficiency in writing: Quantitative corpus linguistic approaches*. Cambridge University Press. https://doi.org/10.1017/9781108770101

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press. https://doi.org/10.1017/9781316584880

Egbert, J., & Schnur, E. (2018). The role of the text in corpus and discourse analysis: Missing the trees for the forest. In C. Taylor, & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 159–173). Routledge.

Ellis, R. (2010). Second language acquisition, teacher education and language pedagogy. *Language Teaching, 43*(2), 182–201. http://doi.org/10.1017/S0261444809990139.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research, 5*(2), 126–158. https://doi.org/10.1075/ijlcr.19001.gab

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.

Gilquin, G. (2020). Learner corpora. In M. Paquot, & S.& Th. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 283–303). Springer. https://doi.org/10.1007/978-3-030-46216-1_13.

Goulart, L. (in press). *Variation in university student writing.* John Benjamins.2024.

Goulart, L., Biber, D., & Reppen, R. (2022). *In this essay, I will …*: Examining variation of communicative purpose in student written genres. *Journal of English for Academic Purposes, 59*, Article 101159. https://doi.org/10.1016/j.jeap.2022.101159. Article.

Granger, S. (2021). Commentary: Have learner corpus research and second language acquisition finally met? In B. Le & Bruyn, & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 243–257). Cambridge University Press. https://doi.org/10.1017/9781108674577.012.

Granger, S., & Paquot, M. (2017). *Towards standardization of metadata for L2 corpora* [Workshop presentation]. In *Workshop on interoperability of L2 resources and tool.* Sweden: University of Gothenburg.

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., Homes, J., & Nesi, H. (2010). *The BAWE Corpus manual. Version III.* https://www.coventry.ac.uk/globalassets/media/global/08-new-research-section/current-projects/bawemanual-v3.pdf.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism: Language and Cognition, 23*(5), 945–950. https://doi.org/10.1017/S1366728919000038

Loewen, S., & Reinders, H. (2011). *Key concepts in second language acquisition.* Palgrave Macmillan.

Lu, X. (2022). What can corpus software reveal about language development? In A. O'Keeffe, & M.& J. McCarthy (Eds.), *The routledge handbook of corpus linguistics* (2nd ed., pp. 155–167). Routledge. https://doi.org/10.4324/9780367076399-12.

McEnery, T., & Brookes, G. (2022). Building a written corpus: What are the basics? In A. O'Keeffe, & M.& J. McCarthy (Eds.), *The routledge handbook of corpus linguistics* (2nd ed., pp. 35–47). Routledge. https://doi.org/10.4324/9780367076399-4.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Routledge.

Myles, F. (2021). Commentary: An SLA perspective on learner corpus research. In B. Le & Bruyn, & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 258–273). Cambridge University Press. https://doi.org/10.1017/9781108674577.013.

Nesi, H. (2022). *Learner corpus research: Some problems, some questions, and some possible answers* [Plenary session]. In *6th Learner Corpus Research Conference*. 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 and English language teaching and learning* (pp. 26–40). Routledge. https://doi.org/10.4324/9781003002901-4.

Paquot, M., König, A., Stemle, E.W., .& Frey, J.-C. (in press). The Core Metadata Schema for Learner Corpora: Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research,* 10(2).

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation. (2016). OJ L 119/1 https://gdpr-info.eu/.

Spada, N. (2015). SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching, 48*(1), 69–81. https://doi.org/10.1017/S026144481200050X

Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2000). *Methods of text and discourse analysis: In search of meaning.* Sage.