2

DESIGNING AND CONSTRUCTING A CORPUS OF SPOKEN SPANISH AS A FOREIGN LANGUAGE: THE PRACOMUL CORPUS OF STUDENT TRANSCRIPTIONS

Damjan Popič, Univerza v Ljubljani

This chapter outlines the design, development, and outcomes of the PRACOMUL platform, a dedicated website and hub created as part of the PRACOMUL project (Pragmatic Competence from a Multilingual Perspective). The PRACOMUL platform connects learners from diverse linguistic backgrounds, enabling them to interact with students from various universities in different countries. Its primary focus is on facilitating the acquisition and development of pragmatic competence in Spanish as a second language.

Keywords: *spoken corpora, Spanish as a foreign language, transcription, corpus linguistics*

1. Introduction

Spoken corpora, collections of transcribed spoken language, have become indispensable tools for a wide range of linguistic research, including phonology, syntax, semantics, pragmatics, and discourse analysis (Adolphs and Carter, 2013). They provide a rich source of data on the

natural use of language, allowing researchers to study how language is used in real-world contexts.

In linguistic research, the study of how (non-native) speakers acquire and use a second language offers invaluable insights into the processes of language learning, adaptation, and variation (Baker, 2010). This article introduces a study centred on a corpus of spoken Spanish, uniquely compiled from dialogues and recordings of non-native speakers — specifically, students learning Spanish. This demographic, often overlooked in traditional linguistic corpora, provides a fresh perspective on the dynamics of Spanish language use beyond its native contexts.

The PRACOMUL corpus was constructed for this end, though with a special emphasis on a particular setting. The corpus thus consists of dialogues recorded by students who are randomly paired to record the dialogues and then transcribe them, thus constructing the corpus in the process. The corpus, purposefully collected from students of diverse nationalities and backgrounds, is enriched with metadata including their university affiliation, nationality, (perceived) proficiency level in Spanish, and age. This rich dataset serves as the foundation for a novel exploration into how non-native speakers navigate the complexities of Spanish, particularly in their use of discourse markers — a key component in understanding fluency, pragmatics and cultural adaptation in language use. This language resource, however, lends itself to several potential research areas, as well as various pedagogical purposes for teaching Spanish at any level. In addition, since the resource is modular in nature and built to allow further modification and adding data to the corpus.

The significance of this study is manifold. Firstly, it addresses a critical gap in linguistic research by focusing on learner Spanish, an area that has not been extensively explored in existing corpora. Secondly, by examining the speech patterns of non-native speakers, the study sheds light on the processes of second language acquisition, highlighting the variations in language use as influenced by the learners' linguistic backgrounds, proficiency levels, and cultural contexts.

Furthermore, this research has practical implications for language education, offering valuable insights that can enhance teaching methodologies and curriculum development, particularly in areas relating to conversational skills and intercultural communication. The study also holds

relevance for the field of computational linguistics, especially in refining speech recognition and translation algorithms to better accommodate non-native speech patterns.

Methodologically, this study represents a meticulous and ethically conscious effort in linguistic data collection. The process involved obtaining the consent and protecting privacy of the participants, while rigorously capturing and annotating their spoken Spanish. The metadata accompanying each transcription adds a layer of depth to the analysis, allowing for a multifaceted exploration of how various factors (potentially) influence language use.

The article is structured to provide a comprehensive overview of the project and its results. It begins with a background on spoken language corpora, with a specific focus on non-native speakers of Spanish. Following this, a detailed account of the methodology employed in constructing this unique corpus is provided. Finally, the article concludes by discussing the implications of these findings for linguistic theory, language education, and technology, and suggests directions for future research in this field of study.

In essence, this article offers a novel contribution to the study of Spanish as a second language, highlighting the diverse ways in which non-native speakers engage with and adapt to the language. It stands as a testament to the evolving nature of language learning and the rich diversity found within the world of language acquisition.

2. Background

Spoken language corpora play a pivotal role in contemporary linguistic research, providing a wealth of data that is vital for understanding language as it is naturally used. These corpora, which consist of systematically collected and organized recordings of speech, offer invaluable insights into the dynamics of spoken language (Biber, 2020), also in the academic environment (Biber, 2006). The importance of spoken language corpora can be highlighted in several key areas:

1) Real-life language usage: Unlike written texts, spoken language corpora capture language as it is used in everyday conversation. This includes

colloquial language, slang, dialectical variations, and spontaneous speech patterns, offering a more authentic picture of language use.

2) Prosodic features analysis: Spoken corpora allow researchers to study prosodic features such as intonation, stress, rhythm, and pitch, which are crucial in understanding spoken language's nuances. These aspects are often key in distinguishing regional accents, emotional states, and communicative intentions.
3) Sociolinguistic insights: Spoken language corpora are essential tools in sociolinguistics, as they facilitate the study of language variation and change between different social groups and regions. They help in understanding how factors like age, gender, socio-economic status and geographical location influence language use.
4) Pragmatic research: They provide a rich resource for studying pragmatics—the study of language in use and of the contexts in which it is used. This includes how people use language to achieve specific effects, such as persuading, apologizing or requesting.
5) Speech technology and computational linguistics: In the realm of technology, spoken corpora are fundamental in developing and refining speech recognition and synthesis systems. These systems require large amounts of data to accurately mimic and respond to human speech patterns.
6) Language teaching and acquisition: For language educators and learners, spoken corpora provide a resource for understanding how native speakers use language naturally. This is particularly valuable in teaching pronunciation, conversational skills, and listening comprehension.
7) Preservation of linguistic diversity: In a world where languages are rapidly disappearing, spoken language corpora can play a role in preserving linguistic diversity, providing a record of languages and dialects that may be at risk of extinction.

In summary, spoken language corpora are indispensable in linguistics, offering a lens into the living, breathing entity of language as it is spoken. They bridge the gap between theoretical language studies and real-world language use, providing a more comprehensive understanding of the complexities and intricacies of human communication.

2.1 A Brief Overview of Spanish (Spoken) Corpora

In Spanish linguistics, the study of language through corpora has traditionally been associated with the Royal Spanish Academy (RAE)[1] and, more recently, the Association of Academies of the Spanish Language (ASALE)[2] (Sánchez-Gutiérrez, De Cock and Tracy-Ventura, 2022). The collections curated by these entities are broad in scope and organized chronologically, with each corpus covering distinct historical periods. The Diachronic Corpus of Spanish (CORDE, by the Royal Spanish Academy) boasts over 250 million tokens, charting the Spanish language's development from its earliest texts up to 1974. The Reference Corpus of Current Spanish (CREA, by the Royal Spanish Academy) comprises 160 million words sourced from diverse regions and nations, spanning the years 1975 to 2004. Additionally, the 21st Century Spanish Corpus (CORPES, by the Royal Spanish Academy) is designed to capture the wide array of Spanish usage in the 21st century, containing more than 300 million tokens. Despite CORPES' significant effort to include a diverse array of Spanish varieties, with approximately 70% of its data originating from countries outside of Spain, the RAE has faced critique for disproportionately highlighting Spanish variants from Spain (Sánchez-Gutiérrez, De Cock and Tracy-Ventura, 2022), representativeness as such being a lasting issue (Moreno-Fernández, 2005). Against this backdrop, the Spanish Corpus, Web/Dialects (Davies, 2016) emerged as an alternative comprehensive corpus, amassing two billion words from websites across 21 Spanish-speaking countries[3].

In terms of specialized corpora, several have been developed to explore Spanish as it is used in particular geographic regions, ranging from broad areas such as Latin America in the Diachronic and Diatopic Corpus of Spanish in America (*Corpus Diacrónico y Diatópico del Español de América*; CORDIAM; Mexican Academy of Language)[4] to more specific locales such as Mexico in the Contemporary Mexican Spanish Corpus (*Corpus del Español Mexicano Contemporáneo*; CEMC; Colegio de México). Additionally, some efforts have concentrated on urban linguistics, exemplified

[1] https://elex.is/portfolio-item/rae/
[2] https://www.asale.org/
[3] https://www.corpusdelespanol.org/web-dial/
[4] https://www.cordiam.org/

by the Sociolinguistic Corpus of Mexico City (CSCM), which includes sub-corpora derived from interviews with individuals of varying educational backgrounds: high, middle, and low. Beyond focusing on geographic dialects of Spanish, certain projects have delved into specific linguistic registers (Sánchez-Gutiérrez, De Cock and Tracy-Ventura, 2022). For instance, the Valencia Colloquial Spanish corpus[5] (Val.Es.Co) investigates the nuances of casual speech by analyzing spontaneous dialogues in Valencia, Spain. Building on this approach, the America and Spain Colloquial Spanish project[6] (AMERESCO) expands the scope by gathering conversational samples from cities across Mexico, Colombia, Argentina, Cuba, Panama, and Chile, further enriching our understanding of colloquial Spanish variants.

The history of Spanish spoken corpora began in the late 20th century, focusing on specific dialects or regions and different social strata, such as the Corpus Oral de Referencia de la Lengua Española Contemporánea (Oral Reference Corpus of Contemporary Spanish; CORLEC)[7], which was completed in 1992 and included an impressive collection of metadata for its database of roughly 1,100,000 words. This metadata includes the professions of interlocutors, their age, sex, location and setting, etc., as the following excerpt demonstrates.

```
<tape 012>
<ACON012B.ASC>
<14-5-91>
<source=family conversation>
<location=Madrid>
<terms=fair, Sunday, free time, photography, request,
physical appearance>
<H1=Doctor, woman , 28 years old>
<H2=Housewife, 55 years old>
<H3=Philologist, 23 years old>
<text>
<H1> ......................
```

[5] https://www.uv.es/corpusvalesco/
[6] https://esvaratenuacion.es/ameresco
[7] http://www.lllf.uam.es/ESP/Info%20Corlec.html

```
<H2> ..............  .........
<H3> ........................
</text>
<tape 005>
<ccon005c.asc>
<27-1-91>
<source=conversation between friends recorded on a Ma-
drid-Segovia commuter train>
<location= Tablada>
<terms=peseta, lottery>
<H1=woman, c . 45 years>
<H2=male, c. 40 years>
<H3=woman, c. 45 years>
<text></text>
```

This beginning was followed by an investigation into different social, rather than merely geographical[8] denominations by the Corpus Oral y Sonoro del Español Rural (COSER)[9] project (1994) (Fernández-Ordóñez and Pato, 2020), which comprises recordings of the language spoken in rural areas of the Iberian Peninsula. The interviews were obtained with the aim of offering a representative sample of the range of dialectal variation, but they also provide an insight into the ways of life in the countryside in the period prior to agricultural mechanisation and rural depopulation.

The digital revolution (the end of 1990s and 2000s) saw the adoption of digital technology, which significantly influenced the scope of Spanish spoken corpora, giving rise to the PRESEEA[10] project. PRESEEA is a project aimed at providing a framework for a sociolinguistic study of Spanish from Spain as well as the Americas. The goal is to coordinate sociolinguistic researchers from Spain and the Hispanic America in order to enable comparisons between different studies and materials, as well as a basic information exchange.

[8] For the geographical distribution of interviewees, see http://www.corpusrural.es/ING/mapa.php.
[9] http://www.corpusrural.es/ING/
[10] https://preseea.uah.es/

Of special note for the present paper is the Spanish Learner Language Oral Corpora project (SPLLOC)[11], which aimed to promote research on the acquisition of Spanish as a foreign language, by creating a smaller, but high-quality database, thus providing an insight into the process of learning (as well as teaching) Spanish.

3. The PRACOMUL platform and corpus construction

This section describes the inner workings and workflow of the PRACOMUL platform. The platform itself is designed in terms of different "tiers", i.e. user profiles, which also determine the sections of the platform that a particular user is allowed to access or edit, and the entire workflow (REGISTRATION → ROLES → POSTPRODUCTION) for a particular user is dependent on their role. Figure 1 demonstrates different the types of users and their roles:

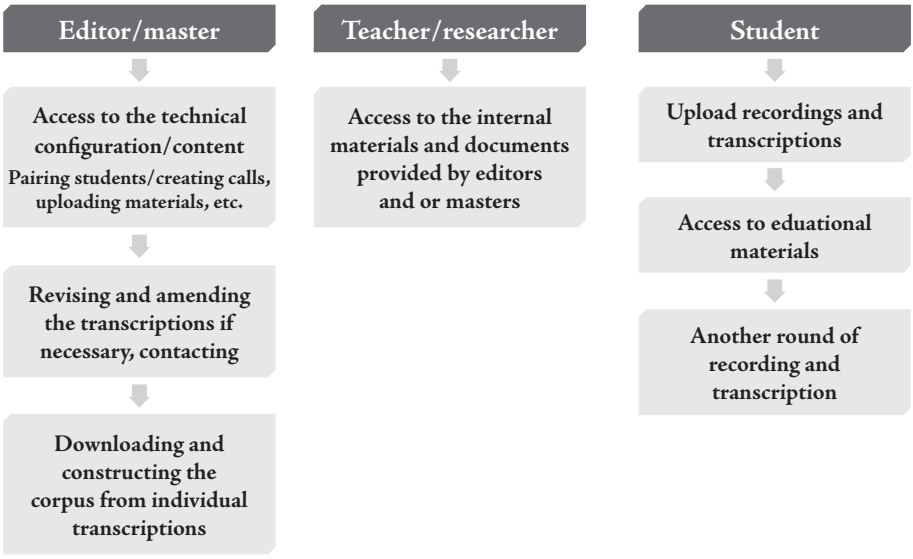

Figure 1: *Outline of the PRACOMUL platform types of users.*

[11] http://www.splloc.soton.ac.uk/

As Figure 1 demonstrates, there are three main user tiers made up of five user roles (master, editor, teacher, researcher, and student). For now, at least several of the distinctions (e.g. the roles of teachers and researchers) serve as placeholders for potential future roles, depending on the content and functions that may yet be deployed on the platform.

3.1 Registration and Metadata Collection

During the registration section of the workflow, the users are requested to submit specific personal and demographic data, i.e.:

- Gender
- Nationality
- Mother tongue
- Place of study
- Age bracket
- Spanish proficiency (self-assessment from A1 to C2)

This information is automatically stored in the system and may then be exported from the system upon creation of the corpus. This essentially means that the text file containing a particular transcription includes a header as in the following illustrative excerpt demonstrates:

```
<doc id="/" age_A="2" gender_A="F" mother_tongue_A="fr"
spanish_level_A="B2" nationality_A="BE" university_
A="Vrije Universiteit Brussel" age_B="3" gender_B="F"
mother_tongue_B="sl" spanish_level_B="C1" nationality_
B="SI" university_B="Univerza v Ljubljani">
```

As we can discern from the actual header (with the doc id number omitted for privacy purposes), we are dealing with a transcription of a dialogue recorded by two female students, the first belonging to the 18-24 age group, from Vrije Universiteit Brussel, of Belgian nationality and with French as mother tongue. The second student is a slightly older (24+) female Slovenian student at the University of Ljubljana, with Slovene as her mother tongue, and with a C1 level of Spanish. This metadata can then

be used in the analysis of the corpus in the SketchEngine concordances, as Figure 1 demonstrates.

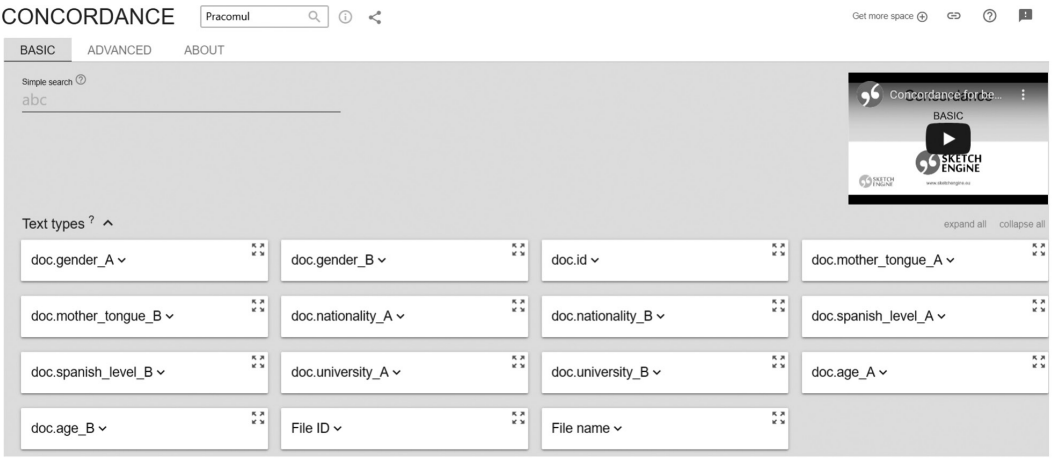

Figure 2: *The SketchEngine text type selection menu in the PRACOMUL corpus.*

Using this metadata, we can configure our search queries and limit the search to a particular subgroup (subcorpus) of the entire student cohort.

3.2 Uploading Transcriptions and Dialogues

Following the registration, all students are (automatically) paired with students of different universities (if possible). They are then required to record a transcription and transcribe it using a template that is available to them[12]. Afterwards, the students upload the audio file and the transcription in the interface, as illustrated in Figure 2.

[12] The template is designed in such a way that a portion of the transcription, containing some personal information, is already filled out as the data is pulled from the user's profile. The main reason for this is that students tend to fill out this information incorrectly, which makes it necessary to manually amend the information.

Figure 3: *The upload/review interface, with personal information omitted.*

As we can discern from the screenshot, the interface allows for simultaneous listening to the audio file and editing the transcript, which is especially useful for the later stages of corpus production, which essentially means manual correction of transcriptions. In addition, we have the option to set the status of a particular transcription and add comments for other editors and masters.

3.3 Educational Materials

This section of the PRACOMUL platform represents a critical component of our approach to language acquisition and education. This curated repository is made exclusively available to students who actively contribute to the corpus by submitting their transcriptions (i.e. access is granted to those who have submitted their transcripts), thereby fostering a participatory learning environment. In an ideal setting, the students are expected to record and transcribe their dialogues anew, after having studied the material available at the PRACOMUL platform. This, of course, gives educators a chance to assess their teaching methods, as well as giving students the opportunity to see if their approaches in learning Spanish are effective.

At present,[13] there are seven teaching units available, relating to the following discursive elements and with varying degrees of difficulty:

- *bueno* (B1)
- *bueno* (C1)
- *entonces* (B1)
- *pues* (B1)
- *pues* (C1)
- *digamos* (B1, B2)
- *por favor* (C1)

The inclusion of multimedia teaching materials is grounded in the pedagogical principle that language learning is most effective when it is interactive, engaging and tailored to the individual's learning journey. These resources are selected and structured to gradually build upon the learners' existing knowledge, reinforcing their skills through practical application and exposure to varied linguistic contexts.

Moreover, the availability of these materials on the PRACOMUL platform serves a dual purpose: it rewards students for their contributions to the growing corpus and simultaneously empowers them to take charge of their learning process. By integrating the task of transcription submission with access to high-quality educational content, we not only enrich our database but also invest in the linguistic and cultural proficiency of our participants. This symbiotic relationship between contribution and learning underscores the PRACOMUL project, making it a cornerstone of our mission to promote Spanish language learning through innovative, technology-driven means.

[13] As of March 2024.

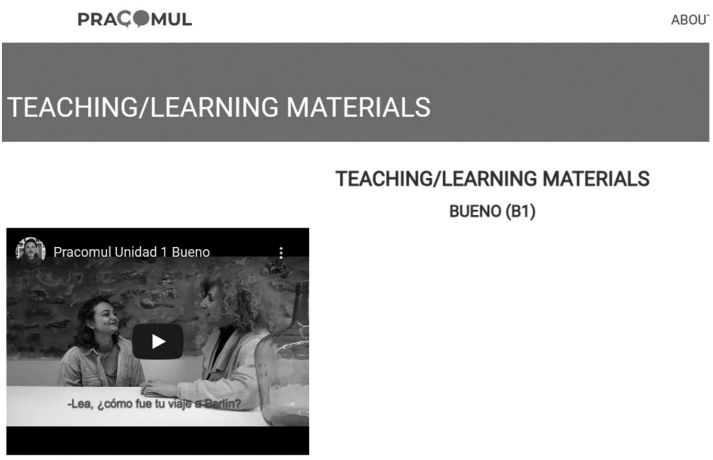

Figure 4: *A screenshot of a teaching material available to students, pertaining to the use of the word* bueno.

3.4 Normalization and Tagging

In the final stages of development and refinement of our corpus, two crucial processes were employed: normalization and part-of-speech (POS) tagging, each serving a distinct purpose in enhancing the linguistic accuracy and utility of our dataset. The normalization process was carried out with the involvement of student contributors, who played a key role in identifying and marking non-standard words within the transcriptions. By appending an asterisk to each non-standard form, students not only highlighted linguistic variations but also provided the corresponding standard spelling. This meticulous process enriched our understanding of language use among the student population and provided insight into the most common non-standard spellings in contemporary Spanish.

Following the normalization, the corpus underwent a POS tagging procedure using SketchEngine[14], a powerful corpus management and analysis tool. This step involved creating a new corpus within SketchEngine, tagging each word with its corresponding part of speech tag to capture the grammatical structure of the language as represented in the texts which we

[14] https://www.sketchengine.eu/

had collected. The outcome of this process was then exported as a vert file, a format conducive to further linguistic analysis and research, as shown in the following excerpt (showing only two sentences: *¿qué edad tienes? ¡Ah sí!*.

```
<s>
<text>
<A>
¿     Fia   ¿-x   ¿     Fia   ¿     0     0
<g/>
qué   DT0CN0      qué-x qué   DT0CN0      qué   F     S
edad  NCFS000     edad-n      edad  NCFS000     edad  F
S
tienes      VMIP2S0     tener-v     tener VMIP2S0
tener 0     0
<g/>
?     Fit   ?-x   ?     Fit   ?     0     0
</s>
<s>
</A>
<B>
¡     Faa   ¡-x   ¡     Faa   ¡     0     0
<g/>
Ah    I     ah-x  ah    I     ah    0     0
sí    RG    sí-r  sí    RG    sí    0     0
<g/>
!     Fat   !-x   !     Fat   !     0     0
</s>
```

4. Analysis of the PRACOMUL Corpus

This section provides an overview and a preliminary analysis of the PRACOMUL corpus. First, basic statistical information about the corpus itself is provided.

4.1 The Corpus in Numbers

In this section, the main characteristics of the PRACOMUL corpus in terms of numbers are presented, making use of the metadata gathered upon the registration of the students, which makes it possible to automatically harvest the information and gain an insight into the contents of the corpus. Table 1 presents the corpus in numbers.

Table 1: *The PRACOMUL corpus in numbers.*

Category	**Count**
Tokens	134,799
Words	108,303
Sentences	5,828
Documents (i.e. transcriptions)	65

As shown in Table 1, the corpus contains 65 transcriptions[15], and a little over 100,000 words, with several more student recording campaigns poised to take place in the coming years. The following table shows the makeup of the 65 transcriptions, with the individual pairs between the interlocutors (A and B) specified.

Table 2: *Language pairs by nationality.*

Mother tongue A	**Mother tongue B**	**Frequency**
SL	IT	17
SL	ES	7
SL	NL	7
IT	SL	6
SL	SL	5

[15] As of November 29, 2023.

Mother tongue A	Mother tongue B	Frequency
ES	SL	4
FR	SL	3
SL	FR	2
ES	NL	2
NL	IT	2
ES	FR	1
SR	NL	1
ES	ES	1
IT	NL	1
RO	NL	1
NL	SL	1
IT	IT	1
ES	IT	1
MK	IT	1
EN	SL	1

The most frequent pairing involves students whose mother tongues are Slovenian (SL) and Italian (IT), with a total of 17 instances of such interactions recorded. This is followed by pairs where one student is Slovenian and the other is Spanish (ES) and pairs of Slovenian and Dutch (NL) speakers, each with 7 instances. Table 3 gives the pairs by mother tongue, regardless of the interlocutors' role (A or B), i.e. the total values for each language pair.

Table 3: *Collated pairs by mother tongue*.

Mother tongue A	Mother tongue B	Total
IT	SL	23
ES	SL	11
NL	SL	8
SL	SL	5
FR	SL	5
ES	NL	2
IT	NL	3
ES	FR	1
NL	SR	1
ES	ES	1
NL	RO	1
IT	IT	1
ES	IT	1
IT	MK	1
EN	SL	1

As Table 3 shows, the most frequent interaction involves Italian and Slovenian speakers, with a total of 23 instances. The second most common pairing involves Spanish and Slovenian speakers, with 11 instances. Dutch and Slovenian speakers interacted in 8 instances, showing another active exchange. There were 5 instances where Slovenian speakers interacted among themselves in Spanish, which suggests that no other pairing was possible at that particular time. French and Slovenian speakers also showed a moderate level of interaction, with 5 instances recorded. Italian and Dutch speakers interacted 3 times, and there were 2 instances of interaction between Spanish and Dutch speakers.

The following pairings each had 1 instance of interaction: Spanish (ES) - French (FR), Dutch (NL) - Serbian (SR), Spanish (ES) - Spanish (ES), Dutch (NL) - Romanian (RO), Italian (IT) - Italian (IT), Spanish (ES)

- Italian (IT), Italian (IT) - Macedonian (MK), and English (EN) - Slovenian (SL).

This initial analysis sets the stage for deeper inquiries into how these linguistic pairings might influence the use of Spanish among non-native speakers, offering a foundation for exploring linguistic phenomena such as language transfer, the adoption of discourse markers, and the negotiation of meaning across different mother tongues.

However, as Table 3 shows, there is an unevenness in the distribution of the language pairs, which was anticipated as it reflects the mother tongues of the PRACOMUL project partners and their (mis)match in schedules. This is confirmed by Table 4, which shows the overall distribution of interlocutors by mother tongue.

Table 4: *Interlocutors by mother tongue (regardless of role, A or B).*

Mother tongue	**Frequency**
SL	58
IT	30
ES	17
NL	15
FR	6
EN	1
MK	1
RO	1
SR	1
Total	**130**

As shown in Table 4, Slovene (SL) is the most represented mother tongue among participants, totalling 58 students. Following Slovene, Italian (IT) is the second most common language, with 30 speakers. Spanish (ES) speakers are also well represented with 17 individuals, and Dutch (NL) speakers make up another considerable group with 15 participants. French (FR) speakers, though fewer in number, add to the corpus's lin-

guistic variety with 6 individuals. Additionally, the corpus includes unique representations from English (EN), Macedonian (MK), Romanian (RO), and Serbian (SR) speakers.

If we compare this with the data on nationality, we get very similar, but not entirely overlapping results, as shown in Table 5.

Table 5: *Distribution of students by nationality.*

Nationality	**Count**
SI	60
IT	30
BE	21
ES	12
MX	2
CL	1
NI	1
PE	1
RO	1
US	1
Total	**130**

As shown in Table 5, the national makeup of the participants in quite varied, with several "outliers" in terms of geographical remoteness (Mexico, Colombia, Nicaragua, Peru as well as the US). For the future, we can only wish to include as many students as possible from different Spanish-speaking countries and thus create a truly balanced language resource.

Another important aspect of the corpus is its makeup in terms of gender, as shown by Table 6.

Table 6: *Distribution of students per gender.*

Female	**Male**
113	17

The gender distribution highlights a significant predominance of female participants, with a total of 113 female students contributing to the dataset. In contrast, the representation of male students is notably lower, comprising only 17 participants. This discrepancy underscores a gender imbalance within the corpus, which, in turn, reflects the distribution of genders in the student bodies of the participating universities.

The distribution of contributors in terms of age is also uneven, with most students belonging to the 18–24 age group, as shown in Table 7.

Table 7: *Distribution of students by age.*

Age bracket	**Frequency**
>18	1
18–24	115
24<	14
Total	**130**

The distribution shown in Table 7 is, of course, an expected result due to the makeup of the student body at the participating universities, which are given in Table 8, along with the total number of students.

Table 8: *Number of students by university.*

University	**Nr. of students**
Universidad de Sevilla	4 (3 %)
Università degli Studi di Palermo	29 (22 %)
Univerza v Ljubljani	66 (51 %)
Vrije Universiteit Brussel	31 (24 %)
Total	**130**

As Table 8 demonstrates, more than half of all transcriptions of the corpus (51 %) belong to students from the University of Ljubljana, with Università degli Studi di Palermo (22 %) and Vrije Universiteit Brussel

(24 %) almost tied in second place. Students of the University of Sevilla make up roughly 3% of the contributors.

Of special note and interest to us are the self-assigned proficiency levels of Spanish of students, as they can then be studied in terms of their self-perceived fluency and their actual output. The levels are given in Table 9.

Table 9: *Distribution of Spanish proficiency.*

Proficiency level	**Frequency (%)**
A2	4 (3 %)
B1	24 (18 %)
B2	58 (45 %)
C1	27 (21 %)
C2	17 (13 %)
Total	**130**

The largest group of students, 58 (45 %), self-assess their Spanish proficiency at the B2 level. This indicates a high intermediate level of Spanish, where students are typically able to handle complex texts, engage in spontaneous conversations, and produce clear, detailed texts on a wide range of subjects. This suggests that a significant portion of participants are confident in their ability to use Spanish effectively in various contexts.

A substantial number of students, 44 in total, or, more than a third (34 %), rate their proficiency at the advanced levels of C1 (27 students) and C2 (17 students). C1 learners can use Spanish fluently and spontaneously, while C2 learners are at a near-native level of proficiency, indicating that these students can understand virtually everything heard or read in Spanish and express themselves very fluently and precisely. This shows that a good portion of the participants considers themselves highly proficient.

24 (18 %) students assess their proficiency at the B1 level. This group represents those who are beyond basic language use but are still developing their ability to navigate more complex language tasks, whereas the smallest group, consisting of 4 students, places themselves at the A2 level, reflecting basic ability to communicate and exchange information on fa-

miliar matters in simple and direct exchanges. This suggests that only a few participants feel they might struggle with more complex language use and interactions.

This means that the corpus includes transcriptions from students with a very diverse range of self-assessed proficiency levels, with a significant skew towards the higher proficiency levels (B2, C1, and C2). This could indicate that most students engaging in these dialogues feel they have a solid foundation in Spanish and that they are capable of complex language use and understanding nuanced aspects of the language.

The presence of students across all proficiency levels from A2 to C2, if taken at face value, offers a rich dataset for analyzing how language proficiency influences the use of discourse markers and conversational strategies among non-native speakers.

The high number of students at the B2 level and above suggests that the interactions captured in the corpus are likely to demonstrate a wide range of linguistic competence, from those who are confidently navigating intermediate language tasks to those who are engaging with the language at an advanced or near-native level. This could provide insights into how proficiency affects language acquisition and usage in an academic context.

The small number of students at the A2 level might offer unique insights into the early stages of language acquisition and the specific challenges faced by beginners in using discourse markers and structuring conversations in Spanish.

4.2 Overview of the Most Common Normalized Words

In this section, we present the most common normalized words to provide some insight into the data that the PRACOMUL corpus provides. In total, 1055 normalized words can be found in the corpus, of which 282 appear more than once. The following table gives the most normalized words (only the lemmas with the frequency above 10 are included, 46 in total).

Table 10: *Most normalized lemmas in the PRACOMUL corpus.*

Normalized lemma	Frequency	Freq. per million words	% of concordance
emm → em	197	1461,43517	6,00793
amm → am	171	1268,5554	5,215
mm → mmm	112	830,8667	3,41568
ehmm → ehm	108	801,19289	3,29369
yy → y	108	801,19289	3,29369
yyy → y	81	600,89467	2,47027
umm → um	63	467,36252	1,92132
ee → eee	58	430,27025	1,76883
uhmm → uhm	55	408,0149	1,67734
ehh → eh	48	356,08573	1,46386
eeh → eh	48	356,08573	1,46386
quee → que	47	348,66727	1,43336
emmm → em	44	326,41192	1,34187
ammm → am	40	296,73811	1,21988
queee → que	36	267,0643	1,0979
Amm → Am	35	259,64584	1,0674
eh → ser	33	244,80894	1,0064
aa → aaa	31	229,97203	0,94541
porquee → porque	27	200,29822	0,82342
pandemía → pandemia	24	178,04286	0,73193
y → eee	23	170,62441	0,70143
dee → de	22	163,20596	0,67094
oo → o	21	155,78751	0,64044
deee → de	21	155,78751	0,64044
mmh → mh	20	148,36905	0,60994

Normalized lemma	Frequency	Freq. per million words	% of concordance
noo → no	20	148,36905	0,60994
hmm → hm	18	133,53215	0,54895
eem → em	18	133,53215	0,54895
peroo → pero	17	126,1137	0,51845
ahmm → ahm	17	126,1137	0,51845
loh → el	16	118,69524	0,48795
Ahh → ah	16	118,69524	0,48795
Hmm → Hm	15	111,27679	0,45746
jaja → ja	13	96,43988	0,39646
Yyy → y	12	89,02143	0,36597
ehhh → eh	12	89,02143	0,36597
Emm → Em	12	89,02143	0,36597
yyyy → y	12	89,02143	0,36597
Uhmm → Uhm	11	81,60298	0,33547
porqueee → porque	11	81,60298	0,33547
ehmmm → ehm	11	81,60298	0,33547
entonceh → entonces	11	81,60298	0,33547
Ehmm → Ehm	10	74,18453	0,30497
lah → el	10	74,18453	0,30497
perooo → pero	10	74,18453	0,30497
eehm → ehm	10	74,18453	0,30497

In this paper, we do not deal with the question whether all the normalizations are "correct" or suitable, we merely showcase the data. However, we can give several basic observations:

1) Frequent hesitation indicators: Words like "emm" (normalized to "em"), "amm" (normalized to "am"), and various forms of hesitation such as "mm," "uhmm," "ehmm," indicate frequent hesitation markers in the dialogues.

2) Lengthening: There's a tendency to lengthen words, as seen in "queee" (normalized to "que"), "nooo" (normalized to "no"), and "muyy" (normalized to "muy"). This phenomenon might reflect a strategy to buy time while thinking about what to say next or to emphasize certain words for clarity or emotional effect.
3) Informal spellings and Internet slang: The presence of terms like "jaja" (normalized to "ja"), which is akin to "haha" in English, points to the influence of digital communication and internet slang on non-native speakers. This reflects an adaptation of informal, colloquial language typically found in text messaging and social media.
4) Dialectal and regional variations: words like "pandemía" (normalized to "pandemia") might indicate either typographical errors or influences from the students' mother tongues. Additionally, variations in the spelling and use of certain terms may hint at the learners' exposure to different dialects or variations of Spanish.
5) Creative spellings reflecting pronunciation efforts: many of the non-standard forms appear to be phonetic spellings based on the students' interpretation of Spanish sounds, such as "emm" for "em" or "queee" for "que". This creativity in spelling underscores the learners' efforts to approximate the sounds of Spanish as they understand/hear them.

5. Conclusion

This study has provided an exploration of the PRACOMUL corpus of spoken Spanish among non-native speakers, primarily university students with diverse linguistic backgrounds. It aims to provide insights into the dynamics of second language acquisition and use, particularly focusing on discourse markers, informal and non-standard language usage. In addition, the paper describes the inner workings of the corpus and its compilation.

The corpus analysis revealed a prevalent use of hesitation indicators and informal spellings among the learners, reflecting the naturalistic, spontaneous nature of their spoken Spanish when conversing among themselves. This finding emphasizes the importance of understanding the fluidity and adaptability in language learning, where non-standard forms and fillers serve not only as linguistic strategies to manage conversations but also as

indicators of learners' engagement with the language (and perhaps also their language competence) in real-time.

The normalization process of the corpus has also shed light on the learners' phonetic interpretations of Spanish sounds, creative spellings, and adaptation to informal communication styles, influenced by digital communication norms and intercultural exchanges. Such insights underscore the evolving nature of language use among non-native speakers and the role of informal language practices in language acquisition.

In addition, the paper describes how the underlying project and the resulting PRACOMUL corpus can contribute to the broader discourse on second language acquisition by illustrating the intricate relationship between language proficiency, informal language use, and the sociolinguistic competence of non-native speakers, by offering a plethora of options for further quantitative and qualitative analysis.

The paper, and the underlying project, is also, in effect and perhaps most importantly, a blueprint for (semi)automatic corpus creation in an educational setting, which may set a foundation for language education frameworks to incorporate an understanding of informal and non-standard language use as essential components of communicative competence. Additionally, the findings highlight the desirability of the inclusion of digital literacy and intercultural communication skills in language teaching curricula, reflecting the contemporary linguistic landscape and the realities of global communication.

* * *

References

Corpus Oral de ELE Pracomul [en línea]. Proyecto Erasmus+ PRACOMUL. www.pracomul.si.

Adolphs, S., Carter, R. (2013): *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Oxford, New York: Routledge.

Baker, P. (2010): *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Biber, D. (2006): *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, D. (2020): "Corpus analysis of spoken discourse". In: O. Kang, S. Staples, K. Yaw, and K. Hirschi (eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*. Ames: Iowa State University, 5-7.

Davies, M. (2016): *Corpus del Español*. http://www.corpusdelespanol.org [1. 6. 2023]

Fernández-Ordóñez, I., Pato, E. (2020): "El *COSER (Corpus oral y sonoro del Español Rural) y su contribución al estudio de la variación gramatical*." In: Á. Gallego and Francesc Roca (eds.), *Dialectología digital del español, Verba* (volumen monográfico), Santiago de Compostela, 71-100.

Moreno-Fernández, F. (2005): "Corpora of Spoken Spanish Language: The Representativeness Issue." In: Y. Kawaguchi, S. Zaima, T. Takagaki, K. Shibano, M. Usami (eds.), *Linguistic Informatics – State of the Art and the Future. The first international conference on Linguistic Informatics*. Tokyo: John Benjamins, 120-144.

Sánchez-Gutiérrez, C., De Cock, B. and Tracy-Ventura, N. (2022): "Spanish corpora and their pedagogical uses: challenges and opportunities". *Journal of Spanish Language Teaching*, 9/2, 105-115.

* * *

ABOUT THE AUTHOR

Damjan Popič is an assistant professor at the Department of Translation Studies, Faculty of Arts, University of Ljubljana, and researcher at the Laboratory for Cognitive Modeling at the Faculty of Computer and Information Science, University of Ljubljana. In his teaching and research, he deals primarily with digital humanities, language technology, corpus linguistics, and designing and maintaining a language-technology ecosystem for Slovene speakers outside Slovenia. He is heavily involved with the Slovene Research Institute in Italy (SLORI, Trieste), aiming to produce a comprehensive infrastructure for the speakers of Slovene in Italy.

e-mail: damjan.popic@ff.uni-lj.si