Corpus-based approaches to spoken L2 production

Evidence from the Trinity Lancaster Corpus

Vaclav Brezina, Dana Gablasova and Tony McEnery
Lancaster University, UK

From the perspective of the compilers, a corpus is a journey. This particular journey – the process of the design and compilation of the Trinity Lancaster Corpus (TLC), the largest spoken learner corpus of (interactive) English to date – took over five years. It involved more than 3,500 hours of transcription time¹ with many more hours spent on quality checking and post-processing of the data. This simple statistic shows why learner corpora of spoken language are still relatively rare, despite the fact that they provide a unique insight into spontaneous language production (McEnery, Brezina, Gablasova & Banerjee 2019). While the advances in computational technology allow better data processing and more efficient analysis, the starting point of a spoken (learner) corpus is still the recording of speech and its manual transcription. This method is considerably more reliable in capturing the details of spoken language than any existing voice recognition system. This is true for spoken L1 (McEnery 2018) as well as spoken L2 data (Gilquin 2015). The difference between the performance of an experienced transcriber and a state-of-the-art automated system is immediately obvious from the comparison shown in Table 1. For meaningful linguistic analysis, only the sample transcript shown on the left (from the TLC) is suitable as it represents an accurate account of the spoken production.

Building a spoken learner corpus is thus a resource-intensive project. The compilation of the TLC was made possible by research collaboration between Lancaster University and Trinity College London, a major international testing board. The project was supported by the Economic and Social Research Council (ESRC) and Trinity College London.²

---

1. We would like to express our gratitude to Ruth Avon (main transcriber) and Alana Jackson, who worked tirelessly to produce high-quality transcripts for the TLC.

2. Grants ES/K002155/1, ES/R008906/1, EP/P001559/1 and ES/S013679/1.

Table 1. Comparison of manual and automatic transcription of spoken learner data

Transcript produced by an experienced transcriber (Trinity Lancaster Corpus)	Automatic transcription (Nuance Dragon Professional v. 15.3 with machine learning technology)
<DISC>	
E: erm we begin with your topic <name> er what topic are we going to talk about?	we begin with your topic other what topic are we going to talk about
S: animals	animals
E: animals all animals	all animals
S: yes	
E: what's your favourite? <.> your tell me first your favourite animals	what's your favourite animals
S: oh	
E: mm	
S: my <unclear=favourite> animals of course is dog	
E: mm	
S: because when I was seven years old	when you resell more dark and laugh
E: mm	
S: my mum and walking with me and and we saw a big patch sheep and and I meet a small dog	my mum and
E: uhu	
S: he er he looks very cute	
E: mm	
S: and like that	
E: oh right uhu yes	
[...]	

This special issue of the *International Journal of Learner Corpus Research* is dedicated to novel explorations of spoken learner language enabled by the TLC. As mentioned above, the TLC is a unique resource in many respects. It is the largest corpus of its kind with detailed annotation of key language learning variables, enabling much called-for studies on, for instance, the effect of task, age of exposure, type of exposure, proficiency, etc. It also brings data from a variety of L1 and cultural backgrounds including backgrounds not previously covered in detail such as Hindi and Mexican Spanish. The ambition of this issue is thus to move the field forward by providing an insight into L2 production that was not previously possible with the level of detail TLC-based research can provide. The TLC offers a large amount of linguistic data combined with rich metadata about the speak-

ers and their performance across different tasks. As Myles (2015: 309) points out, a crucial concern in the field is the “need for good learner data”. The TLC comprises both the quantity, which allows us to draw statistical inferences, as well as quality of the data, which allows us to properly interpret the complexity of the patterns observed. The special issue can thus be seen as a vanguard of the research, which, we hope, will appear with the public release of the corpus.

The research presented in this issue was carried out by researchers who were granted, on a competitive basis, early access to the Trinity Lancaster Corpus Sample, a subset of the TLC. This subset consists of approximately two million words, i.e. just about half of the whole corpus. The distribution of L2 speakers in the TLC Sample across three proficiency levels of the Common European Framework of Reference (CEFR; Council of Europe 2011), L1s and cultural backgrounds is displayed in Table 2. All other characteristics (nature of tasks included, transcription conventions, etc.) are the same as described in Gablasova, Brezina and McEnery (this issue).

Table 2. Number of L2 speakers across proficiency levels, L1s and cultural backgrounds in the Trinity Lancaster Corpus Sample

L1 / cultural background	Proficiency level acc. to CEFR		
	B1	B2	C1/C2
China	80	70	30
India	100	100	30
Italy	100	100	60
Mexico	100	70	60
Russia	40	28	–
Spain	100	100	60

The special issue includes six contributions. The first contribution (Gablasova, Brezina & McEnery) introduces the TLC and contextualises this new resource within learner corpus research on spoken L2. The article addresses key areas in the design of learner corpora of spoken language such as representativeness, amount of evidence available in the corpus (in terms of the number of words, individual contributions and the length of each contribution) and the reliability of the transcription of speech. A detailed description of the corpus is provided in order to maximise its potential of use for the research community. We also highlight some applications of the corpus and outline future directions of our research programme, which involves, among other things, the creation of a comparable L1 counterpart to the corpus. Notably, the Lancaster Spoken Language Transcription Guidelines are provided in full in the Appendix accompanying the article, allowing other

researchers interested in the creation of speech corpora to use a standardised set of conventions for converting speech into a written form; these have been extensively tested in the compilation process of a number of spoken corpora built at Lancaster University, including the TLC.

The second contribution by Götz focuses on filled pauses across L2 proficiency levels, an important issue related to fluency. In addition to L2 proficiency, the study investigates a variety of contextual variables as possible predictors of the frequency of use of filled pauses; these variables include the L2 speaker's age, the age of exposure, the country of origin as well as the effect of the interlocutor (examiner and L1 speaker of English). The article offers a valuable discussion of the inclusion of specific fluency phenomena into the CEFR descriptors.

The third (Gilquin) and the fourth (Römer & Garner) contributions look at verb constructions, a crucial element in the lexicogrammar of L2 production. Gilquin analyses light verb constructions such as *take a walk* or *make a choice*, which are often claimed to present a problem for L2 speakers. The study investigates both the effect of L2 proficiency as well as the effect of acquisitional contexts: EFL (English as a Foreign Language) vs. ESL (English as a Second Language). A comparison is made with native language use in the TLC, included in the form of the examiners' speech. Römer and Garner focus on verb-argument constructions (VACs) such as V *about* N (*care about fashion*) or V *for* N (*ask for a taxi driver*). The study examines the effect of L2 proficiency and L1 background. L2 usage of VACs in the TLC is compared with L1 usage of VACs in the British National Corpus (BNC; BNC Consortium 2007) as a proxy for the target language.

The fifth (Castello & Gesuato) and the sixth (Pérez-Paredes & Díez-Bedmar) contributions address topics in the field of L2 pragmatics with a focus on back-channelling and active listenership, and adverbs of certainty, respectively. Castello and Gesuato's detailed analysis of expressions such as *right*, *of course* and *I know* combines automatic data extraction with meticulous manual checking and categorising of concordance lines to investigate the frequencies and functions of lexical back-channels in L2 speech across different L1 backgrounds. The article also offers a discussion of the implications of the study for the assessment of oral proficiency. Pérez-Paredes & Díez-Bedmar's study explores the usage of *actually*, *really* and *obviously* across proficiency levels by EFL learners whose L1 is Spanish (speakers from Spain and Mexico). It also combines quantitative with qualitative analysis looking into the functions of the selected adverbs of certainty.

Overall, these studies show a range of different possible explorations of the TLC, engaging with its rich metadata. We hope that they will be followed by many more in the future. These studies contribute to our better understanding of L2 speech across different contexts, types of L2 users and proficiency levels (B1-C2).

The studies also demonstrate different types of statistical analysis available for spoken corpus data ranging from complex statistical models to simple pairwise comparisons. Importantly, individual differences between speakers are taken into consideration in all the statistical analyses (cf. Brezina & Meyerhoff 2014). Some of the analyses also choose to increase the precision of the results by considering the effect of multiple variables from the rich metadata of the TLC. Nevertheless, some methodological divergence needs also to be noted. While inferential statistics in contributions two to five are based on relative frequencies, which help to counterbalance the effect of different amount of spoken production in different L2 speakers,³ the final contribution (Pérez-Paredes & Díez-Bedmar) considers absolute frequencies of the target linguistic variables. This methodological decision somewhat changes the context of interpretation of the results, shifting the focus from the target variables to a more holistic understanding of L2 performance per task; this shift of focus needs to be carefully evaluated in any future meta-analyses (Brezina 2018: 267ff).

The TLC was designed as a resource which meets the needs of both learner corpus research (LCR) and second language acquisition (SLA). The amount of evidence about L2 speech that it provides, combined with rich metadata, allow researchers from both disciplines to answer a variety of research questions about L2 use. As discussed in detail in McEnery, Brezina, Gablasova and Banerjee (2019), LCR and SLA have a real opportunity to collaborate provided that both fields reflect on their own assets as well as limitations (see also Myles 2015). Construction of new learner corpora with SLA research questions in mind can thus lead to productive synergy with collective ownership of goals between the two disciplines.

The TLC will be released for non-commercial research purposes in 2019. Our aim is to allow different academic/practitioner groups of users to access the data via various platforms to support different research requirements and needs. These platforms will include:

1. Sketch Engine <https://www.sketchengine.eu/> (with an application for access for a specific project; Kilgarrieff et al. 2014). This form of access is suitable for researchers and advanced users because the system requires specific training and knowledge about how to normalize data and work with subcorpora.
2. Trinity Lancaster Corpus Hub <http://corpora.lancs.ac.uk/trinity/>. This is a brand-new interface developed at the ESRC Centre for Corpus Approaches to Social Science, Lancaster University. It offers user-friendly access to the

---

3. In the TLC the variation in the overall amount of L2 production per task (in number of tokens) is considerable. For instance, the same task (conversation) at the same proficiency level (B1) elicited from 96 to 831 tokens (running words).

corpus without requiring any technical knowledge. Analyses are performed automatically and different visualization options are available. The system is suitable for a wide range of users including researchers and practitioners.

3. Lancaster solutions for corpus linguists such as #LancsBox (Brezina et al. 2015) <http://corpora.lancs.ac.uk/lancsbox/>. These tools have unique features that add extra value to the data by allowing analyses (e.g. collocation networks) that are not available elsewhere.

References

- BNC Consortium 2007. *British National Corpus*, XML edn. Available at <<http://ota.ox.ac.uk/desc/2554>>
- Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brezina, V., McEnery, T. & Wattam, S. 2015. "Collocations in context: A new perspective on collocation networks". *International Journal of Corpus Linguistics* 20(2), 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Brezina, V. & Meyerhoff, M. 2014. "Significant or random. A critical review of sociolinguistic generalisations based on large corpora". *International Journal of Corpus Linguistics* 19(1), 1–28. <https://doi.org/10.1075/ijcl.19.1.01bre>
- Council of Europe. 2011. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe. <https://www.coe.int/t/dg4/linguistic/source/framework_en.pdf>
- Gilquin, G. 2015. "From design to collection of learner corpora". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9–34. <https://doi.org/10.1017/CBO9781139649414.002>
- Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. 2014. "The Sketch Engine: ten years on". *Lexicography* 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- McEnery, T. 2018. "The Spoken BNC 2014: The corpus linguistic perspective". In V. Brezina, R. Love & K. Aijmer (Eds.), *Corpus Approaches to Contemporary British Speech*. New York: Routledge, 10–15.
- McEnery, A., Brezina, V., Gablasova, D. & Banerjee, J. V. forthcoming. "Corpus Linguistics, learner corpora and SLA: Employing technology to analyze language use". *Annual Review of Applied Linguistics*.
- Myles, F. 2015. "Second language acquisition theory and learner corpus research". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 309–332. <https://doi.org/10.1017/CBO9781139649414.014>

Address for correspondence

Vaclav Brezina
Department of Linguistics and English Language
Lancaster University
County South
Lancaster LA1 4YL
United Kingdom
v.brezina@lancaster.ac.uk

Co-author information

Dana Gablasova
Department of Linguistics and English
Language
Lancaster University
d.gablasova@lancaster.ac.uk

Tony McEnery
Department of Linguistics and English
Language
Lancaster University
a.mcenery@lancaster.ac.uk