RiCL Research in Corpus Linguistics

Designing and building SCoPE²: A spoken corpus of Brazilian Portuguese and L2-English

Giovani Santos
Mary Immaculate College/ Ireland

Abstract – This paper presents the process of designing and building a bilingual spoken corpus in order to pragmatically analyse oral L2-English discourse produced by a group of Brazilian university students living in Ireland. It discusses some of the decisions made, challenges faced, and considerations taken while designing a do-it-yourself corpus with a theoretical framework grounded in Corpus Pragmatics. The main objective is to share the lessons learned by examining the steps of designing and building SCoPE², a bilingual spoken corpus, including the selection of participants, gathering data, and challenges in transcribing and coding spoken language with pragmatics in mind.

Keywords – bilingual corpus; spoken corpus; L2 corpus; corpus design; corpus construction; Corpus Pragmatics

1. INTRODUCTION

Answering many linguistic questions can be done considerably more easily using the increasing range of excellent and freely-available corpora. However, some specialised questions can only be answered by do-it-yourself (DIY) corpora. These are carefully designed and custom-built to assist in answering the questions and meeting the needs of specific research. The reasons to build a DIY corpus can be quite diverse, yet the analysis of such corpora can be very fruitful. Nevertheless, care and attention to corpus design is a must, as ill-conceived or poorly designed corpora can have, among other consequences, a serious negative influence on the research output.

This paper, resulting from a PhD study, presents the process of designing and building a bilingual spoken corpus in order to pragmatically analyse English discourse produced by a subgroup of native Brazilian speakers living in Ireland, with a specific focus on pragmatic markers (PMs).

The study is set within the fusion of Corpus Linguistics and Pragmatics, resulting in the emergence of a relatively new field termed Corpus Pragmatics (Aijmer and Rühlemann 2015). This is a blending which has evolved considerably over the last decade and which has proved to be a beneficial approach to investigating and understanding the usage of many linguistic features of real language-in-use evidenced by the rich expanding body of available studies.[1] Yet, most corpora are not designed for the considerations of pragmatics research and the analysis, categorisation and sometimes tagging of pragmatic features is conducted *post hoc*. It is in this context that there is need for careful consideration when designing a corpus for pragmatic research questions.

2. DIY: A BILINGUAL SPOKEN CORPUS FOR CORPUS-PRAGMATIC RESEARCH

The bilingual spoken corpus presented in this paper is targeted and designed with a context-specific research question in mind, namely to establish whether Brazilian university students in Ireland have accommodated particular features of the Irish English pragmatic system through their exposure to the native environment.

The main objective is that of investigating and describing the presence and function of PMs[2] in the participants' first (L1) and second (L2) languages. In addition, the study aims to compare and contrast the use of these features in the participants' L2 against their mother tongue (Portuguese) and the target language (English), so as to gain insights into the influence of both their L1 and the cultural immersion on their development of an L2.

Due to the specificity of the research question and aims of the study, a DIY corpus was necessitated, this being a bilingual corpus comprised of two sub-corpora, namely L1-corpus (Brazilian Portuguese) and L2-corpus (English as a Second Language), amounting to over 200,000 words across both languages. On the usefulness of small, carefully designed and targeted corpora, McCarthy and O'Keeffe (2010) observe that such corpora show evidence of being powerful tools to investigate and elucidate specific language use.

Another advantage of compiling a small corpus is the fact that the compiler and the analyst are often one-and-the-same: this gives an insider perspective to the analysis,

[1] For studies on different pragmatic phenomena through the lenses of Corpus Pragmatics see Romero-Trillo (2008) and Aijmer and Rühlemann (2015), among others.

[2] This paper subscribes to Fraser (1996) on the concept of PMs as a major umbrella under which many types of PMs constitute the class of "items which mark speakers' personal meanings, their organisational choices, attitudes and feelings" (Carter and McCarthy 2006: 207).

bolstering in the interpretation of the data. This close relationship between analyst and data, as well as language and context, makes small corpora a perfect fit for pragmatic studies. This advances the case for a Corpus Pragmatics (CP) framework (Rühlemann and Clancy 2018), which analyses the data both in a vertical quantitative manner (Corpus Linguistics), and in a horizontal qualitative manner (Pragmatics). This advantageous synergy makes CP a significant framework for the reliable, context-specific analysis of language use and language development.[3] Within the CP framework, the analyst can generate software-driven statistics while, concurrently, undertaking a detailed interpretation of the data considering the three major contexts in Pragmatics: situational, background knowledge and co-textual (Cutting 2002).

According to Jucker *et al*. (2018), CP is one of the three empirical methodological approaches for the analysis of pragmatic phenomena (the others being experimentational pragmatics and observational pragmatics). What is more, PMs are perceived as one of the key areas of corpus-pragmatic research (Clancy and O'Keeffe 2015), which includes, but is not limited to, PMs across languages (see contributions in Aijmer and Simon-Vandenbergen 2006), in L2 (Veiga 2016; Santos 2019), and in comparisons between native and non-native speakers (Aijmer 2004; Fung and Carter 2007).

The following sections describe the phases entailed in the designing and building of a bilingual DIY corpus, namely the *Spoken Corpus of Portuguese and English as a Second Language* (SCoPE²), which include: participant selection, and data collection and transcription. Owing to space constraints, this paper is limited to the description of the L2-corpus.

3. CORPUS DESIGN AND DESCRIPTION

There is common consensus that strict criteria to select the corpus type and participants are fundamental to answer the questions set in the research (Granger 2002; O'Keeffe *et al.* 2007; Adolphs and Knight 2010). A first step of considerable importance to take when designing a corpus is the matter of representativeness. Adolphs and Knight (2010) caution that it is the compiler's responsibility to plan and predict any factor that may be a case for inconsistency and non-homogeneity in the corpus. This process involves not only the

[3] See Clancy and O'Keeffe 2015 for a critical review on research regarding key areas of CP.

selection of the participants, but also the type of material to be produced by them for data collection, as well as elements such as environment and technology.

Regarding this, the criteria for the current research to select the participants and achieve corpus representativeness are as follows:

- L2 context: University students in Ireland
- Mother tongue: Brazilian Portuguese
- Level of English Proficiency: C1-C2 CEFR (Common European Framework of Reference for Languages)

In addition to the criteria described above, the participants must be considered to be in the category of Successful Users of English (SUEs) which, according to Prodromou (2008), entails being able to engage accurately and fluently in different contexts with both L1- and L2- speakers of the target language, but does not necessarily equate to being native like.

The data in SCoPE² represents informal real language-in-use. However, Granger (2002) notes that collecting authentic data of L2 speakers might be a major challenge, as this kind of data is normally collected during task-based activities in class. To avoid the unnatural characteristic of classroom-elicited data, unscripted on-line language in use was collected with the participants' consent during casual meetings between the researcher and friends or fellow university students. The meetings took place in informal public and private settings according to the participants' convenience. The interactions were recorded using the iPhone Voice Memos app, which produced high audio quality (especially when recording dyadic interactions as interlocutors were near the microphone in contrast to one multiparty interaction with participants overlapping while conversing and occasionally moving in the room).

Table 1 details the corpus design in a data matrix.

Number of participants	17
Gender	Mixed (11 female and 5 male)
Participants' age	Adults (20-35 years old)
Nationality	Brazilian
L2 Context	University students in Ireland
Level of L2 Proficiency	C1-C2 CEFR; SUE (Prodromou 2008)
Type of data	Audio recordings of unscripted informal conversations
Type of interaction	13 dyadic (participant + researcher) and 1 multiparty (3 participants + researcher)
Average duration	30 minutes in each language (L1 and L2)

Table 1: Data matrix

These informal conversations have a main topic in common. The participants are mostly discussing their experiences and perceptions of travelling in Brazil and all around the world, though many other topics (such as personal relationships, future goals, etc.) may arise in the conversations due to their natural and unscripted nature. When both contributions (L1 and L2) from a participant were recorded on the same occasion, the researcher greeted them in English and used the L2 as an icebreaker before the actual recording, rather than their common L1. The L2 was thus recorded first, followed by the L1.

It is important to note that the L2 sub-corpus described in this paper is not a learner corpus but, in fact, a corpus of proficient L2 language. This is in view of the fact that all participants in this research have successfully completed their English language programmes, have achieved an internationally recognised certificate of either advanced or proficient English, and communicate efficiently within an international environment both with L1- and L2-speakers from different first-language backgrounds, be it at personal, professional or cultural levels. This perspective of an L2 corpus, rather than a learner corpus, accords with that of Prodromou's (2003) as well as with his search and categorisation of SUEs to build such a corpus.

4. TRANSCRIPTION CONVENTION

The collected data comprises informal real spoken language-in-use, and thus its transcription is required for further corpus-based analysis. As noted by Kirk and Andersen (2016: 295), transcriptions are a representation of spoken language which is subjected to a process of "selection, abstraction and omission." It is, therefore, the transcriber's responsibility to ensure that the transcription is as truthful to its spoken version as

possible, though the amount of detail, marking and coding will always depend on the specific research needs (Adolphs and Knight 2010). This section presents the rationale behind the choices for and needs of the transcription of a corpus designed to analyse PMs within a context of L2 development.

4.1. Transcribing spoken language

Due to the multimodal nature of spoken language, the practice of transcription can pose a real challenge when it comes to deciding what is to be included from the array of detailed layers which can be extracted from an interaction (Adolphs and Knight 2010). Spoken interaction is not simply comprised of utterances alone, but these function alongside several extralinguistic features such as tone, rhythm, laugh, eye gaze, and body movement, in order for a speaker to convey a message and their listener(s) to infer the intentional meaning in the speaker's proposition. As Adolphs and Knight (2010: 44) aptly put it, spoken interaction features "a careful interplay between textual, prosodic, gestural and environmental elements in the construction of meaning."

Some researchers may wish to build sound-text aligned corpora, providing resources to undertake studies focused not only on the structure of the spoken language, but also on its prosodic features (see contributions in Cresti and Moneglia 2005). Others may focus on a wide range of pragmatic phenomena, thus annotating their corpora in order to investigate pragmatic intent (Kirk 2016). For those who are interested in the understanding of talk in interaction, Conversation Analysis provides a thorough transcription convention in order to analyse interactive features such as overlapping, prosody and non-verbal elements (Liddicoat 2007).

Adolphs and Knight (2010) flag the individuality of each research project and the importance of identifying with precision the purpose of the study prior to selecting the type of transcription for the study in question, noting that "[i]t is advisable to identify the spoken features of interest at the outset, and to tailor the focus of the transcription accordingly" (2010: 44). With that in mind, SCoPE² was built to analyse PMs in L2, with a view towards the influence which the speakers' L1 and their exposure to the target language may have on their production of such linguistic features. This means that a detailed annotation of different pragmatic phenomena was not necessitated, as the type of corpus methodology adopted was that of a form-to-function, rather than function-to-form

(O'Keeffe *et al*. 2019). In other words, frequency and keyword lists, as well as previous studies on the linguistic feature under investigation, are used as starting points, thus a broad transcription with minimum annotation was sufficient.

4.2. Transcription convention: SCoPE²

The transcription convention used in the structure of SCoPE² (e.g. codes, tags, and punctuation of both sub-corpora) was adapted from that employed in the transcription of the *Limerick Corpus of Irish English* (LCIE; Farr *et al.* 2004) which, in turn, was based on the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE; McCarthy 1998). This was a natural choice since the LCIE is used as a reference corpus in the study for which SCoPE² was built, as it represents the variant of English to which the participants have been exposed. Similarly, the transcription convention for the linguistic material of the L2-corpus of SCoPE² (e.g. filled pauses, backchanneling, contractions, etc.) was adapted from that of the LCIE. However, the transcription convention for the linguistic material of the L1-corpus was adapted from that of the C-ORAL-BRASIL (Mello *et al*. 2012), chosen as a model due to its thorough description of decisions and rationale when working with spoken Brazilian Portuguese.

Although some adaptations were needed, and a parallel transcription was not necessarily required between the LCIE and SCoPE² (due to different research foci), it is useful to try to bring a corpus as close as possible to its reference when it comes to the transcription, as this will facilitate future data reading and analysis. Most of the codes used in the LCIE transcription maintained their original functions in both SCoPE^2 sub-corpora in their original functions, while others were slightly adapted within their existing functions. It is also important to note that, despite obvious variations regarding the transcription of linguistic material between two languages (e.g. English backchannelling *mmhm* versus Brazilian Portuguese backchannelling *hum hum*; English language contractions such as *it's* versus Brazilian Portuguese apheresis in the conjugation of the verb *estar* 'to be' in nearly all of its conjugations, e.g. *tá*, *tava*, *tô* etc.), the two parts of SCoPE² are fully comparable in their structure and codes. The remainder of this section is devoted to describe these adaptations and their justifications.

4.2.1. General codes

For the identification of speakers, each participant is given a number according to their gender and order of file transcription, the odd numbers being males and the even numbers females.

To ensure the anonymity of anyone mentioned throughout the conversations, their names are replaced by either the speakers' own identification numbers (e.g. $2 if speaker two mentions her own name) or a name that reflects the culture and/or language of the names that are mentioned. The string (1) gives an example where speaker $2 is talking about her trip to Cuba when she mentions her hosts' names, who are both Cuban, and had their names replaced by names which are also Spanish sounding.

(1) <$2> Yeah er the the family we stayed with the first one **<$E> pause </$E>** **Juan** and **Rosa** uhm she was a psychologist and she used to get paid uhm eighteen euros a month.

As shown in (1), extra linguistic features are transcribed within the **<$E> </$E>** codes. Not only do they include significant extra information that happens during the interaction (e.g. pause, laugh, etc.), but they also include contextual and/or cultural background information (e.g. explanations of word play that draw on common cultural understandings).

As far as extra linguistic information is concerned, two features –laughs and pauses– which are natural to spoken interaction required further description. Laughing is thus divided into three categories: *laugh*, which refers to a loud and free expression of amusement; *chuckle*, a shorter and inner type of laugh; and *giggle*, describing an even shorter and lighter type of laugh. In the L1-corpus, these pieces of information are transcribed as *risada*, *risos* and *risadinha*, respectively.

Although pauses, at first glance, seem to be quite a straightforward feature to transcribe using the pairs short/long and filled/unfilled as a reference model, a system based on specific criteria needed to be developed to ensure consistency and ease of reading throughout the data. Having said that, it is important to note that filled pauses are only transcribed one way in each language, namely *uhm* for English and *eh* for Portuguese, rather than trying to develop an infinite and exhaustive list of different sounds produced by speakers when filling a pause (especially in L2 production). While *uhm* and *eh* may not cover all sounds produced by the participants either in English or Portuguese, they represent the act of filling space in language use while speaking. The choice for *uhm*

and *eh* over many other graphic forms of filled pauses is due to the unlikelihood of their being similar to any written word in either languages in the data, thus avoiding any possible misinterpretation when reading the texts.

Unfilled pause codes, on the other hand, serve to mark where pauses take place, either within an utterance or after it. Examples (2) and (3) present speakers' turns where it is possible to see two different functions in the use of such a feature. In (2), speaker $2 makes use of a short pause to give emphasis to the adverb *actually* when asking speaker $1 if he had been to Paraguay, whereas in example (3) a pause is employed to allow time to the speaker in order to restructure a sentence. These are examples of short pauses (up to three seconds). Longer pauses are coded according to their length, e.g. <$E> pause of six seconds </$E>.

(2) <$2> +but <$E1> chuckle </$E1> did you go there **<$E> pause </$E>** actually?

(3) <$1> I think there is always **<$E> pause </$E>** there are always both sides you know+

A break between two or more utterances produced within a speaker's turn with no interruption is marked with a full stop. A full stop, therefore, marks where one complete utterance finishes and another starts, as well as where a speaker's turn ends, as seen in example (4). Complete utterance refers to a complete thought, which can be either a full sentence or simply a phrase, and can occur as a full turn or within a turn. Alternatively, when a speaker's turn is not finished but still slightly interrupted by a short answer or backchannelling, the plus symbol (+) is used to mark such a phenomenon of speech continuity (see (4)).

(4) <$3> I don't know it's a tough question <$E> pause </$E> now I'll continue the list and **then+**
<$1> Oh please yeah **yeah.**
<$3> **+and** then I think about the favourite **place.** Okay <$E> pause </$E> after Lithuania I went to Tunisia+

Following the same rationale behind the use of a full stop to mark the end of an utterance, the equal symbol (=) is employed to identify where incompletions take place in the conversations. At a lexical level, an incomplete word can be either followed by its full form, as in the truncation seen in (5) where speaker $5 is talking about free walking tours, or followed by another different word altogether, as in (6), where speaker $3 restructures his sentence, changing its aspect.

(5) <$5> But you but you kinda feel bad when they like they don't charge anything but **the= they** ask for tips+

(6) <$3> +if you compare like to England. I **I don= I've** never been to England to be honest yeah no.

Alternatively, at a speaking-turn level, a word or an utterance may be simply interrupted by another speaker, as shown in (7).

(7) <$4> Yeah. **So**=
<$1> How old are you?
<$4> I'm twenty-four.

4.2.2. Transcription issues regarding PMs

Considering that SCoPE^2 was designed with the particular research purpose of analysing PMs, two issues required particular attention during the transcription process, namely overlapping and ambiguity.

Although overlapping is an important feature in spoken interaction, the research for which SCoPE^2 has been designed is not grounded on Conversation Analysis, where features such as overlapping, as well as pauses, intonation, repair, etc. must be carefully marked. Therefore, overlapping was not marked because, besides being a time-consuming task, that would result in an over-coded text rather than aiding in the analysis.

Nevertheless, overlaps still pose transcription challenges in terms of how to maintain a natural and truthful flow of spoken language without marking where the overlaps take place in the conversation. An attempt to try to overcome some of these challenges is to follow the natural sequence of events in the conversation. Extract (8a) demonstrates where the overlaps take place (marked with the <$O> </$O> codes), while extract (8b) presents the same piece with the overlaps replaced by interruption marks (+ symbol) breaking the interaction in a natural sequence of events. In this interaction, speaker $2 is explaining to speaker $1 where the state of Paraná is located in Brazil.

(8a) <$1> So you are from the last state like the last one in the south.
<$2> Mm **<$O1> uhm </$O1>.**
<$1> **<$O1> Paraná </$O1>** <$E> trying to locate Paraná in an imaginary map in the air </$E>.
<$2> No no no it's **</$O1> Paraná </$O1>** Santa Catarina and Rio Grande do Sul.
<$1> **<$O1> Rio Grande do Sul </$O1>.**

(8b) <$1> So you are from the last state like the last one in the south.

<$2> Mm **uhm**.
<$1> **Paraná** <$E> trying to locate Paraná in an imaginary map in the air </$E>.
<$2> No no no it's **Paraná+**
<$1> Rio Grande do Sul.
<$2> **+Santa Catarina** and Rio Grande do Sul.

Special attention, however, had to be given to the position of PMs in the utterances when breaking overlaps into turn continuity. Occurrences of minimal response tokens (e.g. *yeah*, *okay*, *mmhm*, *mm*) overlapping PMs, for example, were transcribed having PMs as priorities due to their multi-functionality. Considering the importance of the relationship between position and function when it comes to PMs, two linguistic phenomena were assessed: the co-occurrence of PMs and their prosodic aspects (e.g. short pause before or after the PM, rhythm, stress, pitch movement when delivering the PM). Extract (9a) illustrates a case where *you know* and *like* co-occur when speakers $1 and $2 are talking about how difficult flying to Cuba is. Speaker $1 overlaps $2 in between the pair of PMs:

(9a) <$2> Mm yeah so going to Cuba is sort of tricky **<$O1> you know </$O1> like** there are some since of the the embargo like the the United States embargo in Cuba <$O1> like </$O1> it's sort of like not every uhm=
<$1> **<$O1> Mmhm </$O1>.** <$O1> Ah okay </$O1>. It used to. Not anymore.

To avoid a break between the original co-occurrence of two PMs which, in turn, avoids misinterpretation of their functions, the response token (RT) was placed in the next line after the PMs, as illustrated in (9b):

(9b) <$2> Mm yeah so going to Cuba is sort of tricky **you know like**+
<$1> **Mmhm.**
<$2> +there are some since of the the embargo like the the United States embargo in Cuba like+
<$1> Ah okay.
<$2> +it's sort of like not every uhm=
<$1> It used to. Not anymore.

Likewise, PMs were kept in their final or initial positions based on their prosodic aspects or way in which they were originally uttered by the speaker. This is illustrated by extracts (10a–b), where speaker $2 is talking about her motivation to go to Cuba.

(10a) <$2> +already and uhm I had a job in Brazil I I graduated from law school and got a job as a legal adviser and I knew I was going to move to Europe to do a Masters and I I said "oh I'm going to be poor <$E1> laugh </$E1> in a while" <$E> pause </$E> like not poor like <$E> pause </$E> but "I'm going to be a student again <$O1> in a while </$O1> so now that I'm like getting **<$O1> paid"</$O1> you know** "if I don't go to Cuba now"+

<$1> <$O1> Yeah </$O1>. **<$O1> Mmhm </$O1**>.

(10b) <$2> +already and uhm I had a job in Brazil I I graduated from law school and got a job as a legal adviser and I knew I was going to move to Europe to do a Masters and I I said "oh I'm going to be poor <$E1> laugh </$E1> in a while" <$E> pause </$E> like not poor like <$E> pause </$E> but "I'm going to be a student again in a while+
<$1> Yeah.
<$2> +so now that I'm getting **paid" you know**+
<$1> **Mmhm**.
<$2> +"if I don't go to Cuba now"+

As seen in (10a), the overlap takes place before *you know*. However, moving the position of the PM from after the word *paid* to before the word *if* would result in a change of focus on the content being delivered. By listening to the recording, the researcher was able to ascertain the position of the PM within the speaker's turn by taking into consideration prosodic cues. In this specific case, the word *paid* is linked to *you know* in the rhythm of the speech delivered, and the PM *you know* is then followed by a short pause before the speaker continues with her speech. Therefore, again, the position of the PM had priority over the position of the RT, as illustrated in (10b).

A case of potential ambiguity for the interpretation of PMs surfaced during the transcription process. The transcription of SCoPE² has only three pieces of punctuation (following the convention of the LCIE), the question mark being for questions, inverted commas for quotes and the full stop to close complete utterances. Consider examples (11) to (13) below:

(11) <$1> You don't **like** dogs?

(12) <$2> I think I've been **like** to twenty twenty and a few countries.

(13a) <$3> Uhm it's just **like** a big city.

In (11), it is clear that *like* functions as a verb, while in (12) it functions as a PM. However, it is difficult to assign either a grammatic or a pragmatic function to *like*, as in (13a) it could be either of the options. Once again, during the transcription, prosody played an important role when differentiating between grammar and pragmatics. If the sentence in (13a) were uttered straight on with no hesitation, then it would be a case of *like* as a preposition. If the example were uttered with a slight pause between *like* and a *big city*, though, then this would be a case of PM, thus being transcribed with the **<.>** code as illustrated in (13b).

(13b) <$3> Uhm it's just **like** <.> a big city.

4.2.3. Customised new codes

Two more codes were included in the transcription convention in order to attend to the research design of SCoPE[2], namely code-switching (**<$CS> </$CS>**) and language error/correction (**<$X> | </$X>).**

The first one, **<$CS> </$CS>**, refers to any code-switching occurring in either sub-corpus. The second one, **<$X> | </$X>**, is employed when a language error or mistake needs correction to avoid misinterpretation or confusion. The interpretation and correction of such mistakes were facilitated by the fact that the transcriber shares with the participants the same L1. The use of an error code was not a surprise due to the nature of the English sub-corpus, where the participants, although at proficient level of competency in English, were still developing their L2, and thus were not expected to be totally error-free.

Extract (14) illustrates a case where an entire string is composed of a series of mistakes which confuse the message. Without correction, it is unlikely that it would be understood by either an L1-user of English or L2-users with different language backgrounds. Speaker $4 is talking about her willingness to come to Ireland for a year of study-abroad experience and her decision to break up with her boyfriend if he had decided not to come with her. Everything added after the vertical bar symbol (|) is the correction offered by the researcher:

(14) <$4> +"if you're not we are going to **<$X> finish | break up </$X>** right away here because I'm not going to **<$X> sustain | maintain </$X>** a relationship **<$X> so far| over such a long distance </$X>**+

5. CONSIDERATIONS AND CONCLUSION

As seen throughout this paper, the process of designing and building a totally new spoken corpus entails a series of steps that progress from the research questions to the transcription of the data. The type of research one is conducting plays a significant role in this process and so does the nature of spoken language itself, which can, at times, be incoherent and challenging, and thus pose many difficulties when giving it a written form.

To achieve a satisfactory standard in the data, a meticulous review of the work must be conducted with trials and plentiful patience. A number of pilot transcriptions must be undertaken and reviewed before a final and appropriate model can be achieved. The model described in this paper has been reviewed during its development in different ways:

- by reading a sample without listening to the audio in order to check if the transcription of spoken language is coming across as a true representation of the recorded interaction;
- by reading a sample while listening to its audio in order to double check the impressions during the first reading;
- by testing the tags/codes with CL software;
- after attending conferences and meetings on the field in order to gain feedback on the proposed coding system;
- by testing the convention against existing ones.

Another factor to consider when dealing with spoken language is that data collection may affect the transcription process and therefore special attention must be paid to elements such as background noises, the type of recording devices used and the number of speakers in the recorded conversation. For example, considering the natural and sometimes chaotic nature of spoken language, a multi-party conversation may actually be quite fruitless if not well managed, since the level of overlapping can be higher in comparison to a dyadic interaction, and the speakers may go off topic by interacting with different parties at the same time. All this natural messiness and diversity in language might be valuable for Conversation Analysis, but poses many challenges in collecting and maintaining a certain level of quality and truthfulness to spoken language.

Delaying the transcription of interactions can also cause significant problems. As previously mentioned, being the one-and-same data collector and transcriber contributes valuable insights and depth to the conversations and interactions recorded and transcribed. However, it is good practice and strongly recommended that transcriptions be carried out in the shortest time possible after the material has been collected, especially as many of the resources used to aid in online communication, such as body language and facial expressions, are still fresh in mind and, therefore, most helpful in aiding the transcription process and resolving any possible uncertainties or ambiguities that may arise.

In conclusion, though laborious, the DIY is a worthwhile type of corpus. The entire process requires one to reflect on the research questions and contemplate the analysis phase of the research. In addition, although these corpora may often be small in size, a wide range of linguistic features can be extracted for detailed analysis, making such corpora a rich source for corpus studies.

Each DIY corpus compiler may have their own rationale and theories to ground their design and construction. This paper has provided some practical experience and reflections on how to design and build a bilingual corpus to conduct language analysis within a Corpus Pragmatics framework. It is hoped that this work may open more discussions on how to achieve truthfulness and accuracy as close as possible to spoken language. Moreover, it is hoped that this may be a guiding resource for those aiming to venture on a DIY corpus journey for the first time, doing so in a critical way.

REFERENCES

Adolphs, Svenja and Dawn Knight. 2010. Building a spoken corpus: What are the basics? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 38–52.

Aijmer, Karin. 2004. Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies* 3/1: 173–190.

Aijmer, Karin and Anne-Marie Simon-Vandenbergen eds. 2006. *Pragmatic Markers in Contrast*. London: Elsevier.

Aijmer, Karin and Christoph Rühlemann eds. 2015. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Clancy, Brian and Anne O'Keeffe. 2015. Pragmatics. In Douglas Biber and Randi Reppen eds. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 235–251.

Cresti, Emanuela and Massimo Moneglia eds. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

Cutting, Joan. 2002. *Pragmatics and Discourse: A Resource Book for Students*. London: Routledge.

Farr, Fiona, Brona Murphy and Anne O'Keeffe. 2004. The Limerick corpus of Irish English: Design, description and application. *Teanga* 21: 5–29.

Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6/2: 167–190.

Fung, Loretta and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28/3: 410–439.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Granger, Sylviane, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 3–33.

Jucker, Andreas H., Klaus P. Schneider and Wolfram Bublitz. 2018. Preface to *Methods in Pragmatics*. In Andreas H. Jucker, Klaus P. Schneider and Wolfram Bublitz eds. *Methods in Pragmatics.* Berlin: Mouton De Gruyter, x.

Kirk, John M. 2016. The pragmatic annotation scheme of the SPICE-Ireland corpus. *International Journal of Corpus Linguistics* 21/3: 299–322.

Kirk, John M. and Gisle Andersen. 2016. Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21/3: 291–298.

Liddicoat, Anthony J. 2007. *An Introduction to Conversation Analysis*. New York: Continuum.

McCarthy, Michael. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, Michael and Anne O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3–13.

Mello, Heliana, Tommaso Raso, Maryualê M. Mittmann, Heloísa P. Vale and Priscila O. Côrtes. 2012. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: Critérios de implementação e validação. In Tomasso Raso and Heliana Mello eds. *C-ORAL-BRASIL I: Corpus de Referência do Português Falado Informal*. Belo Horizonte: Editora UFMG, 125–176.

O'Keeffe, Anne, Brian Clancy and Svenja Adolphs. 2019. *Introducing Pragmatics in Use* (second edition). London: Routledge.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Prodromou, Luke. 2003. In search of the Successful User of English: How a corpus of non-native language could impact on EFL teaching. *Modern English Teacher* 12/2: 5–14.

Prodromou, Luke. 2008. *English as a Lingua Franca: A Corpus-based Analysis*. London: Continuum.

Romero-Trillo, Jesús ed. 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente.* Berlin: Mouton de Gruyter.

Rühlemann, Christoph and Brian Clancy. 2018. Corpus linguistics and pragmatics. In Cornelia Ilie and Neal R. Norrick eds. 2018. *Pragmatics and its Interfaces*. Amsterdam: John Benjamins.

Santos, Giovani. 2019. Second language pragmatics: A corpus-based study of the pragmatic marker *like*. *Letrônica* 12/4: 1–16.

Veiga, Nancy V. 2016. Discourse markers in CEDEL2 and SPLLOC corpora of learner Spanish: Analysis of some lexical-pragmatic failures. In Margarita Alonso-Ramos ed. *Spanish Learner Corpus Research: Current Trends and Future Perspectives*. Amsterdam: John Benjamins, 267–297.

Corresponding author
Giovani Santos
Mary Immaculate College
Department of English Language & Literature
South Circular Road
V94 VN26, Limerick
Ireland
e-mail: giovani.santos@mic.ul.ie

received: February 2020
accepted: March 2020