Scouting for Data: Creation Process of a Non-Native Speakers' Specialized Spoken Corpus

En búsqueda de datos: Proceso de creación de un corpus oral especializado de hablantes no nativos

Claudia Mejia De la Pava
Universidad de Salamanca
España
clm.delapava@usal.es
https://orcid.org/0000-0003-4173-4900

Resumen

La disponibilidad de material oral especializado en el campo del vuelo recreativo en general y, específicamente, en el subcampo del ala delta, es escasa. Por lo tanto, para realizar un estudio del inglés hablado por los hablantes no nativos al interior de la comunidad internacional del ala delta, fue necesario que la autora recopilase sus propios datos y compilase el corpus especializado que procedería a analizar. El proceso, que involucró varias fases, comienza con la definición de la población universal y la muestra, continúa con el diseño e implementación de una entrevista y es seguido por la transcripción de las grabaciones, que condujo a la creación de los archivos finalizados. Este artículo presenta una descripción detallada del proceso de compilación de este estudio basado en corpus y abarca material que podría ser útil para otros investigadores que estén llevando adelante o estén considerando embarcarse en proyectos similares. Incluye además algunos ejemplos de resultados preliminares, dado que se trata de un proyecto en curso.

Palabras clave: estudio basado en corpus, ESP, corpus especializado de inglés hablado, compilación de corpus, transcripción.

Abstract

The availability of specialized oral material in the field of recreational flying in general and, specifically in the sub-field of hang gliding, is scarce. Therefore, in order to conduct a study of spoken English produced by non-native speakers within the international hang gliding community, it became necessary for the author to collect her own data and compile the specialized corpus she would then analyze. The process involved several phases, which

began with the definition of the universal and sample populations, then the design and implementation of an interview, followed by the transcription of the recordings leading to the creation of the finalized files. This article presents a detailed description of the compilation process of this corpus-based research and covers some material which could be helpful to others pursuing or considering to embark in similar projects. It also includes some examples of preliminary findings, since it is still an ongoing project.

Keywords: corpus-based study, ESP, specialized spoken English corpus, corpus compilation, transcription.

Introduction

The study of the non-native speaker (NNS) English spoken by a specific, international and non-academic community, is at the base of this research project. From this perspective, among others, it can be considered a study within the intersection of different areas, such as: ESP (English for Specific Purposes) because, this community's, interactions lie in the specific field of recreational aviation, ELF (English as a Lingua Franca) since English is used as a vehicular language during interactions between members who do not share the same native language (Mauranen, 2003); also, corpus linguistics, multicultural communication and specialized translation. From the very beginning, one of the key aspects of this project was to implement an ethnographic approach, considering that "language use cannot be properly described and understood outside its context of social use" (Dressen-Hammouda, 2013, p. 501).

The intention is to gain understanding of the features and dynamics found within the language variety used by this specific group of speakers and apply the findings to aid the development in areas such as specialized translation, multicultural communication and learning English as a foreign language, keeping in mind that "[t]he teaching and learning of writing / reading and speaking / listening skills can also gain from the availability of corpora in the classroom" (Bernardini, 2003, p. 534). All this, accompanied by implementing and adhering to the methodologies of the different aspects this study covers, in order to avoid or minimize certain "theoretical and methodological problems that are relevant to the whole of the social sciences" (Hammersley, 2007, p. 691). Because of the lack of material available in this specific field, it was necessary to collect our own data in

order to be able to carry out the study. For this article, we intend to focus on presenting the methodology used during the different phases carried out so far, with certain emphasis on the collection and processing of the data. Moreover, in order to cover a probable lack of knowledge of the potential readers, it was deemed important to begin by presenting some relevant general information regarding the community under study, in order to help the reader, get situated and have a better understanding of this kind of recreational aviation and its community. In order to be consistent with our presentation of the methodology, this first part, will also include a summary of the process followed to define the sample population. The following sections of this article aim to provide a deeper look into the data collection and processing phases, with emphasis in the latter. The last section of the paper will cover some of the preliminary findings (the analysis has not yet been concluded) and it will end with the forthcoming steps required to bring this process to its culmination.

1. Hang Gliding: A Community within the Sport

Bearing in mind that this study is not centered around a strictly academic community and that hang gliding (HG) might not be well known by readers, let us begin with a description of this practice and move on to portraying the community that revolves around this sport.

1.1. The Sport

Hang gliding can be defined as the "sport of flying in lightweight unpowered aircraft which can be carried by the pilot" (Whittal). It is governed by the *Commission Internationale de Vol Libre* (CIVL) and the *Fédération Aéronautique Internationale* (FAI / World Air Sports Federation) both of which are headquartered in Lausanne, Switzerland and recognized by the International Olympic Committee.

In hang gliding, a pilot may launch on foot by running off a mountain launch ramp, or on aerotow from an airstrip, where the hang glider is towed by a light powered aircraft or a winch. Once in the air, pilots look for ascending currents of air and try to stay aloft either for as long as possible or as long

as necessary to either cover certain distances in the fastest time possible, or to fly as far as possible. Pilots lie horizontally inside a harness and steer the hang glider solely by means of shifting their own weight; there are absolutely no control surfaces on a hang glider.

There are numerous international hang gliding events every year, hosted in many countries around the world. Some are merely recreational, but most of them are competitive events that vary in duration between three and fifteen days, depending on their level. Competitive events typically range from a weekend regional league or beginners' meet to a world or continental championship. The frequency and variety of organized events help support the description of this community as an active and cohesive one; for instance, in 2017 – the baseline year of this research – the official FAI/CIVL records showed that, in a span of the previous five years, a significant number of competitions took place as, illustrated by **Figure 1**: Over 100 officially registered events took place every year, averaging 133 per year, not taking into account any of the other smaller unsanctioned gatherings organized by different associations.
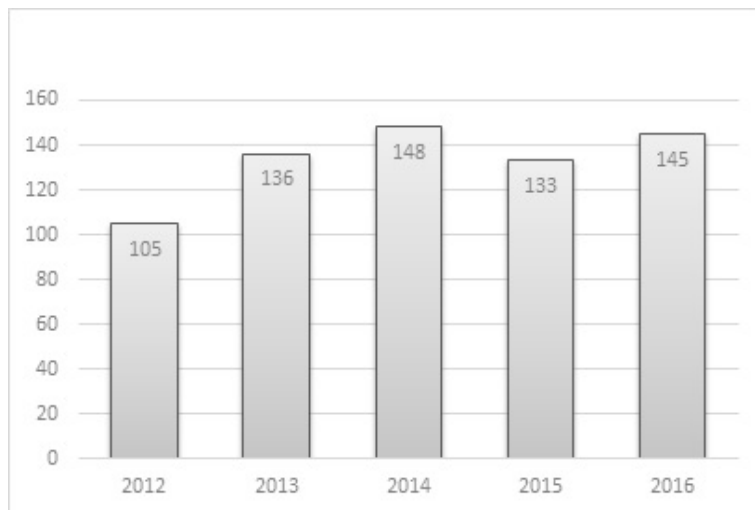

Figure 1. 2012-2016 International Hang Gliding Competitions[1]

---

[1] Designed by the author with data taken from the FAI/CIVL 2016 Competition Coordinator Report.

1.2. The Community

According to the same official FAI/CIVL database, more than two thousand pilots were registered in 2017 from fifty countries altogether. This information, supports the description of this community as being an international and multicultural group, as depicted by **Figure 2** below:
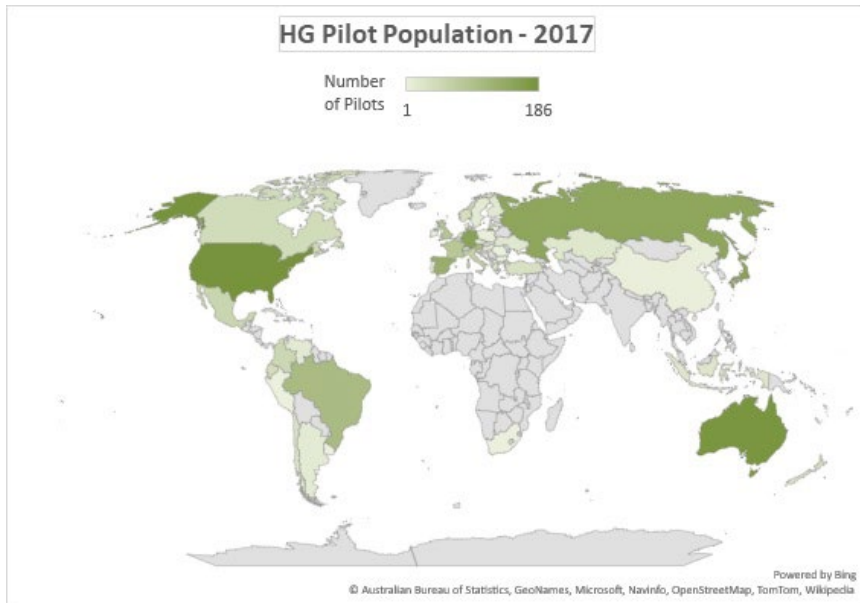

Figure 2. FAI/CIVL Hang Gliding Pilot World Population[2]

The HG community is comprised of current and/or former pilots who may also play other roles such as manufacturers, instructors, event organizers, staff volunteers as well as hang gliding enthusiasts; all this helps

---

[2] Data taken from the FAI/CIVL World Pilot Ranking System (http://civlrankings.fai.org/?a=326& ladder_id=1&ranking_date=2017-03-01&) and processed into the map in **Figure 2**, which is powered by BING and uses data from the Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom and Wikipedia.

corroborate that this community can also be defined as heterogeneous and cohesive.

Moreover, its members have various linguistic backgrounds; yet, as dictated by the FAI/CIVL rules and regulations, the official language of the international events is English. The regulations state that within the CIVL sanctioned competitions, despite the geographic region they may take place in, "[t]he rules, regulations and information circulated to NACs [National Aero Clubs] and competitors or issued during the event shall be in English" (FAI, 2018, p. 38). This includes local event rules, publication of provisional and final results, official communications, briefings among others, which also lead, informally, to making English the contact language between the participating members who do not share a common native language. Many of these members have studied English formally, while several others have acquired most of their knowledge of the language independently over the years. Continuing education opportunities arise for all NNS members of the hang gliding community during their participation in multiple events and through direct interactions with their peers.

These characteristics account for the researcher's interest in taking an in-depth look at the English-language interaction and communication dynamics of a representative sample of NNS members, looking for data that could be expanded to help develop fields such as translation and teaching.

After this short introduction to the sport of hang gliding and a brief description the community around it, highlighting the international, multicultural, multilingual, active, cohesive and heterogeneous nature group, we will describe the methodology applied to select a sample population for the study.

2. Sample Population

After outlining the overall community and making the decision of studying a sub-set of the population as opposed to all of its individuals (Gobo, 2004, p. 437), it was important to streamline the population in order to have a

representative sample group to work with. As Neuman explains, "we select some cases to examine in detail, and then we use what we learn from them to understand a much larger set of cases" (2014, p. 246). In order to facilitate a better understanding of the methodology followed to attain this sample group, a summarized description of this process follows.

The information used for this purpose came mainly from the official databases of the FAI's/CIVL's official pilot registry, complemented with in-depth knowledge and field observations by the researcher throughout many years. The list of countries represented within the community, in March 2017, obtained directly from the hang gliding registry is illustrated by **Table 1**:

Table 1. Registered Countries in the FAI/CIVL (alphabetical order: top to bottom/left to right)

Argentina	Croatia	Hungary	Mexico	South Africa
Australia	Czech Republic	Indonesia	New Zealand	Spain
Austria	Denmark	Ireland	North Macedonia	Sweden
Belgium	Ecuador	Israel	Norway	Switzerland
Brazil	Finland	Italy	Peru	Netherlands
Bulgaria	France	Japan	Poland	Turkey
Canada	Germany	Kazakhstan	Portugal	Ukraine
Chile	Great Britain	Korea	Romania	United States
China	Greece	Lichtenstein	Russia	Uruguay
Colombia	Guatemala	Lithuania	Slovenia	Venezuela

On the basis of language, number of registered pilots and active participation in international events, some countries falling within one or several of these categories were filtered out. In other words, those countries with a) English as their main/official language, b) with low

participation (3 events or less – taking into account the records of 2016 up to the first quarter of 2017, along with field observations of previous years) and c) with less than 0.1% of registered members (3 pilots) were removed from the list: **Australia**, **Bulgaria**, **China**, **Croatia**, **Finland**, **Great Britain**, **Indonesia**, **Ireland**, **Israel**, **Kazakhstan**, **Korea**, **Lichtenstein**, **Lithuania**, **New Zealand**, **Peru**, **Romania**, **South Africa**, **United States** and **Uruguay**.

The remaining 31 nations were then grouped into languages, based on each country's official language(s) and information collected through field observation. This change from 'nations' to 'languages' was necessary in order to accomplish the fundamental shift from a geopolitical to a, primarily, linguistic scope; for example:

(1) Brazil and Portugal ➔ Portuguese

(2) Argentina, Guatemala and Mexico ➔ Spanish

(3) Belgium ➔ Dutch and French

(4) Switzerland ➔ French, German and Italian

The outcome was a total of 20 languages, which were then divided into 3 subgroups according to hang gliding population size. Finally, a minimum and a maximum number of participants was set for each subgroup with the purpose of having a range of possible participants for the study. The resulting subsets of languages and number of interviews are represented in **Figure 3** below:
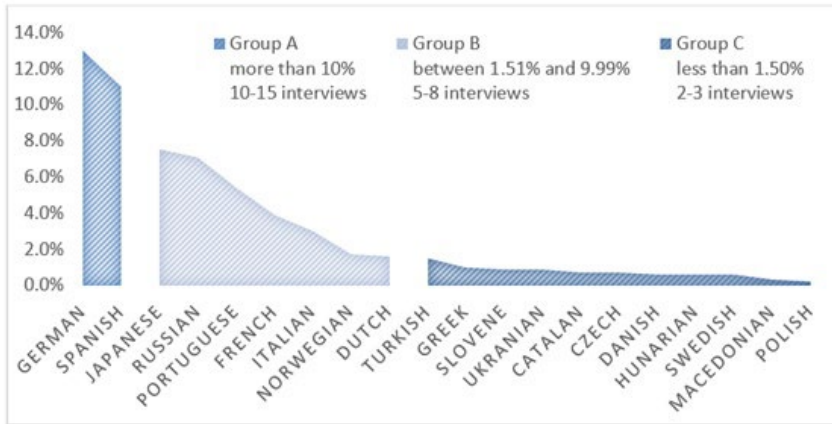
Figure 3. Language Subgroups by Percentage and Range of Possible Interviewees per language

Therefore, it can be said that this sequence of steps led to an initial sample size ranging from 77 to 119 participants in total, from the 3 different subgroups: Group A (more than 10% of registered pilots), Group B (between 1.51% and 9.99% of registered pilots) and Group C (less than 1.50% of registered pilots). The itemization of the subgroups is shown in **Table 2**:

Table 2. Language subgroups and number of interviews (range) per group and total range

Group	Number of Interviews
A	20 – 30
B	35 – 56
C	22 – 33
A + B + C (Total)	77 - 119

Once the sample group outline and size were defined, in order to "reflect particular features of or groups within the sampled population" (Ritchie *et*

al., 2003, p. 78), the next step was to move on to the actual interviewing phase, also referred to here as the data collection process.

Let's proceed to the description of the methods implemented to collect the data, which at a later stage was used to create the specialized corpus mentioned during the introduction to this article.

3. Data Collection Methods

As mentioned above, this section covers the process and methods followed during the data collection phase of the project, i.e., the interviewing process, which is comprised of three stages. In an attempt to provide a more straightforward way to present the information collected, we have grouped this section into two different segments: the design of the questionnaire and the planning and implementation of the actual interviews.

By implementing an ethnographic approach for this research project, the aim was to be able to "combine participation and observation so as to become capable of understanding the experience as an insider while describing the experience for outsiders" (Genzuk, 2003, p. 2). For this purpose, an interview-based method was chosen as the main approach, supported by field observations given that, as several researchers have indicated, among them Gobo (2011), "a questionnaire or an in-depth interview, if conducted correctly, is more likely to obtain similar replies (reliability) regardless of who the interviewer is" (p. 28).

3.1. Interview Design

Taking into account that an interview could serve as a tool to focus on the topic of what the interviewees would probably say or talk about, it would most likely yield a fairly homogeneous content, which could also enable a possible comparative view of the English used by the different participants. Nevertheless, the main purpose of the interview itself was to elicit the participants to speak about hang gliding using different language registers

as well as covering various specialized aspects of the activity. With this in mind, the first choice made was to use solely open-ended questions avoiding any type of question that could trigger a single-word type of answer such as 'yes', 'no' or 'maybe'.

It was also essential to select or design questions which ensured that every participant would be familiar with the contents, knowing the answer/possible answers to all of the questions and that each question would have different possible answers; thus, creating questions that would yield a full answer but that were not intended to influence the answer itself (Legard et al., 2003, p. 154). This was crucial in spite of the impossibility of anticipating the speakers' ability to express that knowledge in English. This decision was mostly driven by two factors: the first one was the intention to avoid adding insecurity and stress to the already somewhat nervous speakers – modest volunteers – while potentially instilling a sense of confidence. The second factor was to create an opportunity to observe whether there would be cases of calques, neologisms or code-switching. To put it in Schilling's words (2013), it was "designed to steer interviewees away from focusing on specific linguistic forms, toward producing connected speech about topics of interest" (p. 108).

Another central issue was the duration of the interview, which should consider above all, the schedule and availability of the participants. In order to be able to conduct in-person interviews, it was established that the most convenient and efficient way to do so would be to attend a large number of the programmed events where the concentration of members of this community would be the highest and could provide the best opportunity of achieving the greatest number of interviews in the shortest time possible. However, one of the drawbacks of this option was that every person attending the event would be on the same tight agenda, allowing for only a few relatively short openings each day to conduct interviews. Therefore, it was deemed appropriate that the entire duration of each interview should be kept under an hour, 45 minutes ideally, in order to maximize the process while being mindful of the interviewee's limited time availability and willingness to participate.

The interview design process culminated with the development of a standardized introductory statement (to collect metadata like date, location and interviewee's name) plus the seven questions, which were designed to prompt answers such as: definitions of general concepts (yet specific to this sport), description of particular phases of flight, accounts of technical aspects and requirements as well as narratives of specific and memorable moments in their lives as hang glider pilots. This range of answers, content and register was estimated suitable for the purpose of collecting data that would allow the creation of the specialized corpus. These questions were also integrated into a PowerPoint slideshow to be used as a visual aid during the interviews.

One more important element of the process was the use of two PDF forms that each volunteer was required to read, fill out and sign. These forms – consent and personal information – were also designed and created prior to the beginning of the actual interviewing process. The first one, pertinent to the research ethics of this study as suggested by Dresing et al. (2015), explained the purpose of the interview to the participants, informed them how their data would be treated and their privacy respected, and requested contact information in case it would be necessary to get any subsequent permissions at a later stage; it was presented to them before the actual interview. The second form would serve to collect some basic demographic details with the intent of creating a database allowing a deeper look into the community's characteristics. It was presented to them after the end of the interview in order to, as encouraged by Schilling (2013), reduce speaker self-consciousness about factors such as age or socioeconomic status. These additional details could serve as a source of supplementary material to the study.

3.2. Interview Planning and Implementation

For the actual interviews, the tools used were a voice-recording software on a handheld device, a microphone attached to the device and a computer with the two PDF forms and the PowerPoint presentation mentioned above.

In order to launch the interview implementation process, the first necessary step was to review the international competition calendar and begin planning – as much as possible – which events would be viable visiting. A second option for performing interviews would be additional encounters with participants both on the way to or from the different venues as well as during personal and work trips, "taking advantage of unforeseen opportunities as they arise during the course of fieldwork […] using available encounters and events as they arise" (Ritchie et al., 2003, p. 81). This method of selecting the interviewees, together with the steps summarized in Section 2 above, regarding the language groups and range of possible interviews, formulates the main type of sampling used, which would seem to fall under the category of 'Quota Sampling', better outlined by Neumann's words, where:

> we first identify relevant categories among the population we are sampling to capture diversity among units (e.g., male and female; or under age 30, ages 30 to 60, over age 60). Next we determine how many cases to get for each category–this is our "quota". Thus, we fix a number of cases in various categories of the sample at the start (2014, p. 249).

In this case, the categories are the different languages and the cases per each category are those defined in **Table 2**, participant ranges (section 2 above).

A relevant situation directly affecting the initial outline of the research project was the onset of COVID-19, which caused the in-person interviewing process to be abruptly interrupted. As time went by and in the middle of the uncertainty, it became obvious that, in order to resume and finalize the interviewing process, an alternate method was necessary. Therefore, the option of conducting virtual interviews was formulated and implemented and most of the remaining interviews were carried out remotely using two different remote communication tools: Skype and WhatsApp.

The outcome of what can be considered a three-year 'nomadic' interviewing process is recapped in **Table 3**:

Table 3. Interviews: quantity, type and locations (2018-2020)

	2018	2019	2020	Totals
In-Person (iP)	61	42	9	112
Virtual (V)	0	0	27	27
iP + V	61	40	38	139
Lost	0	23	0	23
Locations (iP)	12	7	3	22

There was a total of 140 interviews completed and 116 used for the corpus compilation. **Table 3** also shows that the in-person interviews took place in different locations all over the world, including Asia, Europe, North and South America, as well as Australia.

This concludes the illustration of the interview preparation and execution process; the next section will focus on the data processing phase of the study.

4. Data Processing Methods

This section will focus mainly on the steps undertaken to tackle the actual transcription process of the recordings gathered during the interviewing phase, which represents another fundamental portion of this ongoing project. Transcribing the interviews was a necessary step because it would allow to have a digitalized written version of the spoken data collected, necessary in order to proceed to the actual compilation of the specialized corpus.

4.1. Transcription: Transcriber and Technique

In larger projects with vaster resources, it seems that entrusting transcription to a junior researcher, is the common praxis (Bailey, 2008, p. 129); however, that was not an option in the case of this smaller-scale project. Using a transcriber other than the researcher herself was not considered feasible for yet another reason: the high content of specialized

language as well as the strong accents of some interviewees, would have made it excessively challenging to find an external person suitable for this task.

Another choice to be made was whether to adopt an automated/semi-automated transcription method, or to transcribe manually. Even though specialized software can transform the audio files into written text, a manual review of the transcriptions would nonetheless be imperative in order to verify the accuracy of the resulting text, while complying with methodological guidelines. Once again, this circumstance called for the consideration of the role played by the content of specialized language and the, often, strong accents of the interviewees in the precision of the transcription's outcome.

Despite being aware of the shortcomings in transcription software where "current speech recognition software is not able to accurately convert the spoken files into text files" (Reppen, 2012, p. 34), three tests were carried out, using two different interviews, in order to support the decision regarding the transcription method. These two interviews represented the furthermost ends of the spectrum: one by a speaker (referred to as S74) without a strong accent and an advanced level of the English while the other speaker (referred to as S36) had a heavy accent and displayed more difficulty in the use of the English language.

The first test was to have the computer transcribe the interviews. **Figure 4** below displays a screenshot of the outcome (excerpt) where it is possible to see that the transcribed text was rendered as a single paragraph without any distinction to the speaker turns. This created an issue right from the start because the text would have to be manually formatted in order to separate the turns so as to be able to clearly identify what is being said by the interviewer and by the interviewee. Other issues illustrated by **Figure 4**, are examples where the software was unable to recognize when the speaker used non-English speech (highlighted in yellow), it was unable to decode the speaker's accent (highlighted in blue) or the use of technical vocabulary (highlighted in green) as well as false starts highlighted (in grey). Moreover, this automated transcription yielded fragments which were not

only completely off, but made absolutely no sense, such as (line 8): […] when you sing the Chinese group you tried to run Hyundai on lighting […]. Also, the software skipped quite a few utterances, which is not directly illustrated by this figure. All this means that, even after a thorough review of this type of automated transcription, it is quite likely that the level of fidelity would have been drastically diminished, not to mention how much more laborious the process would have been.

Moreover, the omission and/or distortion of several foreign and specialized words and phrases resulting from the automated transcription, would result in the researcher's impossibility to detect, for example, instances of code-switching and self-correction, among other occurrences, which become especially significant during the data analysis phase.

Today is August 26th I'm in Lorain with tonsillitis and we're going to begin our interview hello please answer the following 6 questions what is a thermal how much is a is a win-win-win place because the ground is cold and we have what's ability to to go in his direction to clean sweep next question how does a cross country competition tasks work Lieutenant between the fighter who has good to go some Tito Mesa Envy Hair and they have to try to do this Mufasa next question describe how you take off on a hang glider okay today, so you need a good swing the witness to come before you and to take off is there a lot of attention the choice the time and when you sing the Chinese group you tried to run Hyundai on lighting and supply okay next question what is the necessary equipment to practice hang glide spells to practice and Lightning McQueen and I'm driving and after you need the box when you can leave under the glider the guy that was your body when you hope to stay in your hair it's better that you have a volume sometimes the ride you must be crazy and I did and your your good luck to take off at night what are the parts of a hang glider and how does it fly which process do the living legend I'm the sweetest you take me Christian you are so did you catch the same and the you are the prison under the same and Trinity the boss. I kind of stood next question about your first flight it's supposed to be smaller and smaller one minute you just take off straight down display and Landing I wasn't because my legs and the last question tell me about your best life and why was it the best one they have two children Lady Justice lie in the month of May in shamoni this was not so nice and we are all the moves on land but it was I was young and it's because I saved this picture in my head I said I had a lot

Figure 4. Example of the result of the automated transcription by Google (highlighted by the author)

To further portray the situation, here are some examples of segments from the automated transcriptions (AT) vs the initial draft of the manual transcription (MT). These examples will further support the superiority of

using MT for this project (misrepresentations in the automated version are in bold and underlined text):

> (5) S36 – Answer to question #2:
>
>> **MT**: OK a task in competition is uhm a challenge between uh pilots who has to go some kilometer in the air and they have to try to do this kilometer the more fast- faster than than they could [laughs]
>>
>> **AT**: **Lieutenant** between the **fighter** who has **good** to go some **Tito Mesa Envy Hair** and they have to try to do this **Mufasa** (*note: the rest of the answer was totally omitted by the software*)
>
> (6) S74 answer to question #7 (partial):
>
>> **MT**: [...] in uh October November around I was towed by a ultralight ... and it was a partly couldly day but the pilot of the ultralight was that cool that he towed me in between the clouds [...]
>>
>> **AT**: [...] in October November around I was **told** by a **tree light** and it was ∅ partly cloudy day but the pilot of the ultralight was **Deadpool** and he **told** me in between the clouds [...]

Additionally, the chart in **Figure 5** below provides supplemental information regarding the disparity in transcription techniques, as it presents a summary of the different results obtained by using the automated transcribing tool Google Transcribe (referred to as AT) versus the manual transcription (referred to as MT) of the two interviews chosen for this trial:
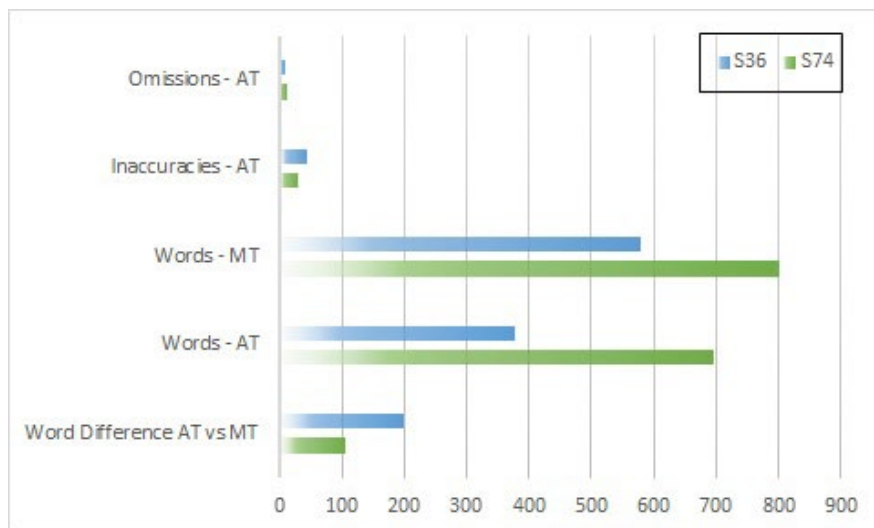
Figure 5. Test of Manual (MT) vs Automated (AT) Transcription Methods

It is worth mentioning that, besides the number of omissions and inaccuracies, there was a significant difference in the final word count where the automated system – in both occasions – rendered fewer words than the manual transcription: 34.6% fewer utterances in interview S36 and 13.2% in S74.

Besides these two tests, one last attempt was performed, using interview #36, in order to see if removing the variable represented by the foreign accent would change the outcome. This time, the interviewer (NS) read the text originally spoken by S36 out loud for the Google dictation (GD) software to transcribe. Once again, there were several words/phrases that were transcribed inaccurately and parts of the dialogue that were omitted. **Table 4** below shows some of the differences, giving us an idea of the discrepancies and the implicit extra editing work this kind of transcription would require:

Table 4. Examples from Google dictation software and manual transcription – interview S36

GD	MT
at asking	a task
Goodwins	good wind
you have to do chores	you have to choice
10 guiding	hang gliding
Valium 80	variomètre

These tests provided ample evidence that, despite the apparent ease of having the computer convert the spoken interview into written text, a substantial amount of extra work would be necessary, in order to amend the omissions and inaccuracies which would also impoverish the quality of the transcriptions.

Despite the fact that the prospect of a manual transcription also implied a large amount of work, there would be much less effort related to correcting words or 'making things right' and this provided a sense of relief, of a more agreeable work flow than constantly searching for misinterpretations and omissions that would not contribute anything positive to the process. It became clear that even if it has been estimated that the task of transcription can be between 4 to 10 times the length of the recorded audio text, meaning 4-10 hours to transcribe 1 hour of speech (Nagy and Sharma, 2013, p. 251), a manual transcription was the appropriate option. This decision was supported by embracing the idea that, the much lengthier MT, also provided the possibility of gradually getting acquainted with the content and gaining in-depth knowledge of the texts, which would provide an interesting insight and, most likely, be useful during future stages of the study while "bearing in mind that linguists are never as close to their object of study as when they are transcribing" (Nagy and Sharma, 2013, p. 237).

4.2. Transcription: Tools and Method

This section presents the details of the actual transcription process. The essential tools used were a laptop computer, a transcription software, MS Word and headphones. After looking into several specialized software alternatives such as Transcribe[3] and Dragon Speech Recognition[4], the final decision was to use InqScribe[5] to carry out this task. Some of InqScribe's most appealing features included the following:

- Supports a wide variety of voice-only as well as video formats
- User friendly
- Adjustable playback speed and volume
- Timestamps
- Snippets[6]
- Keyboard shortcuts for practical functions like pause, play, skip back, snippets, timestamps, etc.
- Customizable shortcuts
- Free version available

The general process of each interview transcription is comprised of different stages. The first stage using InqScribe, was a complete and thorough transcription with the aim to include as many elements as possible. For instance, besides full words, phrases and sentences, it also included: pauses, repetitions, false starts, background information, contextual events and/or speaker noises. This allowed us to collect as much raw information as possible from the very beginning and then select the level of detail at a later stage. The complete transcription also provided the

---

[3] https://transcribe.wreally.com

[4] https://www.nuance.com/dragon.html

[5] https://www.inqscribe.com

[6] "[s]hort bits of text that you find yourself typing into your transcript over and over again" (Inqscribe)

opportunity of having this material available for future complementary studies or analysis.

Once certain features of the transcription software had been customized – for instance, the skip back times and snippets to indicate speaker turns ("de la Pava" for the interviewer and "HG Person" for the interviewee) – the first step was to drop and drag the recording into InqScribe. Next, the recording playback speed and volume were selected; most of the times, volume was set to the highest position and playback speed was set between the 0.7x-0.9x play rate; however, it was often necessary to bring this rate down to 0.2x-0.4x in order to get a better opportunity of either correctly transcribing the utterances or declaring them as unintelligible. This feature of the transcription software proved extremely useful for the process. The next step was the actual transcribing and the following screenshot (**Figure 6**) illustrates what the outcome looked like; it is possible to see the layout, the timestamps, turns and settings:
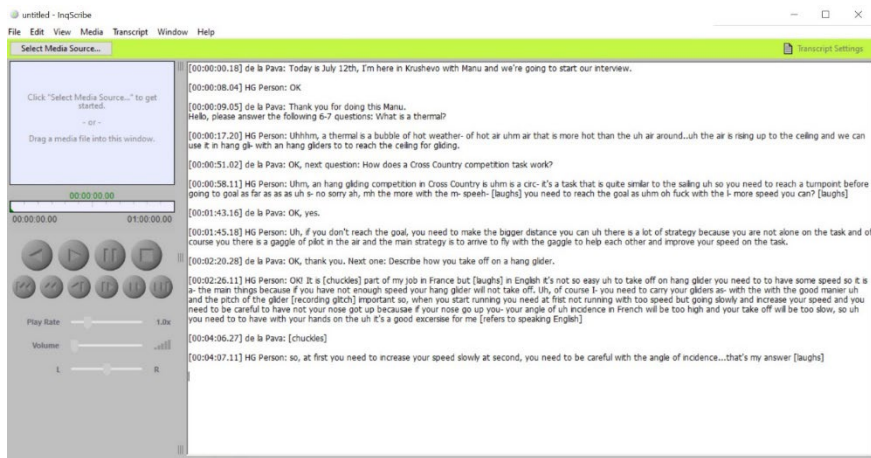

Figure 6. Screenshot that exemplifies the interview transcription process with InqScribe

As soon as each initial draft of the transcribed interviews was completed, the text would be copied and pasted into a separate MS Word document, then saved in a folder titled "Raw Data". This designation was selected to

identify that all the documents in this file contained everything that could be heard on the recording. Moreover, each file name included a unique label assigned to each interview/interviewee: the letter 'S' for 'speaker' and a number allocated to each interview as it was recorded. Most of these numbers were assigned in chronological order (1-116) so that, as the recordings were made, the next number on the list would be assigned to it. The number 0 represents the pilot interview and test run for each of the different steps (e.g., recording, transcription, consent form, personal data, etc.). This type of classification seemed adequate because, on the one hand, it allows us to keep track of all the different documents across the entire process since any record pertaining to each one of the participants includes their uniquely assigned number and, on the other hand, this system facilitates safeguarding the privacy of each interview and the interviewee.

In 2021, after all the initial transcriptions were concluded, a second transcription phase took place, referred to as the "clean up and standardization" phase. For the "clean up" part of this second stage, every document was reviewed, typos were corrected, the questions asked by the interviewer were removed as well as the timestamps and the speaker turn labels, since these would not be relevant this time for the corpus analysis phase. Nevertheless, this information was kept in the original folder, to make it available in case it would be necessary to go back and look for specific parts of the text. For the "standardization" part, the first thing done was to add relevant metadata at the beginning of each file, such as: Speaker ID (Sn), Native Language (L1), Type of Interview (in person / virtual), Place(s) and Overall Recording Duration. Additionally, the beginning and end of each text was marked by the <Sn> </Sn> tags. Then, the transcription guide – created specifically for this study (see below) – was implemented.

Having a transcription guide applied to every document in this study was an important component of the process and the methodology because, not only, would it create a homogeneous content throughout the 116+ transcriptions, but all the transcribed interviews would match and portray

information that would be readily comparable. This would ensure that a high methodological standard was maintained, which has been one of the ultimate goals of the data collection and processing phases of this project. The guide created was based on the knowledge of the material acquired during the initial transcription cycle and on the transcription guides of three different, well renowned international corpus projects: the English as a Lingua Franca in Academic Settings (**ELFA**[7] – University of Helsinki); the Michigan Corpus of Academic Spoken English (**MiCASE**[8] – Michigan University) and the Vienna-Oxford International Corpus of English (**VOICE**[9] – University of Vienna) corpora. Comparing the options found in these guides and the content of the interviews, a number of parameters and conventions were selected, since not every entry in the ELFA, MiCASE or VOICE guides was relevant to the texts under study here. The decision to use several guidelines as a point of departure enabled a wider range of options and possibilities to choose from, allowing the use of the most suitable parameters and conventions for this specialized corpus of non-native and non-academic English. To give an idea of the content of our guide, here are some excerpts of the conventions guide:

(6) <u>Capitalization</u>: Only proper nouns such as names, countries and regions: departments, states, countries, mountain ranges, etc. Brands (Moyes, Wills Wing, etc.) will not be capitalized (see Abbreviations and Acronyms)

(7) <u>Initialisms</u>: All capital letters (GPS, FAI, etc.), except PhD (lower case h)

(8) <u>Numbers</u>: Fully spelled out as words with standard hyphenation

(9) <u>Unintelligible Speech</u>: Represented by 'xx' in between markers: <UN>xx</UN>

(10) <u>Repairs</u>: All transcribed

---

(11) <u>Spoken Mistake / Coinages</u>: Represented by the word between markers: <SIC>text</SIC>. In order to avoid this kind of text to be interpreted as a typing mistake

(12) <u>Contractions / Lexicalization of Reduced Forms</u>: Spelled out (a list is provided)

The next and final phase of the transcription journey included a third and comprehensive review. It was performed by listening to all of the recordings while reading the transcribed texts simultaneously, looking for possible typos and any other inaccuracies that could have slipped through the previous checks. Taking into account that time had gone by between the first transcriptions and the time of this final phase and also understanding that "the same interview can be transcribed in different ways […] even by the same person at different times" (Azevedo et al., 2017, p. 160), made it clear that this final revision should definitely not be skipped. Additionally, two colleagues – one from the academic world and another one from the hang gliding community – were asked to randomly verify some of the transcriptions. They listened to those interviews comparing them to the final draft of the texts, in order to help corroborate their accuracy. One of them reviewed 10 files, and the other one, 15 for a total of 25 revised files out of 116, which accounts for 21.6% of the transcriptions. This step was considered a judicious complement to the methodology and the quality of the verification process followed so far. The very last modification to the files was removing the metadata at the beginning of each one in order to generate the definitive document, saved both in MS Word and Plain Text versions, which would be used to create the corpus itself.

This closes the section describing the procedure followed to achieve the transcriptions later compiled in the Hang Gliding Corpus.

5. Data Analysis and Preliminary Findings

The specialized corpus was finalized and compiled during the summer of 2022 and the data analysis phase is just at its beginning, which accounts for

the decision of including information regarding the initial steps of the analysis itself as well as some preliminary findings.

5.1. Corpus Tools and Analysis

Having recently completed the compilation of the International Hang Gliding Corpus of Spoken English by Non-Native Speakers, developed for the purpose of this study, the next steps will be the in-depth analysis of the corpus, using mainly Sketch Engine. The decision to use Sketch Engine as the primary tool, rather than other tools such as AntConc or WordSmith, is due to two main reasons: First of all, in our opinion, Sketch Engine appeared to be more versatile, offering a wide range of analysis options and, second of all, being able to take advantage of Sketch Engine's institutional login option.

After taking a look to some of the different tools for corpus linguistics, Sketch Engine seemed more suitable for this study because it offers different, user-friendly and graphically attractive ways to search and analyze a corpus; different ways to access the information as well as a number of tutorials (links provided on their landing page) and tips found in each section under the 'About' tab. Also, web-based Sketch Engine allows you to either select a corpus from a list of public corpora or to use your own corpus by uploading your documents and having Sketch Engine compile your specific corpus. This second option is a relatively simple process (once you have the texts selected and formatted, of course) and is exactly what has been done for this project. Nevertheless, we are aware that AntConc and WordSmith also offer numerous possibilities for corpus analysis; therefore, they may be used as complementary tools at a later stage.

5.2. Preliminary Results

Despite being on the early stages of the corpus analysis itself, it was deemed relevant to present some examples of preliminary findings. While providing examples of preliminary findings, it is our aim to pass on

information regarding the experience as a first-time user of Sketch Engine, hoping some readers may find this information useful for future research.

5.2.1. Examples of Specialized Vocabulary

Several examples of specialized language have been detected so far, which enable us not only to delve deeper into the specific struggles of NNS with this kind of terms but also to recognize the potential applications of our findings to the area of translation and teaching of ESP.

Table 5 below shows the results of a comparative search – using Sketch Engine (SE) – between the Hang Gliding Corpus (HGC) with 112, 747 tokens and three of the corpora available in the SE database: British Academic Spoken English Corpus (BASE) with 1,756,545 tokens, the Brown Corpus – group of corpora (BROWN) with 8,099,732 tokens and the Open American National Corpus – Spoken (OANC) with 3,369,613 tokens. This search was conducted to help support the definition of the HGC as a 'specialized' corpus. The words used in this basic search belong to specialized fields, which include hang gliding and aviation, yet the difference in frequency found is worth mentioning. A total of eight words were used for this initial search; **Table 5** presents comparative results of the first four terms and **Table 6** present the results of the other four words, which were found only in the HGC and not in any of the other corpora used as reference. **Tables 5** and **6**, display the overall frequency of the lemmas, the distribution of the frequency per million of tokens and the percentage of the frequency within the given corpora.

Table 5. Examples of Specialized Vocabulary – comparison between different corpora

BASE CORPUS	Batten	Gaggle	Keel	Thermal	OANC CORPUS	Batten	Gaggle	Keel	Thermal
Number of hits	--	--	1	48	Number of hits	--	--	3	1
Number of hits x million tokens	--	--	0.57	27.33	Number of hits x million tokens	--	--	0.89	0.3
Percent of whole corpus	--	--	0.00005693%	0.002733%	Percent of whole corpus	--	--	0.00008903%	0.00002968%
BROWN CORPUS					HG CORPUS				
Number of hits	26	8	40	129	Number of hits	37	9	33	323
Number of hits x million tokens	3.21	0.99	4.94	15.93	Number of hits x million tokens	328.17	79.82	292.69	2,864.82
Percent of whole corpus	0.0003210%	0.00009877%	0.0004938%	0.001593%	Percent of whole corpus	0.03282%	0.007982%	0.02927%	0.2865%

It is worth noting the difference in the relative data, where you may see a similar number of hits within two corpora, yet when looking at the percentage values, there is a significant difference. For instance, on **Table 5**, we can observe that the lemma *Gaggle* appears 8 times in the BROWN corpus and 9 times in the HG corpus, which could appear as a similar result; however, in the former, the 8 hits represent a 0.00009877% of the whole corpus but in the latter, the 9 hits represent a 0.007982% of the whole corpus. Moreover, the concordances within the BROWN corpus are completely unrelated to the fields of aviation, recreational flying or hang gliding.

Table 6 below, shows that while none of these lemmas, belonging to the specialized language found in the HGC, are present in any of the other three corpora.

Table 6. Examples Present in the Hang Gliding Corpus – not in the other corpora

HG CORPUS	Aerotowing	Speedbar	Technora	Turnpoint
Number of hits	13	28	6	81
Number of hits per million tokens	115.3	248.34	53.22	718.42
Percent of whole corpus	0.01153%	0.02483%	0.005322%	0.07184%

The information presented above not only supports the definition of the HGC as a specialized corpus, but it also helps identify a possible need to expand current databases in general to include ESP vocabulary that is not yet accounted for or the need to have more specialized corpora. This complement could lead to an improvement of translation and teaching especially in the field of ESP.

Tables 5 and **6** have been compiled by the author, merging several individual tables automatically generated by Sketch Engine (see screenshot below). This is another example of how versatile and interactive this corpus manager tool is, which makes it highly recommended for fellow researchers looking into carrying out similar research tasks.

Figure 10. Sketch Engine Concordance Search Results - Screenshot

Figure 10 illustrates the pop-up window that appears when you select the information icon (top left side of image, marked with a green arrow by the author) on the results page. Not only is there a table that summarizes the results (used to create **tables 5** and **6** above), but it also automatically creates a graph.

5.2.2. Instances of Technological / Lexical Deviations

Several expressions have been recognized as being 'lexical deviations' or 'inaccuracies'; these two terms are used to avoid stigmatizing such instances because this type of material seems to present an opportunity to learn more about them and potentially use this information, not only to gain deeper knowledge of the NNS specialized variety of English used by the hang gliding community, but also to develop learning exercises or find another way to apply the knowledge gained to the fields of translation and teaching / learning. It is most likely worthy conducting a complementary search to see if it is possible to find more of these instances. This is a partial list of words / phrases representing this category:

    (13)   *Airborned* ('airborne' – without conjugation)

(14) *Cumus* (instead of 'cumulus' – a type of cloud)

(15) *Go-and-back* (the expression regularly used is 'out-and-return' – a kind of distance flight)

(16) *Sustentation* (instead of 'lift' – possible calque of 'sustentação' in Portuguese)

(17) *Runned* (instead of the past tense form 'ran')

(18) *Powered-engined (conjugation of adjectives)*

(19) […] *we climb the mountain with a car* (use of the verb 'to climb')

(20) *Incidence* (possible calque of 'incidence' in French)

These expressions seem to show probable indicators of how the speakers' conjugate verbs (13 and 17); why they choose certain verbs (19); how they create expressions (15, 18) and so forth. This kind of information could be used to develop teaching exercises.

5.2.3. The Use of Articles a/an/ø

Acquiring, or fully understanding, the use of the "English article system" has been widely studied and deemed as one of the most challenging aspects of grammar for learners of English as a foreign language, as recapped by Sun (2016, p. 2). This is also underlined by Sarmento in his study of specialized aviation English, when citing Trimble's account of learner's challenges, he says that the "non-standard use (and absence) of the definite article in descriptions and instructions[10]" (2010, p. 203). In the Hang Gliding Corpus, there seems to be several examples of article use that could be examined as an attempt to gain a better understanding of semi-specialized and specialized English and, maybe, find ways to aid students in their process. Some examples are:

---

[10] "uso não-padrão (e a ausência) de artigo defnido em descrições e instruções" The reverse-translation in the text was performed by the author.

(21)   Omission of the article: *[…] ø kind of explanation […]*

(22)   Use of "a one" instead of "an": […] *there is a one aerodynamic set* […]

(23)   Use of "an" instead of "a": […] *you need an helmet* […]

(24)   Use of "a" rather than "an": […] *I have a obvious idea* […]

(25)   use of "a" / "an" rather than ø: […] *with a foggy glasses* […]

These instances exemplify an alternative use of the articles and could prompt the researcher to look into the interviewees' native language (L1) to try and trace back any influence as in example (22), where the speaker elicits the word "one", which might be the reason behind their choice for "a" instead of "an".

This concludes the presentation of examples of preliminary findings aimed to serve as a preview of some of the analysis possibilities of this study.

Conclusion

The examples presented in the preceding three sections, represent the beginning of a deeper and more thorough search for phenomena within this corpus in search of significant information regarding the use of English within this community, their intercultural communication dynamics and maybe the possibility of identifying a sociolect. This kind of data will provide more knowledge and, hopefully, a better understanding of how non-native speakers naturally produce English speech in specialized contexts of usage. So far, it is possible to say that these preliminary findings help us take into consideration matters such as: a) an opportunity of broadening the databases of specialized vocabulary in the fields of aerial sports and aviation enhancing ESP both in translation and teaching (section 5.2.1); b) gaining insight on the nature of the lexical deviations found in this corpus, which could help lead to an improvement in intercultural communication and mediation (section 5.2.2) and c) taking a closer look at the use of the article 'a/an/∅' within this corpus (section 5.2.3), might

facilitate discerning examples and possible causes of the struggle some NNSs encounter with these articles, which despite not always yielding to miscommunication, may affect their linguistic flow and overall speaker confidence.

This corpus-driven study is being conducted mainly within the specialized field of hang gliding and the results obtained could serve not only to enrich this particular community, but they could also help develop the fields of ESP and specialized translation teaching since this subject has become "a major field of investigation in applied linguistics" (Kübler et al., 2018, p. 810); additionally, it could also lead to enhancements of current specialized translation databases and maybe even make it possible to attempt and aid improve intercultural communication and teaching tools for English as a foreign as well.

Collecting and processing the material for this study has taken about four years and it may seem relevant to state here that embarking on such a process, yielding to the creation of a specialized corpus, has added a very interesting component to the development of this project; nevertheless, as Evans states (2018, p. 4), it "is usually not the kind of project that can be undertaken without a lot of planning and execution time." Fortunately, the challenges presented by this endeavor can be said to be undoubtfully overcome by the rewards, the advantages and added motivation that derive from it. Besides presenting the project, this paper's purpose is also to provide information and some insight into this kind of tasks – specifically conducting field interviews and performing manual transcriptions – that could be useful for other fellow researchers who might consider a similar path.

This paper has covered the introduction of the community under study, it illustrated the data collecting and processing methods put in place, to conclude with the presentation of some of the preliminary results observed so far and a brief description of the following steps.

References

Azevedo, V. et al. (2017). Interview Transcription: conceptual issues, practical guidelines, and challenges. *Revista de Enfermagem Referência* 4(14), 159-168. https://doi.org/10.12707/RIV17018

Bailey, J. (2008). First steps in qualitative data analysis: Transcribing. *Family Practice* 25(2), 127-131. https://doi.org/10.1093/fampra/cmn003

Bernardini, S. (2003). Designing a Corpus for Translation and Language Teaching: The CEXI Experience. *TESOL Quarterly* 37(3), 528-537. https://doi.org/10.2307/3588403.

CIVL – Commission Internationale de Vol Libre. *2016 CIVL Competition Coordinator Report*. Last accessed on: 17 March 2017. Retrieved from: http://www.fai.org /downloads/civl/CIVL_2017_Agenda_Ax20 .

Dresig, T., Pehl, T. and Schmieder, T. (2015). *Manual (on) Transcription. Transcription Conventions, Software Guides and Practical Hints for Qualitative Researchers.* 3rd English Edition. Marburg. Retrieved from: http://www.audiotranskription.de/english /transcription-practicalguide.htm.

Dressen-Hammouda, D. (2012). Ethnographic Approaches to ESP Research. In: B. Paltridge and S. Starfield (Eds) *The Handbook of English for Specific Purposes.* (pp. 501-518). Wiley-Blackwell Publishers.

ELFA. (2004). *ELFA Corpus Transcription Guide*. Retrieved from: https://www2. helsinki.fi/sites/default/files/atoms/files/elfa_transcription_guide.pdf.

Evans, D. (2018). *Corpus Building and Investigation for the Humanities: An on-line information pack about corpus investigation techniques for the Humanities*. Retrieved from: https://www.birmingham.ac.uk/documents/college-artslaw/corpus/intro/ unit2.pdf.

FAI – Fédération Aéronautique Internationale. Sporting Code Section 7. (2018). Last accessed on: 4 October 2018. Retrieved from: https://www.fai.org/sites/default/files /civl/documents/sporting_code_s7-common_2018.pdf.

Genzuk, M. (2003). A synthesis of ethnographic research. *Occasional Papers Series.* Center for Multilingual, Multicultural Research (Eds). Center for Multilingual, Multicultural Research, Rossier School of Education, University of Southern California. (pp. 1-10). Retrieved from: https://web-app.usc.edu/web/rossier/publications/ 33/Ethnographic_Research.pdf.

Gobo, G. (2004). Sampling, Representativeness and Generalizability. In: Seale, C. et al. (Eds). *Qualitative Research Practice* (pp. 405-426)*. SAGE.

Gobo, G. (2011). Ethnography. In: D. Silverman. *Qualitative Research.* 3rd Edition (pp. 15-36). SAGE, https://www.researchgate.net/publication/279373417_Ethnography.

Hammersley, M. (2007). Reflections on Linguistic Ethnography. *Journal of Sociolinguistics* 11 (5), 689 - 695.

InqScribe 2.5 Documentation. Last accessed on: 25 July 2022. Retrieved from: https://docs.inqscribe.com/2.2/snippets.html .

Kübler, N., Mestivier, A. and Pecman, M. (2018). Teaching Specialised Translation Through Corpus Linguistics: Translation Quality Assessment and Methodology Evaluation and Enhancement by Experimental Approach. *Meta*, 63 (3), 807-825. https://doi.org/10.7202/1060174ar

Legard, R., Keegan, J. and Ward, K. (2003). In-depth Interviews. In: J. Ritchie and J. Lewis (Eds). *Qualitative Research Practice. A Guide for Social Science Students and Researchers* (pp. 138-169). SAGE.

Mauranen, A. (2003). The Corpus of English as Lingua Franca in Academic Settings. *TESOL Quarterly* 37(3), 513-527.

MICASE Manual. (2003). The Michigan Corpus of Academic Spoken English. Prepared by R. C. Simpson, D. Y.W. Lee and S. Leicher. Retrieved from: https://ca.talkbank.org /access/0docs/MICASE.pdf .

Nagy, N. and Sharma, D. (2013). Transcription. In: Podesva, R. J. and Sharma D. (Eds). *Research Methods in* Linguistics (pp. 235-256). Cambridge University Press.

Neuman, W. L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches*. Pearson Education Ltd.

Reppen, R. (2012). Building a Corpus. What are the key considerations? In: A. O'Keeffe and M. McCarthy (Eds). *The Routledge Handbook of Corpus Linguistics* (pp. 31-38). Routledge.

Ritchie, J., Lewis, J. and Elam, G. (2003). Designing and Selecting Samples. In: J. Ritchie and J. Lewis (Eds) *Qualitative Research Practice. A Guide for Social Science Students and Researchers* (pp. 77-108). SAGE.

Sarmento, S. (2010). O verbo modal *CAN* em manuais de aviação em inglês: implicações para materiais didáticos. In: C. Lopes Perna, H. Koch Delgado and M. J. Finatto (Orgs). *Linguagens especializadas em corpora: Modos de dizer e interfaces de pesquisa*. (pp. 202 - 232) EDIPUCRS. Retrieved from: https://editora.pucrs.br/edipucrs/acessolivre// livros/linguagensespecializadasemcorpora.pdf.

Schilling, N. (2013). Surveys and Interviews. In: R. J. Podesva and D. Sharma (Eds). *Research Methods in* Linguistics (pp. 96-115). Cambridge University Press.

Sun, G. (2016). The Acquisition of English Articles by Second Language Learners: The Sequence, Differences, and Difficulties. *SAGE Open*, *6*(1). https://doi.org/10.1177/ 2158244016635716

VOICE Project. (2007). *Transcription Conventions* [2.1]. Retrieved from: https:// www.univie.ac.at/voice/ .

Biographical Remark

Claudia Mejia De la Pava is a PhD Candidate at the University of Salamanca in Spain. She earned her Master of Arts, *summa cum laude*, from *the Università degli Studi di Milano*, in Italy. Born in New York City, she is a hyperpolyglot who has lived, worked and studied in multilingual and multicultural environments, for more than 25 years. Her research focuses on intercultural communication and mediation as well as ELF emphasizing on ESP. Since 1997, besides being an English teacher, she has served as translator and cultural mediator in numerous international sport events in the field of recreational aviation and, between 2009 and 2018, she worked for the *Fédération Aéronautique Internationale*.

Nota biográfica

Claudia Mejia de la Pava es una estudiante de doctorado en la Universidad de Salamanca en España. Obtuvo su Maestría en Humanidades, *summa cum laude*, de la *Università degli Studi di Milano*, en Italia. Nacida en la ciudad de Nueva York, es una hiperpolíglota que ha vivido, trabajado y estudiado en ambientes multilingües y multiculturales por más de 25 años. Sus investigaciones están enfocadas en la comunicación y mediación intercultural, así como en Inglés como Lengua Franca (ELF - siglas en inglés) con énfasis en Inglés para Fines Específicos (ESP - sigla en inglés). Desde 1997, además de ser profesora de inglés, ha actuado como traductora y mediadora cultural en numerosos eventos deportivos en el ámbito de la aviación recreativa y, entre el 2009 y el 2018, trabajó para la Federación Aeronáutica Internacional.