Designing and Building the Brazilian Spoken English Learner (BraSEL) Corpus

Mateus Miranda
mm2542@cam.ac.uk

University of Cambridge, Faculty of Education
University of Limerick, Ireland
Federal University of Minas Gerais, Brazil

Introduction

- Through learner corpora analyses, studies have been done on real language in use that contribute to the understanding of learners' difficulties in second language acquisition;
- Studies on learner corpora are mainly focused on academic contexts (Alterberg and Tapper 1998; Díez-Bedmar and Papp 2008; Gilquin and Paquot 2008) with a contrastive paradigm;
- These studies use the learners' year of study as a proxy for development over time. However, year of study can be an arbitrary marker (Tono and Díez-Bedmar 2014). Hence, any corpus based on Common European Framework of Reference (CEFR) levels must be robust in categorizing learners;
- Using examination-based data, advances have been made on written corpora calibrated to CEFR (Capel 2010, Buttery and Caines 2012, Harrison and Barker 2015, O'Keeffe and Mark 2017);
- Meanwhile, research on CEFR-calibrated spoken corpora is in its nascence (Trinity Lancaster Corpus 2010; Jones *et al.* 2018).

Goals and Methods

Part I of this PhD project, located within the field of Corpus Linguistics, involves the compilation of an innovative learner corpus benchmarked to the Common European Framework of Reference (CEFR) and its analysis, offering a sampling frame for further data collection.

Part II of this project is focused on the analysis of hedging devices and their development across the levels in two different ways:

- Top down route by gathering all the forms identified already in the literature as search candidates;
- Bottom up route by looking at frequent items in wordlists, including cluster lists.

- To observe the development of these features, the sub-corpus will be loaded into Sketch Engine that generates word sketches from corpus data (Kilgariff *et al.* 2004).

Building the Corpus

This project involves the participation of 9 universities in Brazil. The criteria to participate are that participants must be 1) an English learner and 2) part of the English without Borders (EwB) Programme in Brazil. Students must also hold an international proficiency certificate.

Corpus information

Size	116 recordings from 20 to 30 minutes each
Transcription	Orthographic
Number of students	116 students
Level of students	A2 to C2 students
Gender	Male and female (balanced)
Age group	18–50 years old

Table 1. Corpus design matrix.

Task Design

The recordings cover different tasks:

- (1) **Free discussion:** An informal conversation about students' lives, interests, hobbies, etc.
- (2) **Thought-provoking picture description:** This part is based on a set of thought-provoking pictures. The interviewer shows the participants some pictures and asks them to choose two to set the scene and describe it, as follows:

Pictures <12> and <5>

In order to identify the pictures that each participant chose to talk about, the pictures received numbers that are included in the transcription.

Topics for picture description and discussion include veganism, poverty, gender equality, immigration, and technology, among others.

Each interview accompanies:

- An audio file in .mp3 or .WAV format;
- A plain text (.txt) with the transcription;
- A plain text (.txt) with the metadata retrieved from a learner profile form:

- The international proficiency exam score;
- Age;
- Gender;
- Exchange or travel experience abroad;
- Number of years studying English;
- Other languages spoken.

Transcription Challenges in a Learner Corpus

- The implementation of a criterion is not an easy task;

Ten years later after Baker (2010) notes on spoken automatic transcription, despite the development of technology, software programs still struggle to offer accurate transcription, especially when two or more participants are involved.

Four software programs were tested - Temi, Trint, Audext and Otter - but the results were not satisfactory. Therefore, manual transcription, despite being laborious, is still the best solution, and training is necessary to reduce disagreements in applying the criteria.

To reduce disagreements related to the corpus transcription guidelines, workshops and training sessions describing the compilation process and guidelines were provided to the assistant research staff.

- The workshops' main goals were to: (a) give an overview on learner corpus research; (b) introduce the BraSEL Corpus architecture and transcription criteria; (c) present the software for transcription (Notepad++, Audacity, Praat and WinPitch); and (d) provide a long-term experience in the field of Corpus Linguistics.

- Many features are represented in the transcriptions (Table 2). The most frequent problems in the transposition of learner speech to written form are those related to truncated words, hesitations, and filled pauses. These are due to the difficulties associated with the recognition of the speaker's uttered material.

Interactional: Speaker turns, overlaps
Verbal: Acronyms; capitalization; code switching; contractions; fillers; lengthening; lexicalized reduced forms; numbers and dates; orthography; punctuation; repetition; unintelligible; word fragments/truncation
Vocal (non-verbal): Exhalation/Inhalation; laughing; pauses
Non-vocal events: Contextual events

Table 2. Overview of the features represented in the transcription.

Validation Process

- Reliability (inter-rater reliability) and the agreement (inter-rater agreement) between transcribers were measured.

After the inter-rater agreement analysis, a feedback section was given to focus on transcribers' difficulties and reduce disagreement in future transcriptions.

- Despite the still nascent amount of research focusing specifically on learner oral corpora across the CEFR, the conclusion of this project will enable further linguistic analysis to better understand learners' spoken interlanguage features.

References

- Alterberg, B. and M. Tapper. (1998). The use of adverbial connectors in advanced Swedish learners' written English, in S. Granger (ed.) *Learner English on Computer*. London and New York: Addison-Wesley Longman, pp. 80–93.
- Baker, P. (2010). Sociolinguistics and Corpus Linguistics. *International Journal of Corpus Linguistics*. 16. 10.1075/ijcl.16.1.06hun.
- Buttery, P. and A. Caines (2012). Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam/ Philadelphia: John Benjamins, pp. 187–204.
- Capel, A. (2010). A1 – B2 Vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Grammar Profile Journal*, 1(1), pp. 1–11.
- Diez-Bedmar, M. B., & Papp, Sz. (2008). The use of the article system by Chinese and Spanish learners. In G. Gilquin, Sz. Papp & M. B. Díez-Bedmar (Eds.) *Linking up contrastive and learner corpus research*. Amsterdam and New York: Rodopi, pp. 147-175.
- Gilquin G. & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction* 1.1, 41-61.
- Harrison, J., & Barker, F. (Eds.) (2015). *English Profile in Practice. English Profile Studies*, Vol. 5. Cambridge: Cambridge University Press.
- Jones, C., Byrne, S., & Halenko, N. (2018). *Successful spoken English: findings from learner corpora*. London: Routledge.
- Kilgariff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105:116– 127.
- O'Keeffe, A., Clancy, B. and Adolphs, S. (2011) *Introducing Pragmatics in Use*. London: Routledge.
- O'Keeffe, A. and Mark, G. (2017). The English Grammar Profile of learner competence: methodology and finding. *International Journal of Corpus Linguistics*. 22(4), pp. 457–489.
- Trinity College London. (2010). *Graded examinations in spoken English—Syllabus from 1 February 2010*. Trinity College London.