The creation of an academic linguistic corpus: linguistic-terminological analysis in Modern Languages for the training of translators, interpreters and English language specialists

Pablo Maersk Nielsen[*]
Agustina Cintia Savini[**]
Eliana Patricia Heinrich[***]
Denise Johanna Boetsch[****]

[*] Docente de Lengua Inglesa en la Escuela de Lenguas Modernas, Delegación Pilar, de la Universidad del Salvador. Profesor Universitario de Inglés graduado de la Universidad del Salvador. Doctor en Lenguas Modernas y en Educación, por la USAL. Especialista en Gestión Universitaria de la Organización Universitaria Interamericana. (Canadá) Desde 1998 a 2008 fue docente de Lengua Inglesa, Fonética y Dicción. En gestión se desempeñó como Colaborador Académico en el Campus Nuesta Sra. del Pilar desde marzo 1993 a diciembre de 2010. Es Decano de la Facultad de Historia, Geografía y Turismo de la Universidad del Salvador desde diciembre de 2010.

[**] Agustina Savini has a degree in Conference Interpreting in English and English-Spanish Interpreting Teaching from Universidad del Salvador. In 2012, she began her work as a university lecturer at Universidad del Salvador. She lived in England for three years where she took a postgraduate course in Phonetics at University College London (UCL); she obtained the CELTA (International House London); she was the coordinator of the Climate Change Conferences organised by ICCYL in Oxford in 2013 and 2014 and was an English teacher at International House London, where she taught EFL, ESL, EAP and Phonetics workshops. At that time, she began working for the Department of Research and Pedagogy at Cambridge University Press, where she still works as a language researcher. Currently, she is also a lecturer at Universidad del Salvador and at Instituto Superior Santa Trinidad, where she teaches Phonetics and Phonology, English Grammar, English Language and Interpreting. She is a freelance interpreter, editor and proofreader. She also develops coordination activities at Universidad del Salvador, and is an oral examiner for Cambridge Assessment English and Michigan Exams at Buenos Aires Open Centre.

[***] Eliana Heinrich has a B.A. in Conference Interpreting, Universidad del Salvador (USAL), 2012 and a B.A. in Legal Sworn Translation, Universidad del Salvador (USAL), 2009. She was awarded the Highest GPA of 2012. She is an Associate Professor in the Undergraduate Degree in Translation and Interpreting at the Universidad Católica de las Misiones (UCAMI), Misiones, Argentina, a position she has held since 2017. From 2012 to 2020, she was an Associate Professor in the Undergraduate Degree in Conference Interpreting at USAL. In addition, she has been teaching the Trados Studio Workshop since 2018, which has been running for 12 editions. She is also a Researcher Fellow in the Research Institute in Modern Languages at USAL. During the 2014-2015 period, she was an Associate Professor in the Interpreting Diploma Course at the Colegio de Traductores de Tucumán, in conjunction with the Universidad San Pablo de Tucumán. She currently works as Language Manager for an IT company based in Washington, United States.

SUPLEMENTO *Ideas*, III, 11 (2022), pp. 1-5

[****] Denise Boetsch has a B.A. in Conference Interpreting, Universidad del Salvador (USAL), 2011 and has completed the Teacher Training Course at Universidad del Salvador (USAL), in 2012. She is an Associate Professor at the Universidad del Salvador (USAL), Buenos Aires, Argentina, a position she has held since 2015.

Bárbara Bortolato[*****]
Ma. Cecilia Frattin[******]

The term 'corpus' derives from the Latin word meaning 'body', and it may refer to any written or spoken text. Nevertheless, in the field of modern Linguistics, this concept is used as a synonym for vast collections of texts representing a sample of a particular variety or use of language/s that are presented in machine-readable form.

The textual corpus has advanced in the linguistic field relatively recently and has clearly transformed it. Despite the existing quantitative projects, the history of quantitative analysis begins with its appearance. The existence of the corpus, moreover, has made it possible to achieve an unprecedented goal: the access to language in its *"raw and real state"*. Never before had it been possible to obtain thousands of examples of authentic use of language elements.

There are various kinds of corpora, which – for example - may contain written or spoken language, modern or old texts, texts from one language or several languages. The documents might consist of books, academic papers, articles, recordings, among others, of varying length.

It is then possible to categorise corpora depending on the material contained in them and the aim to which they are created. **General corpora** consist of general texts that do not belong to a single text type, subject field, or register. An example of a general corpus is the *British National Corpus*. **Sublanguage Corpora** refer to texts which derive from specific dialects or subject fields.

Corpora can also consist of texts in only one language or of texts in more than one language. **Monolingual corpora,** which are probably the most common ones, contain texts in one language only. In the case of *translated* documents into one language only, for instance, into English, we refer to **parallel corpora**, whereas

In addition, she is also a Researcher Fellow in the Research Institute in Modern Languages at USAL. She currently works at Universidad del Salvador and has been an oral examiner for Cambridge Argentina for 8 years.

***** Bárbara holds a degree in sworn English-Spanish translation and a degree in scientific and literary English-Spanish translation from Universidad del Salvador. She works as a freelance translator in the private sector, and as a certified expert witness in federal courts. She is a lecturer at the School of Modern Languages, Universidad del Salvador, teaching English Language to first, second and fourth-year students. She also lectures on legal English reading comprehension at the School of Law of the Universidad de Buenos Aires.

****** M. Cecilia Frattin es profesora en inglés graduada del IES en Lenguas Vivas "Juan Ramón Fernandez". Se encuentra actualmente cursando la Maestría y Especialización en Educación en la Universidad de San Andrés, y la Diplomatura en diseño y gestión didáctica de propuestas mediadas por TIC en USAL. Realizó el curso online Shakespeare's Life and Work en la Universidad de Harvard y tiene un certificado de enseñanza de inglés para adultos de la Universidad de Cambridge. En 2019 realizó el curso de verano de fonética de UCL, en Londres. En USAL es profesora en la cátedra de Fonética y Fonología Inglesas III, tutora de Lengua Inglesa y profesora en los cursos de idiomas. Se desempeña también como docente en el Instituto del Profesorado de Consudec en las cátedras Didáctica II, Lengua Inglesa III, y Laboratorio y su didáctica I y II. Es supervisora y examinadora de exámenes internacionales desde 2001 y 2009 respectivamente.

corpora containing documents in different languages are called **comparable corpora**, such as the ones mirrored on the Brown Corpus of Standard American English.

Corpora also serve as a basis for a number of research tasks within the area of Corpus Linguistics. In the words of Tony McEenery,

'Corpus linguistics is the study of language data on a large scale - the computer-aided analysis of very extensive collections of transcribed utterances or written texts'.

At Universidad del Salvador, since early 2022, we have been carrying out a project based upon conducting terminological-structural research in English language through exhaustive analysis of the linguistic errors produced by USAL students in the different core areas of the undergraduate courses at the School of Modern Languages, namely, interpreting, translation and English language. The main aim is to generate a database that provides reliable information to benefit the teaching-learning processes, research and the creation and adaptation of didactic material. The final result, thus, will hopefully be an academic linguistic corpus in English. We will now try to provide a brief explanation of our work.

The fieldwork of the project is based on the compilation of students' written and oral productions in order to detect both linguistic and paralinguistic errors. We first focus on the most noticeable features, that is, grammatical, semantic, lexical and phonological issues. Through the systematic use of specific conventions, we have been analysing linguistic variation and frequently used language phenomena by language trainees. We will now explain our approach to this process in more detail.

Our research path is divided into different stages: the first one implies collecting the material that is to be analysed. To do so, we are in constant contact with the School lecturers, who provide us with students' real work. All of this work is unmarked since the task we carry out is different from the corrections made by subject tutors. In addition, it is important for us not to be biased by others' perceptions in more subjective matters that may be related to style or register.

Once we gather the material, we divide it according to specific areas: translation, interpreting, phonetics and phonology and English Language. We will now describe how we operate in each of the disciplines.

In the case of translation, these papers are, in turn, subdivided into sworn translation, scientific translation and literary translation.

The interpreting area consists of consecutive interpreting and simultaneous interpreting. In the latter, we include short-term memory operation activities.

Phonetics and Phonology includes all oral tasks, regardless of the subject area to which they belong, and, in the case of English Language, we focus on written material, namely, narrative, descriptive and argumentative essays, short stories and any other written piece created in this area.

The second stage to the process involves applying two main comprehensive guidelines that we have designed before the start of the project. For oral files, we first need to consider the **transcription conventions** file. This document contains various tags which we use to show exactly what the learner has said, including fillers, false starts, self-correction instances and pausing. We use square brackets and a designated letter/letters to show different speech elements.

Once the audio file has been transcribed, we analyse the student's speech linguistically. To do so, we have created a **pronunciation conventions** file and a general **error tagging conventions** file. In the case of the latter, we also use it for written documents. This thorough guideline contains codes to tackle grammatical, lexical and phonological issues.

Lastly, written pieces are also tagged following a specific **punctuation tagging conventions** file. In the case of all of these guidelines, coded errors are embedded within parentheses.

The differentiation between brackets - used for transcribing purposes - and parentheses - applied in the error coding process-, and the existence of a unique coding system stem from the future goal of developing, together with the IT Department, a digital tool capable of spotting and highlighting tags automatically, which is a key requirement in Corpus Linguistics.

Apart from these more general guidelines, we make use of area-specific guidelines, and this is probably the most groundbreaking aspect of the work we are carrying out.

It is common knowledge that each area poses different and targeted challenges, for instance, in translation, transfer from the source text may lead to unnatural wording in the target text, whereas in interpreting, trainees' use of suprasegmental features, such as pausing and intonation, together with speech delivery choices like register will directly affect the quality of their performance. In the area of English Language, USAL students are trained to grasp and write in natural English, thus working on key concepts such as word economy, and applying specific techniques, like stream of consciousness. All of these features, which cannot be analysed within a grammatical-lexical-phonological scope, are contained within specially-designed guidelines.

Each of the members of this research group has different areas of expertise, so we have divided the creation of these documents according to the knowledge and experience that each of us has in these disciplines. We have then shared our contributions to make any necessary amendments and additions.

In most cases, during the first term, USAL translation students work with documents written in English to translate them into our mother tongue. Hence, we are still at the initial stage of the process in this area. Nonetheless, we have already

established some specific codes related to the translation process: *calque/loan*, *proper name repetition*, *false cognate*, *cultural reference* and *phraseology*.

In interpreting, all of the features we tag belong to the scope of interpreting in the sense that they either affect the content or the package of the message being conveyed. As to the content, the specific codes we use are: *wrong meaning*, *wrong technical terminology*, and *wrong figure*. We also make use of other tags that tend to alter the source message in most contexts, such as *omission* and *addition* of information.

Package-wise, we apply codes such as *unnatural rising/falling intonation*, *long pause*, *utterance break*, *hesitation*, *elongation* and *inappropriate register* (in which case learners tend to produce more colloquial or informal expressions than expected). These are essential in grasping the nature of the interpreting task since many of them might be related to non-verbal issues which pose a real challenge to many trainees, namely, dealing with stress, nervousness and adrenaline.

In the case of phonetics and phonology, we use a separate file which outlines characteristic speech features related to pronunciation and prosodic elements. Given that focusing on all speech aspects might pose an added difficulty at such an initial stage, we decided to first concentrate on the following fundamental elements: phonemes, stress, voicing inflection, strong and weak forms, intonation and lateral allophonic variants. This decision was reached considering the main and most basic speech obstacles that River Plate Spanish speakers tend to encounter.

For the English Language section, we use error tags to highlight elements which do not reflect the idiosyncrasy of the language. Some of these are *lengthy statement*, *word choice*, *L1 interference*, *inappropriate register*, *unclear meaning*, among others.

Once we have transcribed and/or directly tagged errors in each of the documents, there comes the revision and editing stage, where we all look at each other's work for accuracy, ensuring that the samples obtained are an actual reflection of the learner's use of the language in different contexts.

We do hope that linguistic data extraction, storage and analysis within the School of Modern Languages at Universidad del Salvador will benefit the training of language professionals and be the mainstay of future research in the area. Thank you.