Federal University of Minas Gerais
Faculty of Languages and Literature
Graduate Program in Linguistic Studies

João Victor Pessoa Rocha

Testing statistical methods for sociolinguistic profiling of Brazilian Portuguese speakers

Belo Horizonte
2024

FEDERAL UNIVERSITY OF MINAS GERAIS
(UFMG)

João Victor Pessoa Rocha

Testing statistical methods for sociolinguistic profiling of Brazilian Portuguese speakers

Thesis submitted to the Examining Board as a requirement to obtain a master's degree in Linguistic Studies from the Federal University of Minas Gerais, under the guidance of professors Heliana Ribeiro de Mello (UFMG) and Crysttian Arantes Paixão (UFBA).

Area: Theoretical and Descriptive Linguistics
Research track: Corpus Linguistics

Belo Horizonte
2024

R672t

Rocha, João Victor Pessoa.
Testing statistical methods for sociolinguistic profiling of Brazilian Portuguese speakers [recurso eletrônico] / João Victor Pessoa Rocha. – 2023.
1 recurso online (120 f.: il., color., p&b.): pdf.

Orientador: Heliana Ribeiro Mello .

Coorientador: Crysttian Arantes Paixão.

Área de concentração: Linguística Teórica e Descritiva.

Linha de pesquisa: Estudos Baseados em Corpora.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 95-106.
Apêndices: f. 107-120.

Corpus – Teses. I. Mello, Heliana. II. Paixão, Crysttian Arantes. III. Universidade Federal de Minas Gerais. Faculdade de Letras. IV. Título.

CDD : 410

Ficha catalográfica elaborada pela Bibliotecária Alessandra Oliveira Pereira - CRB-6/2616

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

Testing statistical methods for sociolinguistic profiling of Brazilian Portuguese speakers

JOÃO VICTOR PESSOA ROCHA

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA TEÓRICA E DESCRITIVA, linha de pesquisa Estudos Linguísticos Baseados em Corpora.

Aprovada em 23 de fevereiro de 2024, pela banca constituída pelos membros:

Prof(a). Heliana Ribeiro de Mello - Orientadora

UFMG

Prof(a). Crysttian Arantes Paixão - Coorientador

UFBA

Prof(a). Flavio Codeco Coelho

EMAp/FGV

Prof(a). Livia Oushiro

Unicamp

Belo Horizonte, 23 de fevereiro de 2024.

Documento assinado eletronicamente por **Heliana Ribeiro de Mello**, **Professora do Magistério Superior**, em 26/02/2024, às 15:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Crysttian Arantes Paixão**, **Usuário Externo**, em 27/02/2024, às 22:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Flávio Codeço Coelho**, **Usuário Externo**, em 05/03/2024, às 09:06, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Livia Oushiro**, **Usuária Externa**, em 06/03/2024, às 20:27, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2929890** e o código CRC **EB712F2F**.

Referência: Processo nº 23072.279407/2023-14 SEI nº 2929890

Acknowledgments

As an amazing partner in this endeavor, Prof. Heliana Ribeiro de Mello was a source of wisdom, ethics, and empathy. She is the epitome of an advisor. Advising is an understatement of what she did for me during our time together. Since our first encounter during my undergraduate years, she has stimulated my curiosity, pushed me to do activities at my utmost performance, and given me perspective on how to navigate the (sometimes egocentric) academic world. She did that with calmness, candidness, and warmth. None of this project would be the way it is without Prof. Heliana's incredible guidance. I could go on and on about her great qualities, but sum up, the words I can give and may represent my thankfulness are "I'm forever grateful for having you as my adviser". In English, there is the "give somebody their flowers" idiom, but I would give you a whole garden for all the teachable, funny, and enriching moments and talks we have had (and because you deserve it!).

In the first half of the Master's, prof. Heliana Mello suggested I should have a second advisor. Not surprisingly, she indicated Prof. Crysttian Paixão. He was very thoughtful, attentive, and patient. He has implemented great techniques to help me understand complex statistics concepts. Also, besides being a teacher and a research advisor, he mentored me in how to navigate in academia; establish healthy partnerships; and reflect on my future. Because of that, without his clear and considerate directions and tips, this thesis would not have been finished.

I also would like to thank:

(i) My parents for their life teachings and sacrifices so that I could have a proper education. Since I was little, they taught me the importance of studying and being committed to what I like! They have been great cheerleaders during my Master's.

(ii) My generous family members, from whom I highlight my cousin Diane (or "Naninha") and her husband Walter for their altruistic help and hospitality; and my aunt/godmother, Jamilia (or "Dinda") and her husband Washington who were the first to stimulate my love for the "letters" and keep inspiring to me!

(iii) My dear friends and colleagues from LEEL. Saulo Mendes, for his kind and selfless assistance and attention through several emails and meetings. Saulo Genghini, Átila Augusto, João Fekete, and Dalmo Buzato for their support, meetings at the lab, outings, curiosity with my research, and, of course, their friendship.

(iv) My fellow beloved friend: Daiane Soares has been with me since the first semester of our undergraduate major. I am deeply thankful for having you by my side since the beginning of my journey at UFMG! Thank you for all the support, laughs, breakdown cries, and friend therapy sessions!

(v) My adorable friends, Natália Coimbra, Laura Rosa, Regina Rosa, Gabriel Bovo, Cristiana Alves, and Leidiane Rezende for their kindness and encouragement.

(vi) My amazing therapist Pedro Ivo, who listened to all my struggles and helped me see healthier perspectives.

(vii) The POSLIN staff, more specifically Felipe and Flávia. They were very friendly and supportive whenever I went to the office to solve a question!

(viii) CAPES — for their funding during the first part of this research.

(ix) Finally, the software people who (in)directly contributed to this work: Google Translate, Overleaf, QuillBot, Thesaurus, Grammarly, Cambridge Dictionary, ChatGPT, Python, R, and Tick Tick (and many more!).

Ser Mineiro é dizer UAI e ser diferente; é ter marca registrada, é ter história. Ser Mineiro é ter simplicidade e pureza, humildade e modéstia, coragem e bravura, fidalguia e elegância[1]

Ser Mineiro — José Batista Queiroz

[1] To be from Minas Gerais is to say 'UAI' and be different; it's to have a trademark, to have a history. To be from Minas Gerais is to have simplicity and purity, humility and modesty, courage and bravery, nobility, and elegance (our translation).

Abstract

This work constitutes a computationally driven and cross-methodological analysis of sociolectal marker recognition, positioning it in the growing area of Computational Sociolinguistics. This research had two main goals: (i) selecting an efficient method for sociolect (dis)similarity recognition; and (ii) describing how speech transcriptions can help profile a speaker. The main term we used to describe an in-group's language was *sociolect* because we believe it is more accurate regarding what sociolinguists deal with. To this end, a spontaneous speech corpus of Brazilian Portuguese compiled according to the Language into Act theory (L-AcT) framework was used to extract the data. This linguistic resource provides, besides the transcriptions, the metadata information about the interaction and the speakers, sound files, sound-text alignment files, and transcriptions annotated with the PALAVRAS parser (Bick, 2000). To achieve the aforementioned goals, three methods were tested: (i) Variation-Based Distance and Similarity Modeling (VADIS) (Szmrecsanyi et al., 2019), (ii) Mann-Whitney test; and (iii) Poisson and Negative binomial (parametric modeling) with Estimated Marginal Means (EMM) (Searle et al., 1980) and Compact Letter Display (CLD) (Piepho, 2004). Each method was assessed in relation to twelve linguistic variables: apheretic forms, apocopated diminutives, foreign words, interjections, reduced and articulated prepositions, pronoun phenomena, rhotacism, pronunciation of *senhor/senhora*, non-standard negation particles, non-standard plural marking in noun phrases, non-standard verb conjugation, and non-standard verb agreement. The VADIS methodology was not successful at fitting our data, because of data conversion from numerical to categorical and the amount of data available. On the other hand, the non-parametric model was able to retrieve significant predictors for ten linguistic phenomena and show the sociolect similarity, but it did not capture any predictor interaction. However, the parametric model retrieved significant predictors for seven response variables and two double predictor interactions, displaying more intricate sociolect groupings. Therefore, according to the findings, the Poisson and Negative binomial models alongside EMM and CLD are productive methods to linguistically profile speakers through speech transcription. Furthermore, our study emphasized the role of sociolects as powerful social markers, uncovering complex relations between society and language. Finally, this thesis advances the sociolinguistics field by the implementation of computational methods in research about Brazilian Portuguese.

Keywords: sociolinguistic profiling; sociolects; speaker; modeling.

Resumo

Este trabalho constitui uma análise computacional e intermetodológica do reconhecimento de marcadores sociolectais, posicionando-o na crescente área da Sociolinguística Computacional. Esta pesquisa teve dois objetivos principais: (i) selecionar um método eficiente para reconhecimento de (dis)similaridade socioletal; e (ii) descrever como as transcrições de fala podem ajudar a traçar o perfil de um locutor. O principal termo que usamos para descrever a linguagem de um grupo foi *socioleto* porque acreditamos que é mais preciso em relação ao que os sociolinguistas tratam. Para este fim, um corpus de fala espontânea do português brasileiro compilado de acordo com a abordagem da Teoria da Língua em Ato foi utilizado para extrair os dados. Este recurso linguístico fornece, além das transcrições, as informações de metadados sobre a interação e os locutores, arquivos de som, arquivos de alinhamento de texto sonoro e transcrições anotadas com o *parser* PALAVRAS (Bick, 2000). Três métodos foram testados: (i) *Variation-Based Distance and Similarity Modeling* (VADIS) (Szmrecsanyi et al., 2019), (ii) Teste de Mann-whitney (modelagem não paramétrica); e (iii) Poisson e Binomial Negativo (modelagem paramétrica) com Médias Marginais Estimadas (EMM) (Searle et al., 1980) e *Compact Letter Display* (CLD) (Piepho, 2004). Cada método foi avaliado em relação a doze variáveis linguísticas: formas aféricas, diminutivos apocopados, palavras estrangeiras, interjeições, preposições reduzidas e articuladas, fenômenos pronominais, rotacismo, pronúncia de senhor/senhora, partículas de negação não-padrão, marcação plural não-padrão em sintagmas nominais, conjugação verbal não-padrão, e concordância verbal não-padrão. A metodologia VADIS não teve sucesso no ajuste dos nossos dados, devido à conversão de dados de numéricos para categóricos e à quantidade de dados disponível. Por outro lado, o modelo não paramétrico foi capaz de indicar preditores significativos para dez fenômenos linguísticos e mostrar a similaridade socioletal, mas não capturou nenhuma interação entre preditores. No entanto, o modelo paramétrico apontou preditores significativos para sete variáveis de resposta e duas interações duplas de preditores, exibindo agrupamentos socioletais mais complexos. Portanto, de acordo com os resultados, os modelos Poisson e Negativo binomial, juntamente com EMM e CLD, são métodos produtivos para traçar o perfil linguístico dos falantes por meio da transcrição de sua fala. Além disso, nosso estudo enfatizou o papel dos socioletos como poderosos marcadores sociais, revelando relações complexas entre sociedade e linguagem. Por fim, esta dissertação traz avanços no campo da sociolinguística pela implementação de métodos computacionais em pesquisas sobre o português brasileiro.

Palavras-chave: perfilamento sociolinguístico; socioletos; falante; modelagem.

List of Figures

1.1 Variety hierarchy circles . . . 15

2.1 Linguistics Circles . . . 23
2.2 Areas of Language and Society . . . 26
2.3 Linguistics subfields . . . 29
2.4 New reading of Linguistics subfields . . . 30

4.1 f0 contour in extract 1 from bfamdl04 . . . 43
4.2 f0 contour in extract 2 from bfamdl04 . . . 43
4.3 f0 contour in extract 1 from bfammn03 . . . 44
4.4 f0 contour in extract 1 from bpubdl06 . . . 45
4.5 Example of metadata file . . . 46

5.1 Sex distribution . . . 50
5.2 Age distribution . . . 51
5.3 Schooling distribution . . . 51

6.1 Corpus linguistics stream steps . . . 54
6.2 Extract from the file bfamcv01.cg.pos.txt . . . 58
6.3 Flowchart of preprocessing and extracting features of the transcriptions . . . 58
6.4 Extract from the data frame with the different configurations of each utterance . . . 60
6.5 Flowchart of preprocessing and extracting features of the header files . . . 61
6.6 Metadata annotation . . . 62
6.8 VADIS workflow . . . 66

7.1 Preposition variable distribution . . . 73
7.2 Interjection variable distribution . . . 73
7.3 Geometric average population growth rate by household status in Minas Gerais . . . 84

A.1 Minas Gerais state map . . . 105
A.2 Belo Horizonte metropolitan area map . . . 105

I.1 Non-standard plural marking in NPs distribution considering sex . . . 118

List of Tables

3.1 Different focuses among the sociolinguistic unit terms . . . 39

4.1 Acts in L-AcT . . . 42

5.1 Comparing C-ORAL-BRASIL I and C-ORAL-ROM . . . 48
5.2 C-ORAL-BRASIL I interaction typology . . . 49
5.3 Age, sex and schooling codes and their description . . . 52

6.1 Model variables . . . 56
6.2 Source of each model variable . . . 61
6.3 Sample numerical description . . . 63
6.4 New age spans after adjustments . . . 65

7.1 Results of the Mann-Whitney test application considering the age variable . . . 75
7.2 Results of the Mann-Whitney test application considering the schooling variable . . . 76
7.3 Results of the Mann-Whitney test application considering the sex variable . . . 76
7.4 List of formulas and models per linguistic variable . . . 77
7.5 Grouping and EMMs of reduced and articulated prepositions . . . 78
7.6 Grouping and EMMs of *senhor/senhora* pronunciation . . . 79
7.7 Grouping and EMMs in non-standard plural marking . . . 80
7.8 Grouping and EMMs in foreign words . . . 81
7.9 Grouping and EMMs in non-standard verbal agreements . . . 81
7.10 Grouping and EMMs in non-standard negation particles . . . 82
7.11 Grouping and EMMs in non-standard verb conjugations . . . 84
7.12 Grouping and EMMs in apocopated diminutives . . . 85
7.13 Comparison between significant predictors in Mann-Whitney and Count Data Model results 86
7.14 Sample data distribution among the social variables in the percentage of words . . . 87
7.15 Summary of results from CLD . . . 88

H.1 Intercept coefficients . . . 117

Glossary

AI	Artificial Intelligence. 15, 18
C-ORAL-BRASIL I	Informal Spoken Brazilian Portuguese Reference Corpus. xi, 16, 18, 28, 37, 41, 45–49, 52, 54, 55, 60, 65, 67, 74, 90, 91
C-ORAL-ROM	Integrated Reference Corpora for Spoken Romance Languages. xi, 45, 47, 48, 54, 55
CL	Corpus Linguistics. 27, 28, 31, 38, 45, 47, 57, 72
CLD	Compact Letter Display. xi, 16, 69–71, 86–88, 91
CompSoc	Computational Sociolinguistics. 18, 30–33, 47, 72, 90
CVL	Corpus-based Variationist Linguistics. 28
DET	Determiner. 56
EMM	Estimated marginal means. 16, 69, 71, 77–79, 81, 82, 85, 87, 88, 91
GID	Group Identity. 21
GML	Generalized Linear Model. 68
L-AcT	Language into Act Theory. xi, 41, 42, 44–46, 90
MLE	Maximum likelihood estimation. 69
N	Noun. 56
NB	Negative binomial model. 69, 77
NLP	Natural Language Processing. 31, 32
NP	Noun phrase. 79, 80, 87
PERS	Personal pronoun. 56
PL	Plural. 20, 56, 81
PO	Poisson model. 69, 77
POS	Part of speech. 18, 57, 59

PT-BR	Brazilian Portuguese. 15, 16, 20, 21, 37, 39, 44, 79, 81, 82, 91, 92, 107
S	Singular. 20, 44, 56, 81
V	Verb. 56
VADIS	Variation-Based Distance and Similarity Modeling. 65–69, 74, 91

Contents

1 Introduction **14**
- 1.1 Some remarks about profiling . . . 17
- 1.2 Organization of the text . . . 18

2 A bird's eye view on the development of Sociolinguistics **19**
- 2.1 What is *Sociolinguistics*? . . . 19
  - 2.1.1 Social pressure and context in focus . . . 19
  - 2.1.2 Speech community and their practices . . . 20
  - 2.1.3 Identity and style . . . 21
- 2.2 What guides Sociolinguistics? . . . 22
- 2.3 A background on Sociolinguistics . . . 24
  - 2.3.1 What about the recent stages of Sociolinguistics? . . . 25
  - 2.3.2 Defining Computational Sociolinguistics . . . 30

3 An agenda for sociolects **34**
- 3.1 Communities . . . 34
- 3.2 Lectal variation . . . 36

4 Utterances, speech, and actionality **41**

5 The data dive **47**
- 5.1 Diastratic variation in the corpus . . . 50

6 Uncovering the method **53**
- 6.1 Corpus linguistics . . . 53
- 6.2 Data engineering . . . 57
  - 6.2.1 Preprocessing and extracting features of the transcriptions . . . 57
  - 6.2.2 Preprocessing and extracting features of the headers . . . 61
  - 6.2.3 Sample description . . . 63
- 6.3 Data science . . . 64
  - 6.3.1 Data visualization and sample testing . . . 65
  - 6.3.2 VADIS — lines of evidence . . . 65
  - 6.3.3 Non-parametric model . . . 68
  - 6.3.4 Count data models . . . 68
- 6.4 Sociolinguistics . . . 70

7 Sociolectal dynamics: mapping out variation **72**
7.1 VADIS assessment . . . 74
7.2 Non-parametric testing . . . 74
7.3 Count data models . . . 77
7.3.1 Reduced and articulated prepositions . . . 78
7.3.2 *Senhor/senhora* pronunciation . . . 78
7.3.3 Non-standard plural marking in NPs . . . 79
7.3.4 Foreign words . . . 81
7.3.5 Non-standard verb agreement . . . 81
7.3.6 Non-standard negation particle . . . 82
7.3.7 Non-standard verb conjugation . . . 84
7.3.8 Apocopated diminutives . . . 85
7.4 Comparison between count data and non-parametric models . . . 86
7.5 Sociolinguistic profiling summary . . . 87

8 Final words **90**

Bibliography **92**

A Minas Gerais and Belo Horizonte maps **105**

B Parsed transcription extract **106**

C List of criteria not implemented **107**

D PALAVRAS tagset **108**

E 3-grams in sociolects divided by sex **110**

F 3-grams in sociolects divided by schooling **112**

G 3-grams in sociolects divided by age **114**

H Intercept coefficients **117**

I Box plot of non-standard plural in NPs according to sex **118**

Chapter 1

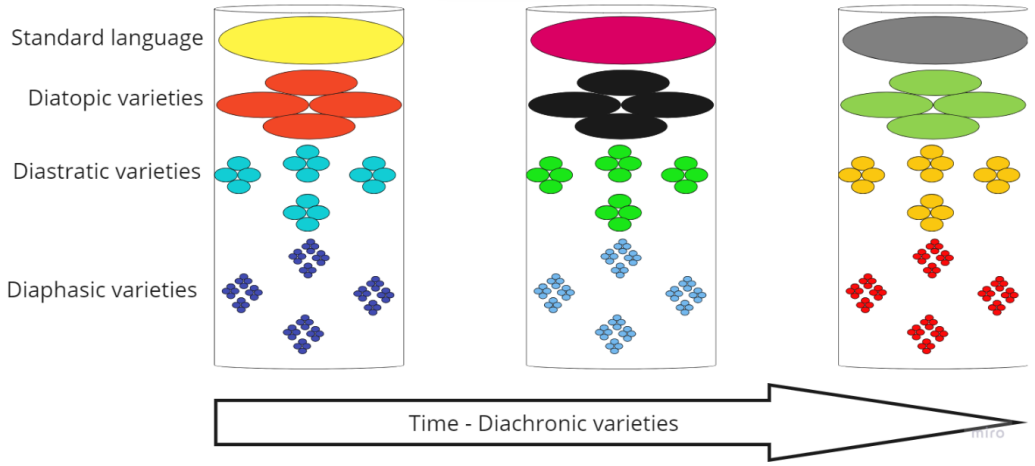
Introduction

One of the most famous research agendas in sociolinguistics is analyzing certain linguistic phenomena as dependent on macro-categories (e.g., sex and education level). Labov (1986) work, one of the seminal papers of modern Variationist Sociolinguistics, analyzed the rhoticity in New York accent based on social class and this approach was used as a basis for the next studies. Since then, sociolinguists have been focusing on how language holds information about its speakers whether on a group or individual level.

However, language variation has not become something easy to examine. There are a lot of factors that interplay in language use, for instance, cognition, register, context, and health conditions. Therefore, the challenge is how to make an affordable research scope without losing important "pieces of the puzzle". In this thesis, the spotlight is on the speaker's social information (discussed in further chapters) due to time and scope constraints. Nonetheless, we acknowledge that other determinants are important to the linguistic variables analyzed. For this reason, alongside this thesis' findings, we would like to invite linguists to integrate other factors into their studies while employing a methodology similar to what was done here.

One clear illustration of such a convoluted network of interactions is the range of language dimensions in which variation can differ. In other words, the web-like structures and processes in language variation are due to the diasystemic structure of languages. A diasystem is a higher-level structural trait shared by several dialects (Weinreich, 1954; Vogelaer, 2013). Therefore, the diasystemic nature of language causes more complex interactions between linguistic dimensions. For example, if two people speak differently because of the region in which they acquired their mother tongue, this is a case of diatopic, geographical, or dialectal variation. Additionally, diaphasia is associated with the variation that is mostly influenced by the communicative situation. Diamesic variation happens due to differences in the medium (speech vs. writing) (Raso, 2013), while diachronic or historical variation indicates differences throughout time. Now, diastratic or sociolinguistic variation refers to the variation that occurs because of the speaker's social background, which is the type of variation investigated here. Adapted from List (2014), Figure 1.1 demonstrates a model of organization for these variations, in which "standard language" is an interdialectal communication variety.

Figure 1.1: Variety hierarchy circles



Adapted from List (2014)

Amidst the intricate web of language varieties and their influencing factors, language analysts are often given the task of assessing linguistic samples to indicate if they were written/said by the same person or to indicate the possible individual's social group. This is either done through qualitative (especially if there are just a few texts) or quantitative methods (in the case of using software or a statistical model).

Thus, to cover such topics, the following research questions will be addressed in the next chapters:

1. What is the most suitable method for sociolect detection?
2. How can we distinguish Brazilian Portuguese speakers from different social groups through spontaneous speech transcriptions?
3. How coherent are these groups' sociolects?

Note that the contribution of this work is mainly methodological. Method validation is at the heart of Computational Sociolinguistics since it is a brand-new area in Linguistics, and the implementation of computational methods has endless methodological possibilities. Moreover, the context in which this research is set is one of exceptionally fast and frequent developments in modern linguistics, primarily due to the introduction of Artificial Intelligence (AI) into Humanities[1].

The general objective of this research is to **elicit an efficient method for sociolect (dis)similarity detection**. Another methodological goal is to describe how speech transcriptions can be source material to indicate different Brazilian Portuguese (PT-BR) sociolects using computational methods. Finally, we aimed to describe the sociolectal markers in PT-BR. For this purpose, the transcriptions, along with the respective metadata, of the corpus of informal speech texts from the C-ORAL BRASIL project (Raso and Mello, 2012) were used as empirical data.

Considering the methodological validation nature of this work, three analyses were performed. The first one was with the Variation-Based Distance and Similarity Modeling (VADIS) (Szmrecsanyi et al., 2019), which is a method that takes on the principles of dialectometry and variationist sociolinguistics in order to

[1] Causing the creation of the area of Digital Humanities.

find differences and similarities between language varieties. The second analysis was with Mann-Whitney, a non-parametric model, which also showed similarities between sociolects as well as provided points for attention for the parametric models. The last method tested was Poisson and Negative binomial models (parametric) with Estimated Marginal Means (EMM) (Searle et al., 1980) and Compact Letter Display (CLD) (Piepho, 2004), which can present complex relations between predictors based on the models of count data.

All analyses were run with twelve linguistic variables, namely:

1. apheretic forms
2. apocopated diminutives
3. foreign words
4. interjections
5. reduced and articulated prepositions
6. pronoun phenomena
7. rhotacism
8. pronunciation of *senhor/senhora*[2][3]
9. non-standard negation particles
10. non-standard plural marking in noun phrases
11. non-standard verb conjugation
12. non-standard verb agreement

The majority of sociolinguistic research targets or concentrates on one or a few factors related to a speaker. It goes without saying that this shows they did not look at other speaker variables when conducting their research. A potential explanation could be the fact that sociolinguists tend to consider the social meanings of the social variables of all informants uniformly and may fail to recognize the interpersonal variations in the meaning of each social variable.

The type of philosophy we follow here is similar to the one in authorship attribution: in order to know what is specific for a group or a person, the analysis must be comparative, not isolated. For example, a sociolinguist can discern sociolinguistic markers in females' speech through a comparison with males' speech. This comparison is essential because speech is the intricate production of multiple dynamic features in sequence, rather than isolated significant variables (Fairclough, 2023). The importance of this comparison lies in the fact that the definition of "young people's sociolect", for instance, is contingent upon its contrast with sociolects associated with older people.

Having that in mind, the following list details the specific goals of the research reported here:

- Extract and clean transcriptions in C-ORAL-BRASIL I.
- Extract and clean metadata files in C-ORAL-BRASIL I.

[2] *Senhor* is equivalent to sir, and *senhora* to madam.
[3] We chose this pair of words because its variants are lexicalized in PT-BR

- Describe the sample[4] (number of words in total, number of speakers, words per utterance, to name a few elements and metrics that will be calculated).
- Correlate speakers, utterances, and social variables.
- Extract and tabulate linguistic variables.
- Systematize and implement statistical models.
- Analyze and report results.

The next section explores the task of sociolinguistic profiling: how it emerged and how it relates to the work done here.

1.1 Some remarks about profiling

In line with authorship attribution analysis and other forensic tasks, sociolect detection also relies on a comparative approach (Coulthard et al., 2010; Gomes et al., 2020). Sociolect, ethnolect, and social dialect studies have focused only on similarities between speakers of the same community, but the studied phenomena can be similarly repeated in other communities. Therefore, it is important to carry out between-group comparative analyses so that it is possible to say which elements are more or less frequent in one sociolect than in another.

Sociolinguistic profiling or **Speaker/Author profiling** is a sociolinguistic and forensic task that has the purpose of finding patterns related to one's sociolect in their language and providing possible social profiles (Perkins, 2021). According to Jessen (2007) and Schilling and Marsters (2015), age, gender, schooling level, dialect, foreign accent, native language, and health conditions[5] are some of the most common social factors examined in this task. Based on that, it is quite clear why this type of profiling is useful in a criminal setting: law enforcement officers can ask an expert to extract the characteristics of an unknown speaker/writer and help find a suspect (Jessen, 2007). Having that in mind, this study is among the first to delve into Brazilian Portuguese sociolinguistic profiling. Thus, we hope that its methodology can help forensic analysts narrow down suspects if they have linguistic data.

One of the most famous cases of sociolinguistic profiling in a real criminal context was the Unabomber's case. Someone, later known as Theodore Kaczynski, was setting up bombs in universities and threatening to explode airlines (FBI, 2016)[6]. Seventeen years after the first bomb explosion, the criminal sent a manifesto that was published in a newspaper. Professor Roger Shuy[7], a linguistics professor, was asked to analyze the manifesto and other notes left by the criminal. Shuy's investigation focused mostly on lexical, collocational, and thematic elements (Shuy, 2001; Leonard et al., 2017). When Kaczynski was finally arrested, they found out that Shuy's profiling[8] was more accurate than the FBI's.

Concerning sociolinguistic profiling in academia, one of the seminal works is Argamon et al. (2003). Using similar categories to Biber's Multidimensional Analysis (Biber, 1988), the authors examined gender and genre differences in a portion of the British National Corpus (Consortium et al., 2007). They found out,

[4] The sample extraction method will be further explained in Chapter 6.

[5] Specially illnesses that affect the vocal or nasal tract. For instance, laryngitis due to smoking can cause a raspy or creaky voice.

[6] https://www.fbi.gov/history/famous-cases/unabomber

[7] Roger Shuy is a distinguished Research Professor of Linguistics at Georgetown University. More information about him can be found at http://www.rogershuy.com/

[8] According to Shuy's profiling report, the Unabomber would be someone with a higher education degree, who lived in northern California for some time, had Catholic roots, and probably came from Chicago.

through Exponentiated Gradient algorithms, that female writers tend to insert more involvedness in their writing style through pronouns, while male authors would likely have more informativeness in their texts because of the use of noun modifiers. They also concluded that these phenomena suffer influence from the text genre as well. For instance, fiction texts, which were part of the subcorpus, are predisposed to have more involvedness as aim to have the reader's attention. From another point of view, the authors showed that non-fictional texts tend to be more informative, which is the case for academic articles.

As speaker profiling became more needed and a worthy challenge, the Author Profiling Task at PAN 2013[9] (Rangel et al., 2013) was created. This task had the goal of evaluating the submitted profiling methodologies and checking which one had the best performance with Spanish and English social media texts. Part of speech (POS) tags, named entities, and sentiment words were some of the categories used to do the profiling. Since its creation, the Author Profiling Task has been happening every year.

Forensic linguistics has also followed the tremendous technological evolution of recent years. More specifically with author profiling, more methods and tools were and have been created nowadays. One example is the Profiling-UD (Brunato et al., 2020), which extracts more than 130 features from texts with the aim of profiling authors. Although it is multilingual due to Universal Dependencies annotation and has a user-friendly web-based interface[10], it only shows the feature metrics and does not do any statistical computing to identify same-author texts or indicate same-group authors.

Even though there have been these endeavors, no method was tested and certified to be a cohesive profiling method in linguistics. This could be faced as something negative, but it opens space for experimentation and development, which is the case for this thesis. Therefore, we invite fellow linguists to be part of this venture, especially now in the era of machine learning and AI in which brand-new methods can be created.

The following section is a summary of each part of this thesis.

1.2 Organization of the text

This thesis is divided into eight chapters. Chapters 2 to 4 present the theoretical foundations of this thesis. In Chapter 2, a critical review of Sociolinguistics frameworks is provided as well as a definition of Computational Sociolinguistics (CompSoc). After discussing the field, the arguments for preferring the term "sociolect" are explained in Chapter 3. In Chapter 4, the main basis of the theory behind the architecture of C-ORAL-BRASIL I is briefly discussed.

Chapter 5 is devoted to explaining the main features of the corpus used here, including its sociolinguistic distribution. Following that, Chapter 6 details all methodological procedures taken here. The results are presented and examined in Chapter 7. Finally, the conclusions are claimed in Chapter 8.

In the next pages, the main assumptions of Sociolinguistics are debated, and a short presentation of CompSoc is provided.

[9] More information at https://pan.webis.de/clef13/pan13-web/author-profiling.html
[10] http://www.italianlp.it/demo/profiling-UD

Chapter 2

A bird's eye view on the development of Sociolinguistics

"Diga-me com quem andas, e eu te direi quem tu és"
"Birds of a feather, flock together"

The quote above is a popular Brazilian proverb that states you can know one's personality through the people they assemble with. This points to the socialization process and group membership that may be transparent in language. Because of that, the social domain of language is already accepted in the Linguistics scholar community[1]. Therefore, a question arises: why do we need a subarea called **sociolinguistics** and what does it entail? To answer this query, this chapter will work on the definitions and the historical intricacies of such a Linguistics discipline as well as present the emerging subfield of **Computational Sociolinguistics**.

2.1 What is *Sociolinguistics*?

According to morphology, "sociolinguistics" can be divided into two parts: "socio", referring to the social element; and "linguistics", concerning communication studies. However, only analyzing the morphemes that constitute the name of the field does not show the potential and principles behind Sociolinguistics.

There was a movement in the past which would use the terms Sociolinguistics and Sociology of Language interchangeably. Yet, they have different focus and purposes. While the Sociology of Language is centered on describing social organization through language, Sociolinguistics concentrates on the language-society relation to better understand language and communication (Wardhaugh, 2006). Nevertheless, "Sociolinguistics" is a broad term that covers several standpoints in regards to the relation language-society (Holmes, 2013). For that reason, the following items will bring a discussion concerning the multiple trends of the area, what they have in common, and Sociolinguistics' main foundational propositions.

2.1.1 Social pressure and context in focus

According to Wardhaugh (2006), Sociolinguistics deals with *"the social distribution of linguistic items, to consider how a particular linguistic variable (...) might relate to the formulation of a specific grammatical*

[1]The *Revista de Estudos da Linguagem, v. 26, n. 2.* has some papers that elaborate on this idea https://bit.ly/3NvwbEx

rule in a particular language or dialect, and even to the processes through which languages change" (p. 13). We can notice that social pressure and context are at the core of this type of definition. In other words, the way we are educated, the place where we come from, the practices in our household, and our economic situation have a great influence on how we behave linguistically. In Oushiro (2015), for example, the author found out that the male *paulistana*[2] community has a greater tendency not to use the Portuguese plural marking in noun phrases as the regular PT-BR grammar would demand. In contrast, they might say:

E1 Example of partial plural marking in PT-BR

(a) *os menino bonito*
the-PL boy-S handsome-S
'the handsome boys'

In comparison with other sociolinguistic factors, such as sexual orientation, the author indicated that heterosexual male people used this morphosyntactic process to reinforce their masculinity, whereas homosexual males and heterosexual females did not follow this tendency. Among other processes, the willingness to belong to a certain group created the pressure on the interviewed heterosexual males to consistently speak that way and the same can be applied to the documented women and homosexual males.

Another example inserted in this definition would be Timbane (2014), which investigated the sociolinguistic differences in Mozambican Portuguese, with a particular emphasis on the incorporation of loanwords originating from the multiple languages in the country. One of his findings was that middle-class people used more loanwords than the lower classes. Although the author does not mention any other social implication of this finding, it is possible to assert that there are sociolinguistic distinctions and such elements shape language use.

Moreover, language evaluation can only happen when contextually inserted, for instance, what is considered rule-breaking in some situations may be the norm in other circumstances. However, in attempting to provoke a change, speakers can use disrupting language in the same community and context that disregards a certain set of linguistic items. As an illustration, some scholars may consider idiomatic language in academic papers as poor writing, while other researchers purposefully use this type of language in their work (Miller, 2020).

In sociolinguistic research, language evaluation is often described as not only the value that lay people attribute to certain varieties (e.g., prestigious and stigmatized varieties) but also the highly institutionalized varieties (e.g., formal language) (Camacho et al., 2004). A notable case of variety prestigiousness is the importance given to British and North American English in additional language learning (Modiano, 1999; David, 2005). These countries have placed political and economic efforts into becoming leading nations and, as a consequence, their linguistic power has also increased. Again, social and somewhat extralinguistic events affect communities and their language.

2.1.2 Speech community and their practices

The ethnographic branch of sociolinguistics has an interest in describing the linguistic behavior that constitutes a given group. Although practitioners of this linguistic area can deal with broad and categorical factors as the one described in subsection 2.1.1, they are also involved in examining specific groups, such as immigrants in a specific city, hence the term "speech community" (further discussed in section 3.1).

It is worth explaining that "speech community" does not have a precise notion in the literature. Despite the recent accounts on that, scholars do not come to an agreement about the definition (Jacquemet (2019)

[2]From the city of São Paulo/Brazil.

and Leuckert (2020) problematize it further). The first critique of this term is that the linguistic behavior of a group is assumed to be stable clues of a community, leaving not so much room for variation and change (Morgan, 2004). The author continues arguing that the second issue is that language is considered only as a representing aspect of a group, thus, to some extent, disregarding the internal linguistic forces in language variation.

For the sake of the explanation aimed in this subsection, we are going to use this term because, as stated by Morgan (2004, p. 18), *"the concept of speech community binds the importance of local knowledge and communicative competence in discursive activities so that members can identify insiders from outsiders, those passing as members, and those living in contact zones and borderlands"*. Moreover, the simple fact that this notion combines linguistic materialization (speech) and the social organization of a group (community) is of great interest to Sociolinguistics.

As an example, Beers Fägersten (2007) conducted a study to find out if there is a correlation between the frequency and the offensiveness of swear words in the university student community. For this, students had to rate swear words and also were interviewed. The author found out that there were similarities across all the interviewed university students, but the degree of offensiveness was also sensitive to gender and ethnicity. In this case, it is possible to describe the general practices of a community but also its nuances at deeper levels.

In other words, scholars are curious about group identity (henceforth GID), which has been an extensively researched topic in the subarea of "speech community" studies. A suitable example for GID investigation would be Labov's famous study in Martha's Vineyard (Labov, 1963) on phonological variation. He discovered that the native Vineyarders had a non-standard dialect as a way of reaffirming their identity and "protesting" against the mainland dialect.

On the other hand, sociolinguists can also take a comparative approach and focus on the differences between communities, instead of each community's linguistic production. Carlos (2005) analyzed some phonetic-phonological phenomena of PT-BR to compare the parents' and their children's language. She observed that the children pronounced [v] instead of [b] differently than their parents in words as in the example below.

(i) Children's pronunciation: [va'χe] 'sweep'

(ii) Parents' pronunciation: [ba'χe] 'sweep'

Therefore, in this branch of Sociolinguistics, membership is at its core. A group can be defined by the linguistic and non-linguistic routines its participants perform. On account of the particular series of patterns, the sociolinguist looks for relatively stable indicators of a (speech) community, i.e., work on this type of Sociolinguistics is often related to the endeavor of describing a group's language.

2.1.3 Identity and style

Although studies in the speech community approach advanced the discussion about language and social groupings, they make identity and affiliation equivalent (Eckert, 2012). This suggests that a community is linguistically stable and its members are consistent in the same way, but this is not how it happens (Grant and MacLeod, 2020).

When proposing the wave trend in Sociolinguistics (corresponding to the one shown in this subsection), Eckert (2012) argues that *"variation constitutes a social semiotic system capable of expressing the full range of a community's social concerns. And as these concerns continually change, variables cannot be consensual markers of fixed meanings (...)"*. This approach succeeds in taking into account a person's agency and, as Eckert (2016) adds, the enactment of social roles through language styles.

In Podesva (2007), the author analyzed the use of falsetto phonation as a resource for social positioning. He further explained through a micro-level method that the falsetto can carry a discursive meaning and be interpreted as part of a "diva" or gay style. As demonstrated by the aforementioned study, this shift in the way sociolinguists view the relation individual-group considers the social meaning, the discursive effects, the contextual devices, and the agency as well as gives space for variation as a not-so-stable, non-linear, and constant process.

In this approach, the social basis of meaning is fronted. Eckert (2016) makes a useful explanation about the foregrounding of social meaning, which implies that, instead of focusing on the variables related to a social factor, linguists should pay more attention to the variables that play a role in the social and contextual situation. Of course social factors and their relation to language may arise, but the focus here is on how people position themselves in the semiotic landscape.

It is worth highlighting that the three branches presented here do not exclude each other and you may even find overlapping in some studies. However, they rather represent the focus and trends in sociolinguistic research. In addition, not coincidentally, these subareas represent the three waves of Language Variation Studies discussed in Eckert (2012). As argued by the author and presented in this thesis, each wave (or branch) examines the meaning construction process under a different lens. The work presented here may be considered a mix of the first and third waves. Because of the way sociolinguistic profiling was conceived, it seeks to have someone's social factors based on language and this is what the first wave intends to do. From another perspective, we also put in the foreground the notion of "lectal coherence" (further discussed in Chapter 3), being the co-occurrence of different variations. Therefore, the third wave, trying to construe someone's performed identity, may have a cohesive integration with co-variation.

In the next section, we will discuss the principles that guide sociolinguistics.

2.2 What guides Sociolinguistics?

Research on Sociolinguistics is based on linguistic principles and in this section, the focus is going to be on the three main ones. They are language as an orderly heterogeneous system (Mesthrie, 2008; Labov, 2010); variation as a carrier of social meaning (Campbell-Kibler, 2010); and community as a locus of research; (Eckert, 2006).

Language as an orderly heterogeneous system. Language, as a complex and dynamic system, exhibits both order and diversity. In this context, "orderly" indicates that language follows a set of patterns that allow for meaningful expression. On the other hand, "heterogeneous" highlights variability within the language system. It acknowledges the existence of multiple linguistic elements, variations in pronunciation, vocabulary choices, and language use across different contexts, regions, and social groups.

As claimed by Guy and Hinskens (2016, p.3),

> *This vision of orderly diversity implies that speech communities are sociolinguistically coherent, in the following sense: the orderly variables that define the community should collectively behave in parallel: variants (or rates of use of variants) that index a given style, status, or a social characteristic should co-occur. Coherent middle class speakers would use all the variants associated with their status, and speakers who are coherently signaling a 'casual' style would use all the 'casual' variants. The collocations of variants of different linguistic variables are social conventions, but they may be at least partly internally (structurally) motivated.*

From that, it is possible to claim that language is rule-based, but not from the normative grammar perspective. In Sociolinguistics, language has rules that are more fixed (general language parameters) and rules that

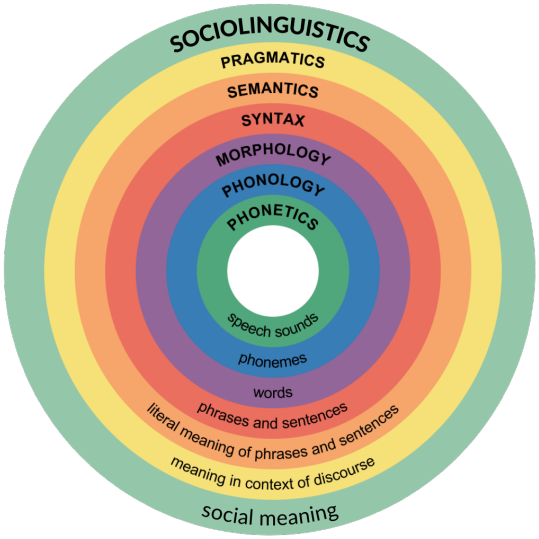
are more variable (lectal variation).

Variation as a carrier of social meaning. As shown in the previous subsections, there are language domains that may indicate group membership and social background. Pronouncing a sentence with a certain intonation may reveal one's age range, and one's lexical repertoire may mark their schooling level. To some degree, that is what we desire to achieve with the research described here. In a nutshell, the examples given so far suggest that the variable forms in language have social meanings. That is to say that social meaning indicates relations between linguistic and social structures that permit speakers to perform their identities and convey their stances (Campbell-Kibler, 2010).

It is noteworthy that having social meaning does not take out the referential and representational functions of language (Myers-Scotton, 2000). With that being said, language receives another layer besides the grammatical, which is the social domain. Figure 2.1, adapted from Thomas and Cook (2005), displays how Sociolinguistics would be inserted into Linguistics in general. This diagram is didactic because it puts Sociolinguistics in the outer circle, indicating that the inner circle areas can be explored by the largest one. Hence, sociolinguistics can look at different linguistic levels: from structural to textual-discursive layers, making it an interdisciplinary area.

To illustrate, the so-called "Valley Girl" accent[3] has a particular set of phonetic features (Nycum, 2018). These traits are so intricate that whoever recognizes this accent thinks, almost automatically, that that individual must come from the South of California. In this case, a person's accent might indicate where they come from, in other words, it gives a hint of a sociolinguistic factor.

Figure 2.1: Linguistics Circles



Adapted from Thomas and Cook (2005)

Community as a locus of research. Sociolinguistics has dealt with the language used among and within

[3] If you are not familiar with the "Valley Girl" accent, check out this short video to know more: https://youtu.be/n_WM-Af5svs

communities. This lies in the fact that not only abstract categories (age and gender, for example) may be a group's feature but also their shared semiotic activities (Eckert, 2006).

In Sociolinguistics, the focus is on shared grammar, that is, what is linguistically common in a certain group that makes it different from another and what social factors are conveyed by these patterns. In addition, the manner people socially position themselves through language is foregrounded.

In Labovian Sociolinguistics, focusing on the community means studying the language used by its members as well as analyzing the sociolinguistic norms that affect their speech (Eckert, 2006). While this statement emphasizes only the group language, it is important to recall that each community has its routine and rules, which are to some extent portrayed in language.

In conclusion, this section has attempted to provide a brief summary of Sociolinguistics as a discipline. Now, the next section will bring an outline of the history of Sociolinguistics and how it has advanced so far.

2.3 A background on Sociolinguistics

Sociolinguistics has become a popular area since Labov's contributions. Several conferences and journals[4] have already spotlighted the area. However, Koerner (1991) observations regarding that linguists underreflect upon the origins and history of their field of research are still true. According to this author, a research agenda can only be perceived as a mature field when its practitioners are familiar with its trajectory. Hence, it is important to bring into discussion a summary of the history of sociolinguistics.[5].

Labov has received the title as the precursor of Sociolinguistics, even though other linguists had contributed to the discussion about language and society previously to Labov. For instance, in the early 1800s, William Dwight Whitney, a skillful lexicographer, claimed that speech is a social and not an individual possession (Alter, 2005). Moreover, Whitney started the conversation about social factors and language use, when he found out that lower-class speech had influenced the upper-class language. Consequently, **one** of the ways that Whitney viewed language was as a concrete social institution like any other, such as science, religion, or law (Lacerda, 2021).

Another significant part of Sociolinguistics history concerns dialect studies. Going against the neogrammarians, scholars believed in the inseparable link between culture and language (Lacerda, 2021). These language researchers also further elaborated the fieldwork to establish dialectal-geographical regions and maps (Koerner, 1991). In this pivotal area, it is possible to perceive that the methodology used by sociolinguists nowadays (which will be discussed further ahead in this section) is a repercussion of this type of study.

Jumping to the 1900s, Ferdinand Saussure, considered as "the father of modern Linguistics", made a clear distinction between what is individual and what is social through one of the famous dichotomies: *langue vs. parole*. The former is the representation of one's thoughts and reasoning and therefore, it is personal and singular; whereas the latter is the abstract system available for its speakers (Lacerda, 2021). Despite the fact that language variation is so dear to Sociolinguistics, Saussure argued that Linguistics should focus on the formal explanation of language, disconnected from the cultural and social aspects (Cabral, 2014). Thus, at the time, the social domain was disregarded as part of linguistic studies.

[4] Access the following links to check numerous conferences and journals related to the area: https://conferenceindex.org/conferences/sociolinguistics (accessed on January 2nd, 2023) and https://onlinelibrary.wiley.com/journal/14679841 (accessed on January 2nd, 2023)

[5] As a personal note, I do believe that in-service researchers should know at least a bit about their field history not only to be aware of where it came from but also to foresee where it can go.

Nevertheless, the social element had arisen again through Saussure's apprentice, Antonie Meillet. He recognized that social conditions shape language as well as such conditions are the reason for linguistic change (Lacerda, 2021). Meillet proposed that language is not autonomous, but a dynamic system in which social facts may have an effect. Although this belief is highly aligned with Sociolinguistics, it stayed at the "academic margins" because of the prevalent vision at the time, which was Saussure's structuralism.

In the case of generative linguistics, the aim was to describe the innate aspect of language through logical and mathematical symbols (Lacerda, 2021). Hence, according to Mesthrie (2011), Chomskyan Linguistics sought to describe the essence of language from a mentalist perspective in which researchers would project an ideal speaker in order to capture the human language capacity. Taking that into account, processes such as socialization, cultural exchange, and nation contact would not be important variables in language studies. However, there is enough evidence that these actions do affect language (for example, check Mayr et al. (2020) who describe language contact in a bilingual context).

Turning now to Labovian Sociolinguistics, Labov highlighted the importance of collecting authentic data (Gordon, 2006), which was left aside in the early stages of generative linguistics. In addition, in this perspective, the speaker not only understands language structure but also computes when and how to use different forms through their competence (Cabral, 2014). Thus, as discussed in section 2.2, language is a heterogeneous system, that can carry social meaning, and can consistently group individuals.

Up to this point, a concise summary of Sociolinguistics background was given. Yet, there is a question: how is modern Sociolinguistics today? What are the issues and solutions regarding methodology within the area? The answers to these questions will be presented in the next subsection.

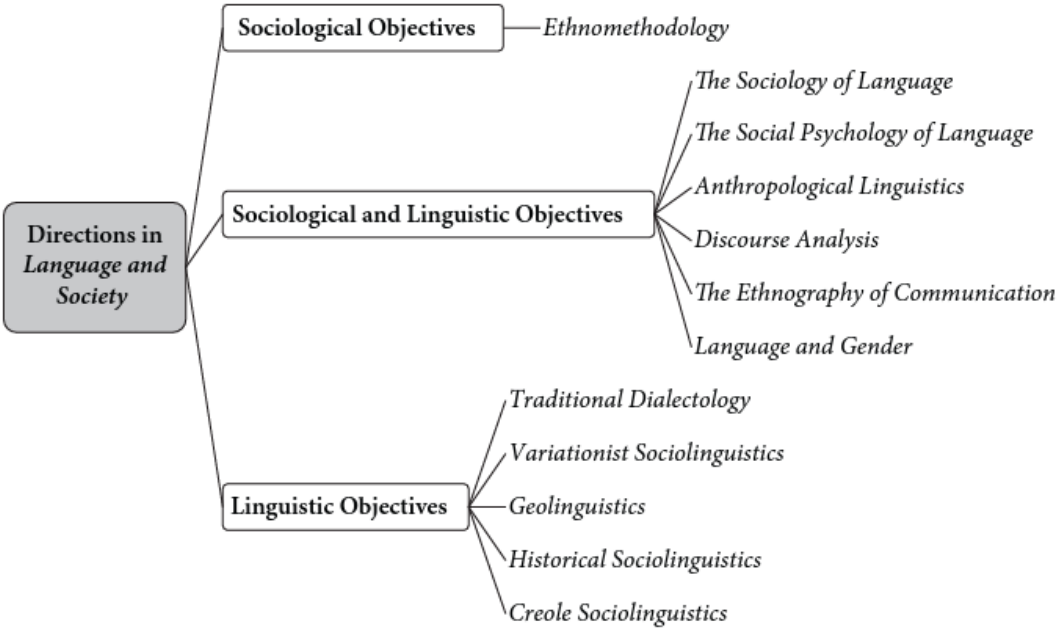
2.3.1 What about the recent stages of Sociolinguistics?

As previously discussed, modern Sociolinguistics had multiple influences, including the areas of Anthropology and Sociology (Shuy, 1990). Of course, this would generate different methods of reporting the relation between language and society. Moreover, depending on the study goals, the expertise area may change. In Figure 2.2 (Hernández-Campoy, 2014), the different perspectives are shown according to their aims. As can be seen, Sociolinguistics is located in the "linguistic objectives" branch. Nonetheless, those areas can complement and, sometimes, overlap with each other. The following part of the text moves on to describe the variety of methodological visions in Sociolinguistics. It is important to highlight that several methods have already been applied in Sociolinguistics, and this subsection does not intend to explain all of them in detail due to time and space restrictions.

Firstly, one of the most famous methods is **ethnographic fieldwork**. It requires the researcher to be *in loco* and it is qualitative since the analyst needs to intensively engage in the social setting (Hernández-Campoy, 2014). Furthermore, in this type of research design, the analyst is interested in participant observation, in which they can notice the community nuances regarding language (Hernández-Campoy, 2014). As Levon (2013, p. 74) pointed out, *"while your goal as a researcher is to become as much of an 'insider' as possible, it is important to realize that most people are not accustomed to having someone observe and comment on what they do"*. Therefore, that can present as an issue, once the participants may not consider the ethnographer as a, at least temporary, member of the community.

To illustrate this method, Sabaté-Dalmau (2016) examined the application of multilingualism in a Spanish university and gauged students' reactions through extensive fieldwork and data collection. She found out that students had both positive and negative attitudes towards such a process. Some of them faced the implementation of English language courses as a way to improve their chances for employment, whereas others claimed that it would make it harder for the survival of minority languages. Such conclusions were possible due to, among other aspects, the intense participant-observation she performed.

Figure 2.2: Areas of Language and Society



From Hernández-Campoy (2014, p.11)

Another well-spread method in Sociolinguistics is the **sociolinguistic survey**. This type of design is basically a multiple-choice test in which the informants choose the option they think best answers the question or proposition. Unlike ethnographic fieldwork, the sociolinguistic questionnaire allows the researcher to gather more data and do quantitative analysis. Notwithstanding, the questionnaire method is not free from having issues. Cooper (1980) and Lieberson (1980) revealed that surveys can have problems in sampling techniques and execution. They went further and explained that these complications may happen owing to the lack of training in survey design.

A useful example of sociolinguistic survey work is Correia and Flores (2021), in which the authors analyzed how children develop their languages in a bilingual setting, more specifically, when there is a heritage language. By using a questionnaire, they were able to create a survey in which researchers could investigate the linguistic experience in children's contexts by asking about the family history, the schooling trajectory, and the exposition of the heritage language. For the intended goal, the sociolinguistic survey was a plausible approach, since they would not gather (social) metadata by recording children's interaction, for instance.

Additionally, **sociolinguistic interviews** are also used in linguistic research. As the name indicates, the researcher or their team interviews the aimed community sample in order to investigate a set of linguistic phenomena. Such interactions can be a structured conversation, an elicited chat, or free communication

(Hernández-Campoy, 2014). One of the issues that may arise while performing this method is that, although the linguist seeks to have authentic data, the interviewee may not feel relaxed enough and can become self-conscious of their language.

Furthermore, Koven (2011) demonstrated that interview stories can have some level of complexity. The author compared interviews with natural conversations and defended that interviews can be used, for instance, for narrative analysis. As can be seen, interviews give new perspectives on language studies, since researchers can either narrow down the topic and the structure or let it be unrestrained, although natural conversations can provide more insights into sociolinguistic phenomena.

Although those methods can provide valuable conclusions about language use, there are a few issues that are not related to the methods themselves but to the practices done by sociolinguists, especially in Brazil. Some of these issues are: (i) data collection and extraction; (ii) data availability; and (iii) under-usage of quantitative analysis (Kendall, 2011; Szmrecsanyi, 2017). The following paragraphs will indicate directions and real examples to overcome those issues, instead of blaming sociolinguists.

Firstly, there is a tendency in Sociolinguistics for researchers to collect a small amount of data (i.g. a couple of hours of an interview) or hundreds of interactions as a result of years of fieldwork (Kendall, 2011). Moreover, only the piece of language considered useful for the study at hand is described, so the rest of the data is somewhat disregarded (Kendall, 2007). This causes two major problems: **bias in research** because the linguist will look only into the data they have transcribed; and **waste of rich data** since only the data for their research is actually available and used. Nonetheless, there are some efforts to reduce these problems. For example, the North Carolina Sociolinguistic Archive and Analysis Project[6] (Kendall, 2007) is a corpus containing transcriptions of whole interactions in addition to the acoustic information. Endeavors like this allow the accessibility of more data and, consequently, lead the way to more research.

From another perspective, sociolinguistic data is usually "locked" in their compiler linguist's archive. That is, the data is collected, transcribed, and analyzed but it is not available for peers to reproduce the study or to further analyze it. This may implicate in "academic bubbles" since only the people involved in the compilation process are able to examine the data. In addition, it may cause a delay in science once other linguists are unable to investigate it. Even so, corpus linguistics (CL) has advanced in a way that corpora with sociolinguistic parameters and theoretical basis are now available, making open-access and machine-readable sociolinguistic data. Some examples of such corpora are[7]:

1. **the CallFriend multilingual corpus**[8]: a phone call archive with metadata of people's sex, age, and schooling time, which is available in American English (Southern and Northern), Canadian French, German, Japanese, Spanish (from Spain and Caribbean) and Mandarin (from Taiwan and Mainland China) (Liberman and Cieri, 1998).

2. **the ESLORA - Corpus for the study of oral Spanish**[9]: a European Spanish corpus composed of semi-structured interviews and spontaneous speech with metadata of participants' age, sex, schooling level, and role (Barcala et al., 2018).

The methods previously listed *(ethnographic fieldwork, sociolinguistic survey, and sociolinguistic interviews)* do not presuppose quantitative analysis since statistics was neglected by linguists because of lack of training, fear, or dislike (Cantos Gómez, 2002). In Sociolinguistics, this has been materialized with over-

[6] https://bit.ly/3s9sCuN

[7] I would like to show my gratitude to the CORPORA mail list, which archived the corpora links in an email thread (https://bit.ly/3OXk1Fe) and also has shared several news and opportunities with linguists for more than 30 years.

[8] https://ca.talkbank.org/access/CallFriend/

[9] http://eslora.usc.es/

qualitative[10] studies and simplification of quantitative analysis, for instance relying merely on mean and median (for example, check Dutra and Simioni (2020), and Ataíde et al. (2021)). Corpus and computational linguistics have developed enough to show that these and other simple statistical measures, although they are important, offer little to get to conclusions.

According to such issues, there is an urgent need, which has been addressed in recent years, for statistical and computational analysis in Sociolinguistics. Grafmiller and Szmrecsanyi (2018), for example, used a series of distance metrics and predictive modeling procedures to analyze particle placement in nine varieties of World Englishes. One of their findings was that there are differences in particle placement use related to the status of the language in the country (as a native language or as an additional language). The analysis they did and the level of granularity and precision they presented the results would not be as it is without the computational methods.

An early attempt to insert statistical methods into Sociolinguistics was through the Varbrul methodology (Guy and Zilles, 2007). According to the authors, *"Varbrul is a set of computational software of multivariate analysis, specifically structured to accommodate sociolinguistic data"*[11] (p. 105). That is, it allows simultaneous analysis of several variables, and it can deal with texts commonly compiled in sociolinguistic research. On the other hand, one can argue that Varbrul makes it hard to compare interactions among factor groups since Varbrul is a fixed-effect model and it can attribute magnitude to a less significant variable. Consequently, that may lead to false conclusions. Nevertheless, Varbrul was a milestone in Sociolinguistics, because it introduced a brand-new way of doing Sociolinguistics research. Moreover, several studies used it and demonstrated relevant results[12].

Moving on now to consider the fruitful symbiosis between Sociolinguistics and CL, in recent years, there has been an increasing trend to building corpora and incorporating them into linguistic analysis (Tagnin, 2018), and with Sociolinguistics it would not be any different. Szmrecsanyi (2017) defends that, although Variationist Sociolinguistics usually deals with unconventional corpora (i.e., sociolinguistic interviews), it is questionably based on corpora[13]. Yet, they can be designed and organized in a way that they are machine-readable and able to undergo corpus analysis.

Another point of view is that both CL and Sociolinguistics are consistently empirical (Szmrecsanyi, 2017). They both require authentic language use compilation. CL has focused more on written corpora, mainly because oral corpora demand more staff, equipment, budget, and reasoning about the ethics of the compilation. Differently, Sociolinguistics has concentrated on spoken texts, more specifically the vernacular variety (Kendall, 2011). Despite the data type difference, interdisciplinary research and connections between these two fields is not only possible but also desirable.

Alongside these lines, there are speech corpora that can be used in both CL and Sociolinguistics. Although Kendall (2011) mentions that standard corpora do not frequently have their contextual information, this is not the case for C-ORAL-BRASIL I (Raso and Mello, 2012), the corpus used in the present study, which contains the topic, the general context and the participants' sociolinguistic information for each recording. Although this specific corpus has not been built having only sociolinguistics in mind, it has a serious compilation that makes it usable for such a field. Thus, efforts have been made and still have to happen in order to combine principles from both areas and fulfill their demands.

Furthermore, Szmrecsanyi (2017) outlines three criteria for a study to be in what he called "Corpus-based Variationist Linguistics" (henceforth, CVL). His criteria states that CVL:

[10] Here we are not diminishing qualitative studies. They are as important as quantitative work.

[11] Original text: O Varbrul é um conjunto de programas computacionais de análise multivariada, especificamente estruturado para acomodar dados de variação sociolinguística

[12] Check Young and Yandell (1999) for the use of Varbrul in Second Language Acquisition Studies; and Matus-Mendoza (2002) for using Varbrul in Sociolinguistic Lexicology.

[13] Whether or not these text databases follow an "adequate" definition of corpora (Sardinha, 2000) is up to discussion.

(i) sheds light on how different people say the same thing in various ways.

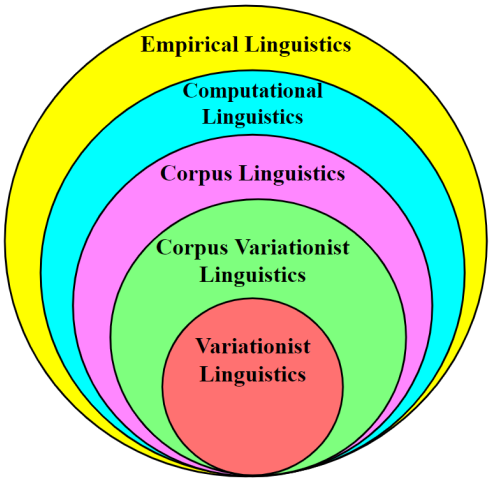
(ii) focuses only on the linguistic decision-making processes.

(iii) has a strong quantitative and statistical basis to explore linguistic data.

It is possible to expand the term and add "Computational" to the name because of the current trend of combining Linguistics and Computer Science: Computational Variationist Linguistics, or putting it even broader, **Computational Sociolinguistics** (this last term will be further explained in subsection 2.3.2).

Adapted from Szmrecsanyi (2017), Figure 2.3[14] attempts to represent graphically how the overlap and in(ter)dependence of each Linguistics subfield are related to Empirical Linguistics and Sociolinguistic Studies (Kendall, 2011). It is important to highlight that the circle sizes do not display the subfield importance or some kind of strict hierarchy. However, not all variationist work is corpus-based, consequently, "Variationist Linguistics" could not be totally circumscribed in the "Corpus Variationist Linguistics". Likewise, the same principle applies to the other subfields. Additionally, Sociolinguistics is broader than "Variationist Linguistics", since the latter is one of its branches (Meyerhoff, 2006).

Figure 2.3: Linguistics subfields

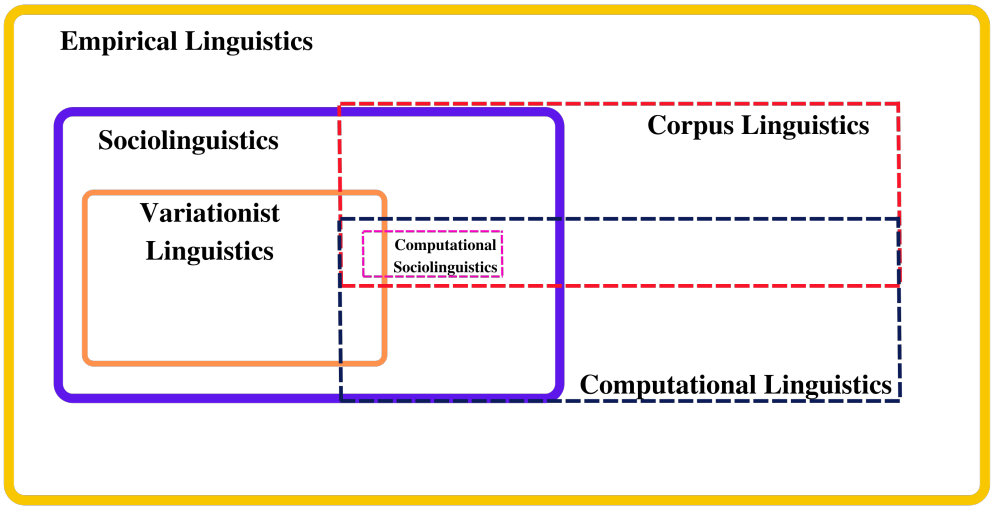


Adapted from Szmrecsanyi (2017)

Having that in mind, Figure 2.4 is an option to depict how these subfields correlate with each other. As demonstrated by Figure 2.3, Empirical Linguistics covers all of the fields shown, since all of them are concerned with authentic data. Moreover, Sociolinguistics is explicitly put as broader than Variationist Linguistics.

[14] We added Computational Linguistics.

Figure 2.4: New reading of Linguistics subfields



Elaborated by the author

Differently from Figure 2.3, in Figure 2.4 Corpus and Computational Linguistics are organized outside of the circumscribed areas, since they can be used by various fields as a methodological apparatus (Sardinha, 2000). Because of that, sociolinguistic work may or may not adopt computational or corpus techniques. Owing to this flexibility, these fields are represented with dotted lines.

Additionally, CompSoc is displayed in the figure as a combination of various Linguistics areas since it is an interdisciplinary field (Nguyen, 2017). Subsection 2.3.2 will discuss the intricacies of this area in more detail, but what is important for now is that CompSoc is also fluid in the sense that different objects of analysis and researcher's preferences influence on which approaches will be used.

This section reviewed some of the contemporary methodological and epistemological discussions around Sociolinguistics. In the next subsection, the subfield of Computational Sociolinguistics will be explored in addition to how its newness is changing the way sociolinguistic work is done.

2.3.2 Defining Computational Sociolinguistics

So far this chapter has discussed several aspects of Sociolinguistics. So, why is there a specific subsection for CompSoc? First of all, although technology and statistics have been used in Linguistics, only in recent years they have become part of Linguistics' "good practices", especially in areas in which researchers are interested in causality or correlation of variables. Therefore, CompSoc is still in its early stages of development and practitioner formation.

Moreover, unfortunately, the research to date has been made almost entirely[15] by Anglo-Saxon and European researchers. Considering the Brazilian context, for instance, the circumstances are unsatisfactory,

[15] Note here that we are **not** criticizing Anglo-Saxon and European science.

being Mello (2023) the first study to articulate this field in Brazil. There are some efforts from universities especially in the southeast of the country to integrate corpus and computational methods into Sociolinguistics and Linguistics in general[16] but the Brazilian Linguistics community has not extensively applied them in their research, when possible. Thus, since CompSoc might be a brand-new topic for the Brazilian community, this subsection tries to introduce it to them as well as to be a way of reaffirming quantitative analysis in Linguistics.

One of the most important aspects of CompSoc is **data collection** (Nguyen, 2017). It is possible to argue that this concern came from the data compilation caution given by CL, since it aims at compiling a great number of texts to represent a language variety (Sardinha, 2000). As a result, the concern is on how data is available in the "wilderness" and what criteria should be taken to form the study sample. One may say that this interest in data collection should be done in all Linguistics fields but CompSoc and CL as a whole take a step further and link themselves to a computer in order to quantify data distribution and execute accurate statistical calculations (Kennedy, 1998). All this without losing sight of the goal of linguistic description.

Likewise, with the advancement of technology, not only the access and storage of data are easily carried out but also the number of tools increased. For instance, SpaCy[17] (Honnibal and Montani, 2017) and Stanza[18] (Qi et al., 2020), which bring updated NLP models, are capable of annotating and describing linguistic data with a high level of efficiency. Therefore, such tools can exponentially enhance sociolinguistic analysis, since it saves us time, money, and workload while dealing with corpora.

Still related to data collection, according to Nguyen (2017), one of the main data sources of CompSoc has been social media website posts. For instance, tweets and Facebook posts are efficiently compiled with the help of web scraping algorithms. Hence, more communicative situations and sociolectal behavior can be described as well as other facets of a certain group can be investigated. For example, a comparison among football fans in computer-mediated and in-person interactions.

As previously discussed in subsection 2.3.1, Sociolinguistics has a lot to offer to CL in terms of types of data in addition to learning from it regarding corpus compilation techniques. However, social media-based corpora and other sociolinguistic corpora raise the question of the ethical implications of such databases. Kendall (2013) alerts that linguists should think deeply about data ownership, copyright, and ethics behind data gathering. He complements by saying that since the beginning of the research planning, these elements must be taken into account. Being accountable for these issues can help preserve the data in the long term and avoid lawful-related complications. As can be seen, data collection is not just a simple phase of research, instead, it is highly important and precious to CompSoc.

From another perspective, the **implementation of computational methods** is also paramount for CompSoc. The thorough work of Nguyen et al. (2016) claims that one of the reasons why computational methods have been implemented in Sociolinguistics is the availability of computer-mediated texts as they are easily extracted from the web and can have some sociolinguistic information about its writer or speaker. For instance, in Szmrecsanyi et al. (2019)[19], the authors used two corpora: the International Corpus of English (Greenbaum, 1991) and the Corpus of Global Web-based English (Davies and Fuchs, 2015), the latter being a corpus of billions of words from online texts from 20 English-speaking countries. Thus, in the case of this paper, the sociolinguistic information the authors had was the country where the text was produced. Other examples would be various outstanding studies in the journal Frontiers in Artificial Intelligence[20] (volume 2, 2019) in which the research topic was CompSoc. One of them is Grieve et al. (2019), in which the au-

[16] Some examples are researchers Lívia Oushiro and Plínio Almeida Barbosa at the University of Campinas; and Ronald Beline Mendes at the University of São Paulo

[17] https://spacy.io/

[18] https://stanfordnlp.github.io/stanza/index.html

[19] This study will be later discussed in depth in the present paper

[20] https://www.frontiersin.org/research-topics/9580/computational-sociolinguistics#overview

thors compared British tweets (1.8 billion words), and a famous dialect survey to outline lexical variation in British English. Studies like these were only possible by means of computational techniques.

Moreover, computational procedures can help systematize oral spontaneous language in a way that would be nearly impossible if done by hand. To illustrate, Knight et al. (2020) reported the computational design of *CorCenCC corpus: The National Corpus of Contemporary Welsh*, in which there are several types of linguistic data, including spoken informal interactions. The authors also described the computational treatment the corpus underwent: pre-processing, storage, markup and annotation, and corpus search interface construction. Likewise, the study reported here is also an example of such potential: the scale and the depth would have been much less than it is if it had been done without Natural Language Processing (NLP) and statistical methods.

One of the outside-of-academia applications of CompSoc would be the creation of chatbots. For instance, Eliza[21] (Weizenbaum, 1966), the first documented chatbot, was designed to emulate a therapist. If there had been a well-built computational sociolinguistic work, alongside other areas, that analyzed therapists' language, Eliza's performance would have been much better than it actually was. Such analysis could incorporate, for example, how the same therapist talks to clients of different ages and genders. Another example would be SimSimi chatbot[22], which is a chatbot for small talk. According to its website, SimSimi can have emotional conversations with people, but the interaction would certainly be enhanced if a language model with sociolinguistic annotation[23] was implemented.

Another aspect is that computer scientists are becoming interested in social events. However, Computer Science alone cannot withstand the complexity of language and society (Nguyen et al., 2016); therefore, it needs the know-how and insights from Sociolinguistics. On the other hand, the authors claimed that they are in symbiosis. Sociolinguistics also needs Computer Science, particularly from a methodological standpoint, considering the implementation of quantitative analysis and computational modeling. Hence, departing from this two-way relationship, **CompSoc refers to the ground-breaking field of the intersection of Sociolinguistics and Computing that seeks to study the social aspects of language through a computational perspective** (Nguyen et al., 2016).

To help us understand how that definition works in research, let us take Morales Sánchez et al. (2022) as an example. The authors developed a "white-box" model to detect gender based on linguistic features, such as orthographic, morphological, lexical, syntactic, digital, and pragmatic ones. They were able to pinpoint the most relevant features in consonance with the decision trees produced. A very important aspect of this study is that a computational method is used to solve a Sociolinguistics problem[24]. Consequently, it is possible to boost sociolinguistic investigations with computational analysis.

Furthermore, the "cutting-edge" element of the field also means "new" and "in need of more work", which relates to the fact that there is no more adequate or perfect computational method for CompSoc research questions. As in any field, the method used depends not only on the observed phenomenon but also on the research design. On the other hand, since this emerging area is brand new, linguists are still experimenting to define what methods suit a specific problem better, that is, there is no established method such as in sociolinguistic interviews and surveys. However, that does not mean that the studies done in this area are unproductive or immature. In fact, what makes CompSoc so important right now is the urgency for its methodological and theoretical development.

Such progress is done mainly using social media data because it

> *has contributed to the insight that text can be considered as a data source that captures multiple*

[21] More information in https://web.njit.edu/~ronkowit/eliza.html

[22] https://simsimi.com/

[23] For example: separating in the model what was said or written by a child, a teenager, and an adult.

[24] I particularly think that this is in need. There should be more unity among computer scientists and linguists.

> *aspects and layers of human and social behavior. The recent focus on text as social data and the emergence of computational social science are likely to increase the interest within the computational linguistics community on sociolinguistic topics* (Nguyen et al., 2016, p. 578).

To illustrate, Ilbury (2020) analyzed orthographic variation in tweets of homosexual British men especially concerning linguistic clues of African American Vernacular English. The author used computational methods to extract data from the social media website, but they performed their analysis manually. According to the author, the interactional component of language would not be well-modeled by a computer. Whether or not they are correct, the point is that CompSoc (and Computational Linguistics in general) is still looking for methods that could answer its questions and that are able to convince the Linguistics community that computers can examine and model socially situated language use.

Another limitation is the amount of sociolinguistic information available in online linguistic data. With Twitter, for example, users can put their location and their birth date. Other information, such as sex, can be inferred by the user's name (Burghoorn et al., 2020). Nonetheless, the granularity and quantity of this information are not as large as those collected in sociolinguistic surveys. Once again, CompSoc urges and stimulates the compilation of large corpora with a design for sociolinguistic work, especially spoken language corpora.

Nevertheless, CompSoc is an innovative approach to studying language and social meaning. As discussed here, it is concerned with data collection and quantitative analysis in order to make robust and accurate conclusions about sociolinguistic phenomena. Because of that, CompSoc may be starting a fourth wave of Language Variation studies. There must be more studies on this area to examine if it is indeed the case of a new trend, an expansion, or an improvement of the three waves.

To sum up, this chapter began by describing the different trends in Sociolinguistics and arguing that, although they focus on distinct elements, they are complementary. Also, it went on to present a review of the history of how linguists thought of language and society. Following that part, the basic and most famous sociolinguistic methods were described. Finally, this chapter has shown the importance of combining quantitative analysis and Sociolinguistics and how that translated to the rise of a new field, Computational Sociolinguistics. The next chapter describes the synthesis and evaluation of the term "sociolect".

Chapter 3

An agenda for sociolects

"You shall know a person by the company they keep"
"Você conhecerá uma pessoa pelas pessoas que ela mantém por perto"

Rephrasing Firth's famous statement about the collocational property of words[1], one may assert that humans also have a "social quality". In Ancient Greece, Aristotle wrote that we are social animals (Aristotle, 2001). Although Aristotle needs an update, several studies have shown that he was quite correct[2]. If socialization indeed influences how we behave; thus, a few questions arise such as: *is language distinct enough between different groups? How can we measure these differences?*.

In order to answer the previous questions, what follows next is an account of the different names for "group languages"[3]. Moreover, the notion of "sociolect" will be explained as well as why it is more plausible than other terms. Finally, the chapter finishes with a brief description of the levels of language variation and how they affect each other.

3.1 Communities

As it was demonstrated in section 2.2, Sociolinguistics aims at describing how language items and processes are common in a group and how they make this group different from others. However, there has not been an agreement on the terminology that should be used to refer to a "group's language". Some researchers use "speech community", while others, using a more specific and narrowed community, adopt "community of practice". Another portion of linguists focuses on the language used per se and they prefer one of the lectal terms: ethnolect, sociolect, or social dialect. In this subsection, a description of these terms will be presented alongside the discussion of the pros and cons of operating with such words.

Let us start with **speech community**, which is a dear concept to Sociolinguistics. According to Morgan (2004), a speech community is a group of people in which their *"language represents, embodies, constructs, and constitutes meaningful participation in a society and culture"* (p. 3). It is possible to conclude that the speech community is, thus, purely related to linguistic features.

[1] Original statement: you shall know a word by the company it keeps (Firth, 1957)

[2] Check Guhin et al. (2021) for a methodological discussion; Wang et al. (2020) for ethnic-racial implications of socialization; and Pardede and Ramadia (2021) for a medical standpoint.

[3] This term is used here as a general notion for socially situated language use. The proper terms will be approached later on in this chapter.

For instance, Bustan et al. (2021) found out that the speech community[4] of Manggarai native speakers in the region of Ruteng (Indonesia) consistently used plant metaphors in their daily language. Furthermore, the authors proposed that this regular use of metaphors with plants is closely associated with the lifestyle of the people of that community. From another perspective, Nance (2022) analyzed sound changes in the Scottish Gaelic speech community. The author discovered that the community may be changing the way they pronounce the lateral sounds due to generational differences (mainly the age gap). These studies are examples of how the diversity of linguistic phenomena makes membership vast.

Because of that, determining a group based merely on their language is somewhat reductionist as pointed out by Wardhaugh (2006), because there may not be a direct correspondence between a set of linguistic features and a given community. Gumperz (2009) claims that people assemble with those who have a shared set of linguistic items in common but that does not account for other membership reasons. A person can participate in a group firstly by personal tastes (arts and entertainment in general), while another can join a group because of their occupation. Although these groups may have a specific language use, that is not the only and main reason for their association.

Moreover, globalization impacts the applicability of the speech community notion. The overemphasis on "sharing" in the speech community fails to deal with situations in which shared knowledge is no longer considered (Jacquemet, 2019). As argued by the author, more recent investigations of language use have to take into account the deterritorialization of communicative practices and social formations, since there has been more mobility of goods, services, information, and people from different backgrounds[5] (Bloom, 2004).

Still on the "sharing" element, a linguist cannot analyze a certain community presupposing that there is a shared norm for every group. Bucholtz (1999) explained that "sharedness" is more linked to a researcher's interest than an intrinsic feature of a community's language. For example, Belo Horizonte[6] cannot be called a speech community (Wardhaugh, 2006). It is too big and plural. One may say that it is a pack of smaller speech communities. Even so, the level of granularity to create criteria that would separate its inhabitants in speech communities would be highly difficult because there are geographical, social, religious, occupational, ethnical, and age factors in play.

In order to avoid using speech community and have only linguistic parameters of membership, some sociolinguists use the term **community of practice** (Lave and Wenger, 1991). A community of practice is an assembly of people who have mutual participation in an endeavor (Eckert and McConnell-Ginet, 1992). Wenger et al. (2002) expand the definition and state that a community of practice is *"a group of people who shares a concern, a set of problems, or a passion about a topic, and who deepens their knowledge and expertise in this area by interacting on an ongoing basis"* (p. 19). This phrase was created by the impact of European authors, such as Foucault and Bourdieu, who considered speech variance as well as semiotic[7] comprehension, power relations, and ideology (Jacquemet, 2019). Because of that, a community of practice is composed of, at the same time, its memberships and the actions taken by its members (Eckert and McConnell-Ginet, 1992).

The term "community of practice" has been extensively used in language learning and business papers. For example, Gray (2005) investigated how informal learning could happen in an online community of practice of Adult Learning Councils. The authors used this term because the professionals involved had similar routines, issues, and ways of thinking at the workplace. From another perspective, Li et al. (2009) did a review of how communities of practice were reported in health and education papers. They identified

[4] This term is used here because the authors explicitly used it.

[5] And perhaps unrelatable as well.

[6] The capital of state of Minas Gerais.

[7] Here we are considering semiotics as a general account for the theory of codes and sign production, in other words, any kind of interactional behavior (Eco, 1979).

that the main features of a community of practice are social interaction among members, knowledge sharing, knowledge creation, and identity building.

At first sight, it seems that the term community of practice solves the problem of using speech community since it does not focus only on linguistic parameters to determine a community (Eckert and McConnell-Ginet, 1992). However, it still builds up on the concept of community, which is again associated with shared knowledge, mutual goal-oriented engagement, and elective membership (Jacquemet, 2019). Moreover, both notions are more linked to the community itself than to the language used by the group members. While "speech community" centers on a stable linguistic community, "community of practice" concentrates on a semiotic group. Thus, sociolinguists had to find ways of overcoming these issues. An alternative is adopting the lectal terms, which, in turn, do not have a consistent definition as well. The next subsection will approach a brief description of these terms.

3.2 Lectal variation

There is a major hurdle that many sociolinguists encounter while seeking to name and describe various subsets of language, such as sociolect, dialect, (social) dialect, style, register, idiolect, and others. The terminological instability is mostly exhibited by the co-occurrence of such ideas (Lewandowski, 2010). As well stated by Zwicky and Zwicky (1982), *"anyone who wants to talk about the many varieties of a language is immediately faced with severe problems, the initial manifestations of which are largely terminological."* (p. 213). This imposes a challenge because sociolinguistic studies can be hard to compare considering their chosen terminology. Therefore, this subsection attempts to elucidate lectal variation: the definition of the "-lect" terms, how they may affect each other, and finally why sociolect is the most suitable term for the research presented here.

"Language is a dialect with an army and a navy"[8]. This is a famous statement commonly attributed to Max Weinreich[9] (Bright, 1997), which raises many issues. Based on this saying, the difference between language and dialect would be merely political, since "an army and a navy" is a metaphor for power (Maxwell, 2018). On the other hand, Wardhaugh (2006) considers language as a single linguistic communication system that encompasses a number of mutually intelligible variants, which, in turn, are dialects. However, there is no straightforward explanation for the difference between language and dialect, because they are a simple division for a phenomenon that is infinitely variable and complex (Haugen, 1966).

Nonetheless, some studies have used "dialect" as a language variety associated with a region, social group, or ethnicity (Wolfram, 2017). As an illustration, Wagner et al. (2014) described how children perceived dialectal variation. The variants they present as stimuli are bounded to specific places: Ohio (USA), Lancashire (Great Britain), and Maharashtra (India) Englishes. From another point of view, Cheshire et al. (2008) analyzed white British people's speech from inner and outer London. They linked dialectal variation to people's ethnicity and social groups. Although their results might seem promising, there is no common ground in defining dialect.

Another idea is "ethnolect" which is defined as a variety of languages that index speakers to an ethnic group (Clyne, 2000). However, one may consider that ethnicity is also a "gatherer" of a social group. An individual can assemble with peers from their workplace as well as from their ethnicity. For instance, an indigenous teacher can be part of (i) a teacher group in their school; while outside the institution, they engage more with (ii) people from their tribe. Therefore, ethnolect poses some issues: (i) may not be centered in

[8] Original text: *A shprakh iz a dialect mit an armey un flot*.

[9] Maxwell (2018) criticizes Weinreich's authorship and gives credit to a student who said that statement during a lecture.

only one ethnicity[10] but it is socially organized, whereas (ii) is both ethnically and socially situated.

One of the problems with ethnolect is that it conveys the idea that a person is conditioned only by their ethnicity. It puts ethnicity in a way that ignores the individual's agency (Carter and Fenton, 2010). A possible solution for this issue may be adopting a socially informed notion that can cover these situations without disregarding culture and ethnicity.

For that reason, other studies have used the term "social dialect"[11], which can be characterized as the language of a social group (Corder, 1971). For our purposes, a group is an assembly of at least two people who gather for a reason: social, religious, political, cultural, familial, and so on (Wardhaugh and Fuller, 2015). For example, Febriani and Jufrizal (2019) compared the differences in diction, stress, and pronunciation between employees and labourers[12] in Padang (Indonesia). In this case, there was a social practice involved (the work itself) and a much smaller community (workers in a specific city). Moreover, they used sociolect as an abbreviation of social dialect. Again, there is the possibility of misconstruing this phrase, because it entails a different idea than a social group's language.

Although social dialect and sociolect are interpreted as the same language subset (Hudson, 1996), sociolect seems to be more accurate and free from preconceptions. The key problem is that using "social" as a modifier of dialect does not hide the assumptions we have about dialects. Dialects are commonly thought of as regional and national varieties (Hudson, 1996; Wolfram, 2017). On the other hand, the word "sociolect" gives the idea of a social group language, since its morphology is simple: socio- (society, social) and -lect (language subset or variety).

As asserted by Lewandowski (2010, p. 61), sociolect is *"the language spoken by a particular social group, class or subculture, whose determinants include such parameters as gender, age, occupation, and possibly a few others."*. This definition stresses the social aspect of language, but it does not state that it must be a stable variety, in contrast with speech community. Moreover, sociolect focuses on language use, i.e. linguistic materialization, which is the center of Sociolinguistics, whereas the notions of speech community and community of practice concentrate on the membership itself. According to Monteiro (2002), sociolect is a set of linguistic traits that are used preferably by a particular layer or social community. He asserts that the more complex the community, the stronger the social stratification and, consequently, the greater the difference in language use between strata. Consequently, sociolect is one of the indicators of the stratification of society.

From another point of view, sociolect allows us to explore the diasystem (Weinreich, 1954) in a community's speech. Based on Weinreich (1954), a diasystem can be defined as a *"higher-level structural feature shared by some dialects, also used metonymically to refer to dialects sharing such a structural feature, and/or to dialects in close contact."* (Vogelaer, 2013). If this concept is expanded to include sociolects, a speaker's various communities and the different ways individuals speak in each group are explained. For instance, the speaker bfamcv01LEO in C-ORAL-BRASIL I (Raso and Mello, 2012) is a male young adult undergraduate student; thus, he may use/activate one sociolect to speak with his peers during class and another to talk to his family at home. However, both sociolects share features mainly because they are based on the same language, PT-BR. Another example is Tavares (2008), which demonstrated that the use of "e, aí, daí, então" (Portuguese linking words) is similar in communities of adults and teens[13] in two capitals of different Brazilian states. This happens because *"speakers are differentiated by gender and age, by education, status or class, by ethnicity, by geography (neighborhood, place, rural vs. urban, etc.), by 'localness' (long established vs. new arrivals), and so on."* (Guy and Hinskens, 2016, p. 3). Thus, a sociolectal diasystem can

[10] There may be Asian, Caucasian, African, and Latino teachers.

[11] Douglas Biber, one of the most famous corpus linguists, uses social dialect in Biber (2019); Finegan and Biber (1994).

[12] A person doing unskilled manual work for wages.

[13] The variables in this study are: age (15 to 50+), sex (male and female), and schooling (less or more literate by years of schooling).

clarify which and how some sociolects interact.

There have been some misrepresentations about sociolects despite the rich insights its investigation may bring. Reynolds et al. (2013) presented a method for detecting sociolects in addition to defending that sociolects are a set of lexical items used by a group that shares similar interests. This specific paper was not authored by linguists and, consequently, the definition is misleading[14]. In fact, sociolect is an arrangement of language structures, which include phonetic, lexicogrammatical, and discursive aspects (Finegan and Biber, 1994), associated with a group (Wolfram, 2004).

A common misunderstanding about social dialect or sociolect is that all members of one social group use the same structures, but members of other social groups never do (Wolfram, 2004). This group-exclusive approach does not properly consider some processes that happen frequently: affiliation, disengagement, and merge, to name a few. People from a social group can join another group, detach themselves from their original group, or, for any reason, two or more groups can combine. In these situations, what Pijpops and Van de Velde (2018) called **lectal contamination** may happen, which is lexically-specific preferences[15] that may emerge as a result of language interaction with another variety that shares the same construction. Since it is a quite recent notion, more research is needed to provide more evidence and go further in the discussion about "contamination".

Another process related to the sociolect and "-lect" terms, in general, is **lectal coherence**, which is defined as the systematic co-variation of variable linguistic elements that share a social attribute (Guy, 2013). In other words, sociolect does not need to be a set of static components, in fact, it is a network of co-occurrent elements. For example, *"speakers can tend towards the high-status end of the social spectrum for one variable, while simultaneously displaying relatively lower status usage of another."* (Guy, 2013, p. 70). This author goes further and compares sociolect to language: language is an array of linguistic particularities co-occurring together used by several people, likewise, sociolect is traceable by the simultaneous co-occurrence of lexical, phonological, morphosyntactic, semantic, and pragmatic features. Considering that, a corpus linguistics approach is suitable for sociolect description since one of the main principles of CL is the co-occurrence of linguistic elements (Gries and Durrant, 2020). *"That is, a linguistic expression E– a morpheme, word, construction/pattern, ...– can be studied by exploring what is co-occurring with E and how often."* (Gries and Durrant, 2020, p. 142).

Still on lectal coherence, Guy and Hinskens (2016) complemented that, just as languages and dialects are clusters of grammars and lexicon, sociolects, and styles are sets of linguistic variables. In this same work, the authors asked the question: *"Do such sets in fact cluster and co-occur, and if they do not, in what sense do the specific varieties exist?"* (Guy and Hinskens, 2016, p. 2). For this reason, a sociolinguist's role is to reveal which and how variables cluster together. Not coincidentally, the research reported here explores this question in the sense that we aimed to provide sociolinguistic profiling describing which variables co-occur with others.

Even so, it is fundamental that we bear in mind that speakers are operating agents, acknowledging their active role in shaping language use through speaking and writing. They calculate and control different variables simultaneously. For example, an individual can use language to manage group membership, situational purposes, identity construction, management of social relationships, and whatnot at the same time (Finegan and Biber, 1994; Guy, 2013). For this reason, any account of sociolects must consider the speakers' agency.

Because of that, register and sociolect are language subsets that can affect each other or overlap. Finegan and Biber (1994) argue that register is the entire range of linguistic variation that is linked to changes in communicative situations, including shared context, mode, and purpose. A similar interpretation was

[14] Here we are not belittling the authors' work. We are showing the importance of a linguist while dealing with language, especially in the hot area of Language Technology and Natural Language Processing.

[15] I believe that other language spheres are influenced by lectal contamination but there is room for research to check that.

proposed by Halliday (1978) who claims that register is dependent on "what you are doing at the time" and (social) dialect is a function of "who you are". With that being said, register is conditioned by the context and sociolect is reliant on a person's identity. Finegan and Biber (1994) went further and explained that the systematic patterns of register variation are functionally motivated, while sociolect variation displays systematic patterns for the same linguistic elements because speakers engage in different series of registers. As a result, the same functional patterns that drive register variation also drive sociolectal variation (Finegan and Biber, 1994).

Table 3.1 summarizes the different focal points among the sociolinguistic units highlighted so far.

Table 3.1: Different focuses among the sociolinguistic unit terms

Language subset	**Focuses on**
Speech community	a linguistically stable group of people
Community of practice	a group of people who share regular semiotic practices
Ethnolect	the language used by people with the same ethnicity
Sociolect	the language used by people with the same social background or in the same social group
Register	the language used by people in a specific situation for specific purposes

Elaborated by the author

As can be seen from Table 3.1, sociolect is language-based and socially situated. The work reported here intends to describe the similarities and differences, if any, between people's languages according to their social background. Thus, if the language in use is at the core and the utterances are gathered based on the speakers' social factors, sociolect is the most adequate theoretical notion.

Although sociolects are socially relevant and salient, they still depend on other systems of communication in order to fully function (Yadlovska, 2022). First of all, sociolect is highly influenced by the immediate national language with which the speakers interact, for instance, the sociolect spoken by young people in Belo Horizonte is going to be PT-BR-based. Moreover, sociolects exhibit features of speech, which can be expressed by words, phraseological constructions, syntax, and pronunciation, to name a few (Yadlovska, 2022).

Furthermore, the term "sociolect" is general enough to cover "ethnolect" without losing its cultural implications. Additionally, there is no requirement for shared knowledge and stability within the group language, in the sense that a social group has an internal diversity (i.e., the members). Despite having a within-group heterogeneity, well-established sociolects are not chaotic; however, there may be some chaos when two or more sociolects are merging [16] (Trudgill, 2004; Mesthrie, 2008). Thus, the sociolinguist's role is to describe how the group behaves linguistically whether or not some coherence is present.

However, sociolects are not immune to criticism. As presented in Table 3.1, the phrase "with the same social background" can be interpreted as overly general[17]. Despite this critique, sociolect can maintain a balance between granularity and generalization. For example, one can say "Belo Horizonte women's sociolect" or "doctors' sociolect", adjusting the social background criteria based on the analytical scope.

In short, this chapter began by describing the concepts of "speech community" and "community of

[16] Check out Trudgill (1999) to see a discussion about lectal chaos before its systematization.

[17] I would like to thank Dr. Mayara Nicolau for pointing that out during the XIV SETED 2023.

practice" and arguing that, although important for Social Sciences, they do not accurately represent what sociolinguists analyze. It went on to suggest that "sociolect" is a more appropriate term while dealing with an in-group's language description, mainly because it is a language subset and not an abstract representation of a community. In the chapter that follows, the main underpinnings of the Language into Act Theory and how they contribute to speech studies will be explored.

Chapter 4

Utterances, speech, and actionality

"Você estuda sociolinguística"
"You study sociolinguistics"

The preceding sequence of words could be uttered in different ways by the same person depending on the context and intention. For instance, it could be uttered as a question "Você estuda sociolinguística?"[1], as a declarative statement "Você estuda sociolinguística."[2] or as a skeptical statement "Você? Estuda sociolinguística?"[3]. In speech, what perceptually determines the speech act (question, declaration, skepticism, and many others) is above all prosodic features (Teixeira et al., 2018). Cresti (2000) was the first that construed a framework based on prosody after extensive research on spontaneous speech corpora. It also provided evidence that it is important to have prosodic segmentation into utterances, alignment between text and sound files, and informational unit annotation. In Cresti (2000) and Cresti and Moneglia (2005), this research scheme is named Language into Act Theory (henceforth L-AcT). The following is a brief description of the theoretical underpinnings that established the C-ORAL-BRASIL I design and architecture, as well as the reason why utterances were chosen as the basic unit for the research reported here.

L-AcT is a theoretical and methodological framework that highlights pragmatic and informational elements in language description through the compilation of spoken corpora (Cresti, 2000). Because of that, it has a foundation derived from the Speech Act Theory (Austin, 1962) for studying spontaneous speech. In light of this, L-AcT is an approach based on empirical data organized in high-standard spoken corpora.

This corpus-driven[4] approach was built upon the urgent necessity to propose a theory that accounts for the grammatical and interactional structure of speech based on oral communication per se and not a transposition of theories of written language. In addition, L-AcT would need to propose, through large datasets, what the basic unit of speech is, that is, the minimal portion of speech that has a complete communicative function (Izre'el et al., 2020).

One of the strongest premises in L-AcT is that speech is characterized by a succession of informational units prosodically marked. Moreover, this reference unit of speech presents an illocutionary nature, in other words, it conveys a linguistic action (Cavalcante, 2020). From that, it is possible to argue that prosody carries the pragmatic meaning in spoken interactions (Raso, 2012a; Cresti et al., 2018; Cavalcante, 2020).

[1] You study sociolinguistics?
[2] You study sociolinguistics.
[3] You? Study sociolinguistics?
[4] A corpus-driven approach uses a corpus as a source of language theory, therefore, it does not use data to explore a pre-established theory or hypothesis (Biber, 2015).

L-AcT establishes the utterance as the basic unit of speech, which is delimited between terminal prosodic breaks (Raso, 2012a), in other words, the boundaries are intonation-based. These "border" areas (i.e. terminal breaks) are perceptually understood by speakers alongside an illocutionary value (Moneglia, 2011). L-AcT presupposes that the interlocutor interprets the terminal profile as an indication that an utterance and, consequently, a linguistic action, is fulfilled (Raso et al., 2007).

As the utterance counterpart (Moneglia, 2011; Rocha et al., 2011; Raso, 2012a) and trackable between terminal breaks, a speech act is the smallest language unit that can be pragmatically interpreted (Austin, 1962). Cresti (2018) divides speech act into three levels, similar to those described by Austin (1962). They are: perlocutive, illocutive, and locutive acts. The first act[5] is an ideational/emotive state as a response to an outside input and is manifested linguistically in an affective manner toward the addressee. The second act refers to the expression of the first act in terms of a particular language action schema, which is often defined as a pragmatic issuing. The third act concerns the linguistic islands that take the shape of phonetic and prosodic transmission channels, which are used to convey the internal action schema. These last elements are the output itself, which is addressed to the listener, and the pragmatically packed mental content, which is arranged in accordance with particular language competence. Table 4.1 presents a summary of these terms (Cavalcante, 2020).

Table 4.1: Acts in L-AcT

Act	**Meaning**
Perlocutive	the drive and desire to act linguistically
Illocutive	speaker's communicative goal with an utterance
Locutive	the linguistic material and its literal meaning

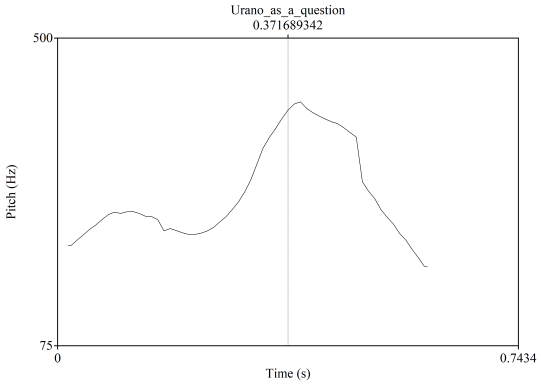
From Cavalcante (2020)

According to L-AcT, prosodic components in the utterance (e.g., pitch and emphasis) are systematically linked to the transmission of the illocution (Raso, 2012a; Cavalcante, 2020). For example, in Figure 4.1, the word "Urano" is pronounced with a question intonation, indicating, for instance, that the speaker is in doubt if they heard the correct word. On the other hand, Figure 4.2 displays a statement intonation with the same lexical item, acting as if the speaker is confirming what they heard. Therefore, there is a correspondence between prosodic elements and utterance interpretation.

a) Question — Extract 1 from bfamdl04[6]
Urano //

[5]Note that this definition differs from that written in Austin (1962)
[6]Listen to the audio in https://bit.ly/3tujXqE

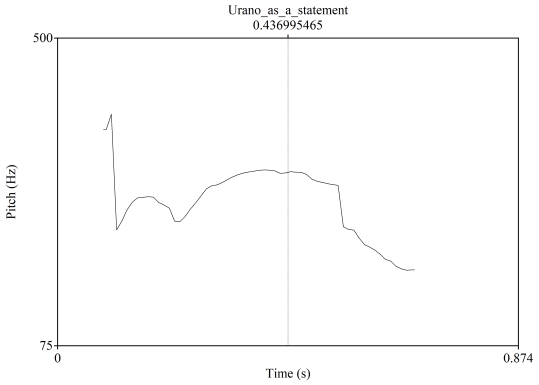
Figure 4.1: f0 contour in extract 1 from bfamdl04



Elaborated by the author

b) Statement — Extract 2 from bfamdl04[7]
Urano //

Figure 4.2: f0 contour in extract 2 from bfamdl04



Elaborated by the author

Departing from these issues, Cresti (2000) proposed four basic requirements to compile golden-standard spoken corpora. They are:

(i) **High diaphasic variation:** if the goal of the corpus is to have illocutionary variation, it is necessary to capture as many communicative situations as possible (Moneglia, 2011). The basic premise for it is that people act differently in various contexts.

[7] Listen to the audio in https://bit.ly/3RNv7zX

(ii) **Prosodic segmentation:** one of the central differences between writing and speech is how we convey meaning. One of the ways we do that in speech is through prosody. Therefore, a spoken corpus should have some kind of prosodic annotation or segmentation (Cresti, 2000).

(iii) **High acoustic quality:** this feature is more related to the technical side of corpus compilation. Many kinds of phonetic and phonological analysis may not be possible if the sound quality is not sharp (Hua et al., 2008; Raso et al., 2016).

(iv) **Text-sound alignment:** it enables practicality and speed in research as the linguist can hear and read the data at the same time. Moreover, it helps the researcher not depend solely on the transcription.

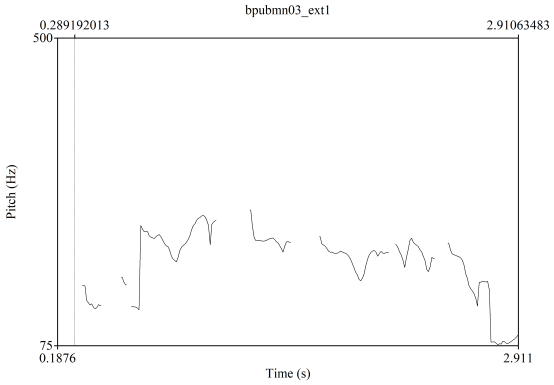
In the scope of L-AcT, a terminal break is transcribed with double slashes (//) and non-terminal breaks with only one slash (/). For example, Figure 4.3 represents an example of a declarative statement in PT-BR with more than one non-terminal break. The utterance in Figure 4.3 presents a rise in the pitch at the beginning of the utterance and a drop at the end. On the other hand, in Figure 4.4, there is a different melodic movement with an utterance with only a terminal break (a simple utterance), in which the speaker finishes the utterance in a relatively high pitch at the peak of the last word, construing a question.

a) Extract from bpubmn03[8]

*ANG: <são> dois pontos / também / que eu queria pegar //
*ANG: <are> two points / too / that I wanted-1st-S-PAST catch-INF //

Adapted translation: these are two points I'd like to cover as well

Figure 4.3: f0 contour in extract 1 from bfammn03



Elaborated by the author

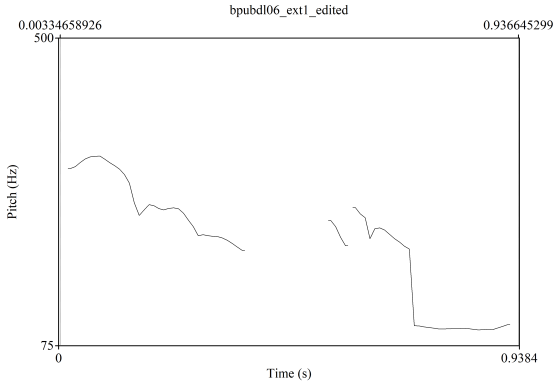
b) Extract from bpubdl06[9]

[8]Listen to the audio in http://bit.ly/49KP7Km
[9]Listen to the audio in https://bit.ly/3SLuJCY

**JAN: tem blusa <de frio> //*
*JAN: have shirt <of cold> //

Adapted translation: is there any jacket

Figure 4.4: f0 contour in extract 1 from bpubdl06



Elaborated by the author

L-AcT offers something desirable in CL, especially for spoken corpora, which is comparability (Mello, 2014). For instance, the C-ORAL-ROM (Cresti and Moneglia, 2005) is a collection of comparable speech corpora of some European Romance languages: Spanish, French, Italian, and Portuguese. These corpora underwent the same process of recording and compilation as well as had the same transcription criteria. For this reason, these corpora can be used in cross-lingual studies. For example, Cavalcante (2020) used the C-ORAL-ROM Italian and Portuguese corpora, the Brazilian Portuguese corpus (C-ORAL-BRASIL I - Raso and Mello (2012)), and the American English corpus (Cavalcante and Ramos, 2016) to analyze the prosodic features of a certain information unit in these languages. This analysis was possible as comparable spontaneous speech corpora allow investigations in several languages through the same standards. In addition, it allows an analysis to determine what is a property of speech in general and what is language-specific (Mello, 2014).

Furthermore, the way that the metadata transcription was done is helpful for linguistic studies. Some of the items registered in the metadata are the number of words in the transcription, comments, and sociolinguistic information. This can make the linguistic research have clearer and more helpful information on the original interaction context. Figure 4.5 is an example of a metadata file[10]. This current research would not be possible without the metadata description.

L-AcT, aligned with the principle of representativeness in CL (Biber, 1993), suggests that its corpora should have a large diaphasia spectrum (Raso, 2012b; Mello, 2014). As pointed out by the authors, compiling a corpus considering diaphasic variation (a.k.a. context variation) leads to diastratic variation (a.k.a. sociolinguistic variation), an important type of variation in a speech corpus. Moreover, diaphasic variation is fundamental, because it enables a speech corpus to cover several different speech acts and, consequently,

[10]The X's replaces the informants' first name. We decided to hide here to maintain anonymity.

Figure 4.5: Example of metadata file

@Title: Soccer championship
@File: bfamcv01
@Participants: LEO, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
GIL, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
LUI, XXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
EVN, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
@Date: 31/05/2009
@Place: one of the participants' home, Belo Horizonte/MG
@Situation: chat among four students who have organized a soccer championship, four clip-on microphones not hidden, non-participant researcher
@Topic: positive and negative aspects of the championship, plans for the next championship
@Source: C-ORAL-BRASIL
@Class: informal, family/private, conversation
@Length: 7' 00"
@Words: 1482
@Acoustic_quality: A
@Transcriber: Bruno Rocha
@Revisor: Heloísa Vale, Bruna Rocha, Tommaso Raso, Bruno Rocha
@Comments: The participants laugh sometimes. At certain points, the participants open plastic wrappings and eat cereal bars. GIL, EVN and LUI pronounce the word "também" as "tamém". GIL pronounces the word "mandar" as "meandar". Apheretic form: tendeu (entendeu).

Elaborated by the author

linguistic phenomena, namely register variation (i.e., formal vs. informal) and lexical “switch” (i.e., doctors at the hospital talking with other doctors vs. doctors at home talking with their family).

In summary, this chapter has shown L-AcT is a corpus-driven framework that was built to deal with the complexity of oral and spontaneous language. Therefore, this chapter has attempted to provide a brief review of the main foundations of this perspective, which are: large corpus compilation; speech unit derived from prosody; prosody as an indicator of an utterance function; and connection between meaning and prosodic units. Details about C-ORAL-BRASIL I will be presented in the following chapter.

Chapter 5

The data dive

"What are the three foundations of Corpus Linguistics? **DATA, DATA, DATA**"
Vaclav Brezina at the 2022 Lancaster Symposium on Innovation in Corpus Linguistics[1]

Corpus Linguistics (CL) is the only subfield in Linguistics in which its data is in its name[2], which is explainable since CL is data-oriented (Stefanowitsch, 2020). It is important to highlight that the definition of corpus adopted here is **a properly compiled collection of texts (a "body" of language) that is kept in an electronic database for the aim of supporting linguistic study on a language or linguistic variety** (Sardinha, 2000; Baker et al., 2006). Moreover, as discussed in section 2.3.2, CompSoc is an area in which data quality is highly important. In order to achieve that, the data must be well-described and explored. Thus, this chapter attempts to summarize the rationale of the corpus used in the present research, which is the Informal Spoken Brazilian Portuguese Reference Corpus (Raso and Mello, 2012), or C-ORAL-BRASIL I.

C-ORAL-BRASIL I[3], available at https://www.c-oral-brasil.org/, satisfies all the criteria for spoken corpus compilation discussed in chapter 4, which are: high diaphasic variation, prosodic segmentation, high acoustic quality, and text-sound alignment. The corpus intends to record spontaneous Brazilian speech, with an emphasis on the dialect of Belo Horizonte and its metropolitan area in the state of Minas Gerais[4]. Furthermore, it is important to emphasize that C-ORAL-BRASIL I is only composed of **informal** texts. The formal texts, which make up the C-ORAL-BRASIL II, are going to be released soon.

C-ORAL-BRASIL I is the fifth branch of C-ORAL-ROM, a multilingual corpus from four of the major European Romance languages (Italian, Portuguese, French, and Spanish) (Cresti and Moneglia, 2005). The same architecture, segmentation, and transcription criteria[5] were applied to C-ORAL-BRASIL I to ensure compatibility with the corpora that make up C-ORAL-ROM. Each text in the corpus has:

(i) a sound file in *.wav* format;
(ii) a transcription file in *.txt* format;
(iii) a text-sound alignment file in *.xml* format;

[1] See more information about the event at: https://bit.ly/3JZ3s9N
[2] This was claimed by prof. Vaclav Brezina at the 2022 Lancaster Symposium on Innovation in Corpus Linguistics.
[3] The corpus is also available for online queries in http://www.c-oral-brasil.org/db-com
[4] Appendix A contains two maps indicating these regions in Brazil.
[5] Some criteria were modified in C-ORAL-BRASIL I, but they do not interfere greatly in the comparability of the corpora.

(iv) a header with the interaction and participants' metadata in a *.txt* file[6], and;

(v) a file with the parsed transcription with POS tagging made with the PALAVRAS parser (Bick, 2000). An example of a parsed transcription is in Appendix B.

Table 5.1, adapted from Raso (2012b), shows the basic dimensions of C-ORAL-BRASIL I in comparison with the informal part of the corpora in C-ORAL-ROM. The information about the participants' sex corresponds to the informal part of the C-ORAL-BRASIL project and both the formal and informal parts of C-ORAL-ROM.

Table 5.1: Comparing C-ORAL-BRASIL I and C-ORAL-ROM

Corpus	**Number of recording files**	**Duration**	**Number of utterances**	**Number of words**	**Number of male speakers**	**Number of female speakers**
PT-BR	139	21:08:52	34,167	208,130	158	203
PT-EU	86	14:56:31	21,949	165,436	144	117
SPA	89	14:36:21	21,618	168,868	247	163
FRE	98	12:09:54	10,517	152,385	154	150
ITA	92	16:50:49	23,085	154,967	276	17

From Raso (2012b)

As can be seen in Table 5.1, C-ORAL-BRASIL I surpasses the corpora in C-ORAL-ROM in all dimensions. However, this does not mean that one corpus is better than the other, instead, one may claim that more accurate results with C-ORAL-BRASIL I can be retrieved because of its size and balanced dimensions.

Some factors that the sociolinguistic tradition has identified as important for comprehending the structure of speech were chosen to direct the design of the corpora (C-ORAL-BRASIL I and C-ORAL-ROM) to assure comparability (Cresti and Moneglia, 2005). The first factor is the contrast between informal interaction (face-to-face) and formal communication + media + telephone. For each category, a subcorpus was created (C-ORAL-BRASIL I and C-ORAL-BRASIL II).

Another relevant feature of the corpus is the variety of interaction types, which are: monologues, dialogues, and conversations (Raso, 2012b; Rocha, 2016). According to the authors:

(i) **Monologues** are interactions in which one participant's speech is dominant. Moreover, the intervention from an interlocutor is little, causing small or no changes in the textual flow.

(ii) **Dialogues** are chats between two people, which build the communication together.

(iii) **Conversations** happen when three or more people are engaging in the chat.

According to Raso and Mello (2010), the monologic type includes a variety of subject matters (life narratives, interviews, work monologues, etc.), variation in the profile of the recorded persons (family, friends, clients, etc.), and variety in the locations where the recordings were made (workplace, friends' homes, restaurants, etc.). Additionally, a wide range of activities being carried out during the interactions is present in dialogues and conversations, for example, two or more people cooking together, two or more people working at a computer together, an individual explaining to others how a technological device works, in addition to having the same above-mentioned variations for individuals recorded and places where recordings were carried out.

[6]The headers in C-ORAL-BRASIL I are in distinct files whereas, in C-ORAL-ROM, they are integrated into the transcriptions.

The corpus also provides a division between public and family-private contexts, which is explained in Raso and Mello (2010) and Raso (2012b) as dependent on the speakers' role in that particular situation. The place does not dictate if it is private or public, but how the speakers interact does. For instance, if a speaker engages as a friend (family-private) or as a professional (public). Table 5.2, adapted from Raso (2012b), displays the distribution of interaction types and contexts in the corpus.

Table 5.2: C-ORAL-BRASIL I interaction typology

Public			**Family-private**		
48,766 (23.46%)			159,364 (76.56%)		
Dialogues	**Conversations**	**Monologues**	**Dialogues**	**Conversations**	**Monologues**
17,997	14,547	16,222	55,361	51,887	52,116

Adapted from Raso (2012b)

In terms of transcription, C-ORAL-BRASIL I uses semi-orthographic criteria to identify certain potential lexicalization or grammaticalization processes in Brazilian Portuguese. The corpus transcription takes into account a number of phenomena, including apheresis (e.g., *está* > *tá*), rotacism, pronominal reduction (e.g., *ela* > *ea*, *ele* > *e'*, etc.), loss of plural marking (e.g., *as meninas* > *as menina*), and others (Mello et al., 2012).

Likewise, C-ORAL-BRASIL I contains a high level of diaphasic variation, which was a crucial stepping stone while compiling the corpus (Raso and Mello, 2010). Considering that this corpus aims at representing Brazilian spontaneous speech, this justifies why the team sought to cover as many situations as possible. Some examples are: a personal training session, an amateur football game, a visit to a patient in a hospital, the make-up session of a drag-queen before her show, and a mother-to-daughter cooking lesson[7].

In short, C-ORAL-BRASIL I was compiled through the following methodological techniques:

(i) Spontaneous interactions were recorded in natural contexts, with high-quality wireless equipment.
(ii) Recordings were transcribed by expert transcribers according to the criteria established in Mello et al. (2012).
(iii) Transcriptions were reviewed.
(iv) Transcriptions were statistically validated[8].
(v) Text and sound files were aligned.
(vi) Corpus lexically and morphosyntactically tagged with the PALAVRAS parser (Bick, 2000);
(vii) Subcorpus composed of 20 texts informationally tagged according to the Language into Act Theory[9].
(viii) Informal language corpus and informationally annotated subcorpus were made available online.

To conclude, in this part of the chapter, the main features of the C-ORAL-BRASIL I were shown. In the next section, the sociolinguistic variation in the corpus is presented.

[7] This was possible because of the high-quality equipment used while recording, which was **wireless** monodirectional microphones and a mixer for recording conversations.

[8] Whenever a new group of transcribers and annotators are trained, their transcriptions go through a Kappa test (Fleiss, 1971) to check the degree of agreement between the marked prosodic breaks in each transcription.

[9] For more information about that, check Raso (2012a)

5.1 Diastratic variation in the corpus

This thesis' research was done considering the speaker's sociolinguistic description present in the metadata files. Because of that, this section will present the corpus dimensions in regard to the diastratic variation. As presented in Figure 4.5, the headers have:

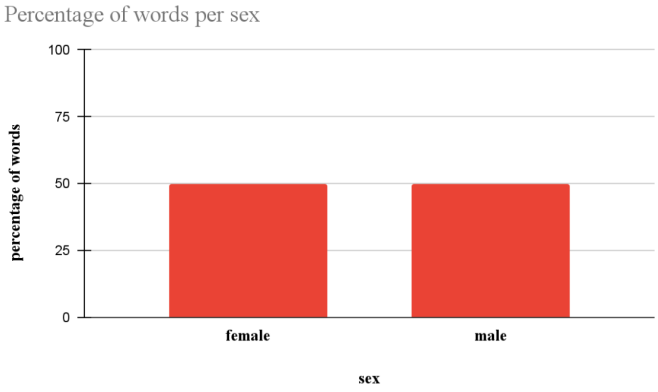
(i) Place of interaction;
(ii) Situation
(iii) Topic
(iv) Participants' information, which is:

  (a) Name abbreviation
  (b) First name
  (c) Sex
  (d) Age
  (e) Schooling level
  (f) Occupation
  (g) Role
  (h) Place of origin

The focus of this research is on **sex, age**, and **schooling level**. Almost sixty-nine percent (68.23%) of speakers on the corpus are registered with their complete sociolinguistic information. The rest was not labeled as their participation in the interaction was not foreseen.[10] (Raso, 2012b). The following charts show the distribution of the diastratic dimensions on the corpus according to the percentage of tokens/words.

The number of male speakers is 158, while the female is 203. Only one speaker who said one word was marked as unknown. According to Raso (2012b), the balance in the sex category is almost perfect, considering the number of words produced by each group. Figure 5.1 brings a graphical representation of this division.

Figure 5.1: Sex distribution

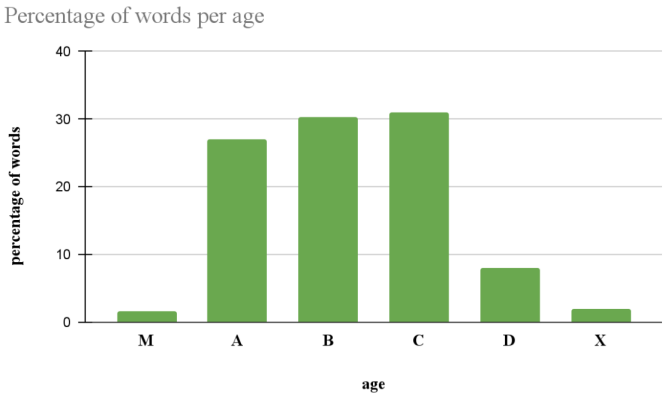


Adapted from Raso (2012b)

[10] These refer to the people who passed by the recording site but are unknown.

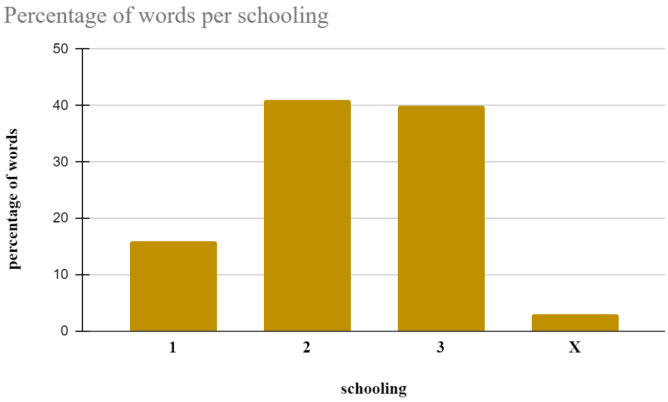
Figure 5.2 displays the age arrangement in the corpus. **Age M** is less than 18 years old. **age A** refers to between 18 and 25, **age B** is between 26 and 40, **age C** is between 41 and 60, **age D** is more than 60 and **age X** is unknown. As it can be seen, the corpus is well-balanced considering the speakers who are between 18 and 60 years old. Moreover, it follows the age distribution in the country, most specifically in the state of Minas Gerais in regards to the last census in 2010[11].

Figure 5.2: Age distribution



Adapted from Raso (2012b)

Figure 5.3: Schooling distribution



Adapted from Raso (2012b)

Figure 5.3 shows the schooling distribution in the corpus regarding the number of words. The caption needs some explaining: **group 1** is made of people who do not have any schooling or had less than or equal to 7 years of schooling; **group 2** concerns people who had taken a university major but their jobs do not require

[11] Age pyramids from Minas Gerais and Brazil can be found in https://bit.ly/3AWtAyY

such degree; **group 3** is formed by people who had taken a university major and their job requires this level of degree. Finally, **group X** is unknown. Table 5.3 summarizes the social factor codes and descriptions.

Table 5.3: Age, sex and schooling codes and their description

Factor	Code	Description
Age	A	$18 \leq age \leq 25$
	B	$26 \leq age \leq 40$
	C	$41 \leq age \leq 60$
	D	$age \geq 61$
	M	$age \leq 18$
Schooling	1	no education or up to 7 years
	2	with university major but job does not require it
	3	with university major and job requires it
Sex	male	-
	female	-

Elaborated by the author

In regards to the occupation distribution on the corpus, 42 different jobs were identified, e.g. teacher, physician, driver, housewife, and retired. This is not documented in Raso and Mello (2012) but the occupation numbers were useful for one of the results later discussed in chapter 7.

This chapter has reviewed the key aspects of the C-ORAL-BRASIL I, which is the corpus used as the data source in this investigation. Firstly, the technical elements were presented and they were followed by the most important components of its architecture. Furthermore, we also showed the methodology of compilation, and finally its diastratic distribution. The next chapter describes the procedures and methods used in this research. The actual data used in the models are going to be presented in section 6.2.3.

Chapter 6

Uncovering the method

"So we mathematically define classes of linguistic representations and formal grammars [...] that seem adequate to capture the range of phenomena in human languages."
Ryan Cotterell in a Medium post[1] *about Computational Linguistics*

As explained by the chapter title, the methods will be presented and discussed below. To begin with, the procedures taken here can be divided into four streams:

(i) **corpus linguistics**: getting to know the type of data that is going to be dealt is of utmost importance to any kind of data-oriented work. Thus, one of a corpus linguist's most essential jobs is learning about the raw data and its corpus.

(ii) **data engineering**[2]: since speech transcription is a type of unstructured data, the researcher must adapt it to a structured data format without losing its complexity so that a model can capture its nuances.

(iii) **data science**[3]: after converting the raw data into a model-readable structured format, we tried to look for models that fit the types of data we had, run the statistical model, and watch its performance.

(iv) **sociolinguistics**: having the results on our hands, it is time to interpret them considering the sociolinguistic analysis. In this part, the linguist has to answer questions such as "What do the findings X and Z tell me about sociolects A and B?".

Although some streams can be more Linguistics- or Statistics-related, the objectives and research questions guided the steps of this analysis in order to have a seamless and interdisciplinary methodology. A careful description will be made for each part in the following subsections.

6.1 Corpus linguistics

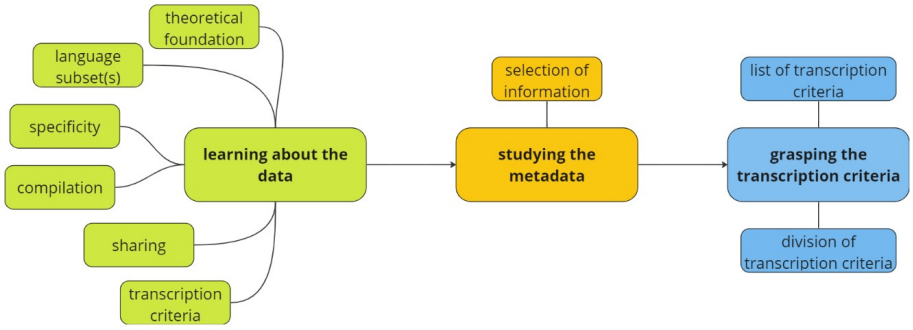
Since a corpus presupposes a large amount of data (Reppen and Simpson-Vlach, 2019; Stefanowitsch, 2020), a corpus linguist has to be familiar with the features and challenges the data poses. That may include understanding the corpus:

[1] *Check the post here:* http://bit.ly/3jiVTBN
[2] A **data engineering** professional is in charge of preparing the data for analytical and operational uses.
[3] Someone who works with **data science** is responsible for extracting meaningful insights from the data.

1. **theoretical foundation**: What theory of language was the basis for the corpus compilation?
2. **specificity**: How specific is(are) the theme(s) in the texts? Are they about the same topic or do they cover a range of topics?
3. **variety(ies) covered**: What language dialects, registers, and styles are included in the corpus?
4. **compilation procedures**: How is the data collected and stored? What and how is the metadata stored and shared? What materials are going to be used? How many teammates are needed? How will the workflow be from recording/extracting data to sharing? The answers to these and other questions impact the corpus design.
5. **data sharing**: How can other linguists access the corpus? Where should the corpus be available?
6. **transcription criteria (for speech corpora)**: What criteria were used to transcribe the spoken data? What phenomena did the team choose to focus on?

As the methodology for this stream, the following steps were adopted: (i) learning about the data; (ii) studying the metadata information; and (iii) grasping the transcription criteria. The first part was addressed in chapters 4 and 5. The main point in this part is that the researcher must know their data well, especially when they did not participate in the data gathering and compilation processes[4]. Some of the actions that it might encompass are reading papers and books about the corpus, studying the corpus documentation, getting training from the compilation team, and making queries in the corpus. These steps are essential to learning the corpus challenges, complexities, advantages, and uniqueness. Figure 6.1 displays all the steps taken in this stream.

Figure 6.1: Corpus linguistics stream steps



Elaborated by the author

In this section, the elements selected to incorporate into the model are introduced. It is worth mentioning that some of the linguistic variables were chosen because they were part of the annotation and transcription guidelines in the C-ORAL-BRASIL I, that is, one would need a corpus with similar protocols[5] in order to run the model we did.

[4]This was one of the remarks during our presentation in the 1st International Conference on Data and Digital Humanities Text Mining and Multimodal Storytelling — in 2023 (http://bit.ly/3yR5Fz7).

[5]That would be the case of the C-ORAL-ROM corpora with small adjustments.

After learning more about the corpus, the metadata files were analyzed regarding which piece of information would be incorporated into this study. The metadata file contains plenty of information about the speakers and the interaction. The chosen sociolinguistic features were **sex**, **age**, and **schooling** because they have extensively been explored by the sociolinguistic literature and, thus, could be easily compared with other studies. It is important to comment that the impact of sex on linguistic units is generally negligible unless the context involves phonetic phenomena[6]. Nonetheless, we will keep the sex variable to examine how the linguistic elements behave according to this predictor as well as to use as many data as possible from the metadata files.

Even though the corpus displays other information, such as profession and place of origin, these pieces of information would be too granular for the small number of speakers, or they could not be enough. For example, there are more than 40 different professions represented by the speakers; in order to include this piece of information information in our analysis, there should be a higher number of speakers distributed in those 40+ jobs. Regarding the place of origin, the analysis would be biased as most of people are from Belo Horizonte. Besides that, in general, people in nearby towns around Belo Horizonte may speak a bit differently than those in Belo Horizonte. (check Reis et al. (2011)). Therefore, there should have been a balance in that regard as well. Nevertheless, it is relevant to mention that C-ORAL-BRASIL I was built to follow the C-ORAL-ROM guidelines and it is a well-compiled corpus considering its purpose. Nevertheless, it must be taken into account that it was not primarily created to serve sociolinguistic research.

Following that, all transcription criteria were listed. The focus of this research was on a set of grammaticalization and lexicalization phenomena marked in the transcriptions of C-ORAL-BRASIL I. Our intention was to verify if these processes were also marked sociolinguistically. From that list, the treatment each criterion should receive was described. For example, initially "paralinguistic sounds with informational relevance" should be counted per speaker and per utterance. Later on, the number of these sounds was so small that they were taken out of the pre-set list of variables. Some of the transcription criteria were not considered in our analysis right from the beginning, such as acronyms (e.g., "CEMIG") and onomatopoeias (e.g., "au"), which are more related to the interaction topic than to the sociolect per se.

It is possible to divide these criteria into three categories: model variables (check Table 6.1), sociolect simple metrics, and not implemented. A description of the first two categories is going to be introduced next. The last category has its elements listed in Appendix C.

[6]Physiological aspects, particularly those associated with the formation of the vocal tract, such as the thickness of the vocal folds, assume significance in this regard.

Table 6.1: Model variables

Phenomena	Sociolinguistic element	Example
apheresis	the lower the schooling level, the more tendency of saying apheretic forms. (Mollica et al., 1998; Pezarino et al., 2023)	"bora" [ˈbɔ.ɾə] *(let's go)* instead of "embora" [ĩ.ˈbɔ.ɾə] or [ẽ.ˈbɔ.ɾə]
rhotacism	the older and less educated a person is, the more the tendency of saying non-standard forms (Oliveira et al., 2022)	"pranta" [prɐ̃ˈ.tə] *(plant)* instead of "planta" [plɐ̃ˈ.tə]
senhor/senhora pronunciation	the lower the schooling level, the more the tendency is the non-standard forms (our hypothesis)	"sior" [si.ˈor] *(sir)* instead of "senhor" [si.ˈɲor]
pronunciation of diminutives	it may be frequent in all sociolects studied here (Rodrigues, 2019)	"jeitim" [ʒeɪ.tʃĩˈ] *(way, solution)* instead of "jeitinho" [ʒeɪ.tʃĩˈ.ɲʊ̃]
verb conjugation	our hypothesis is that it would not be sensitive to age, schooling, or sex because words such as "vamo" are learned since childhood in Belo Horizonte	vamo-1st.PL.present (go)
verb agreement	the higher the schooling level, the more frequent the standard agreement is (Monte, 2019; Oliveira and Santos, 2020)	eles-PERS.PL.3rd vai-V.S.3rd *they go*
plural marking in noun phrases	the lower the schooling level, the higher the number of non-standard plural marking is (our hypothesis)	os-DET.PL.MALE menino-N.S.MALE *the boys*
foreign words	people with higher levels of schooling tend to use more foreign words	"because" (English) and "anche" (Italian)
pronominal phenomena	younger people tend to use the reduced forms more than older people (Peres, 2006)	"ocê" [o.ˈse] or "cê" [ˈse] instead of "você" [vo.ˈse]
reduced and articulated prepositions	the higher the schooling level, the more the preference towards the standard variants (Silva, 2010) the older the speaker, the higher is the preference for no-standard variants (de Souza Santos and Silva, 2021)	"pr'"[pɾ], "pra" [ˈpɾa] or "pa" [ˈpa] instead of "para" [ˈpa.ɾə]
negation particle (also including double negation)	younger people prefer the reduced forms (Ramos, 2002; Avelar et al., 2013)	double negation, "nũ" [nũ] or "n'" [n] instead of "não" [nɐ̃ʊ̃]
interjections and exclamations	it may be frequent in all sociolects studied here	"Nu'", "Nossa", and "No'"

Elaborated by the author

Simple metrics

1. Frequent tokens,
2. Frequent n-grams,
3. Number of utterances,
4. Number of tokens per utterance.

As can be seen in Table 6.1, some phenomena already have literature stating that they are affected by specific sociolinguistic variables. However, since there is no work like the one reported here on Belo Horizonte or Mineiro Portuguese, all linguistic variables will be used in the model for all external variables. It is also important to mention that the metrics in List 6.1 are textual and numerical features from the corpus that were used to get more information about it since they are part of the good practice guidelines in CL (Evison, 2010; Paquot and Gries, 2021), but they were not implemented into the computational model.

In the next section, the procedures for preprocessing and transforming the corpus into a structured machine-readable format will be described.

6.2 Data engineering

There are many different formats in which data can be found, including structured and unstructured tables, pictures, texts, audio files, and videos. For a machine to read and process them, it is crucial to first clean any noisy and unwanted elements (Hemalatha et al., 2012) and then convert the given data into a machine-readable format (Maharana et al., 2022). With that in mind, this section is going to address the procedures for performing such tasks. Firstly, the operations done in the POS-tagged files are described, and they are followed by others carried out in the header files[7]. The Python coding[8] can be found at `https://github.com/joaoprvictor/sociolinguistic-profiling-pt-br-masters`.

6.2.1 Preprocessing and extracting features of the transcriptions

The POS-tagged files in txt format show tokens, lemmas, and POS tags. Figure 6.2 displays an extract of a POS-tagged file in the corpus. Moreover, Appendix D presents the tagset lists used in PALAVRAS (Bick, 2000), which was the parser used in the corpus. The POS-tagged files (in txt format) were chosen because they have more information besides the simple textual flow. The steps in this phase are displayed in Figure 6.3.

[7] In this thesis, "metadata file" and "header file" are considered synonyms.

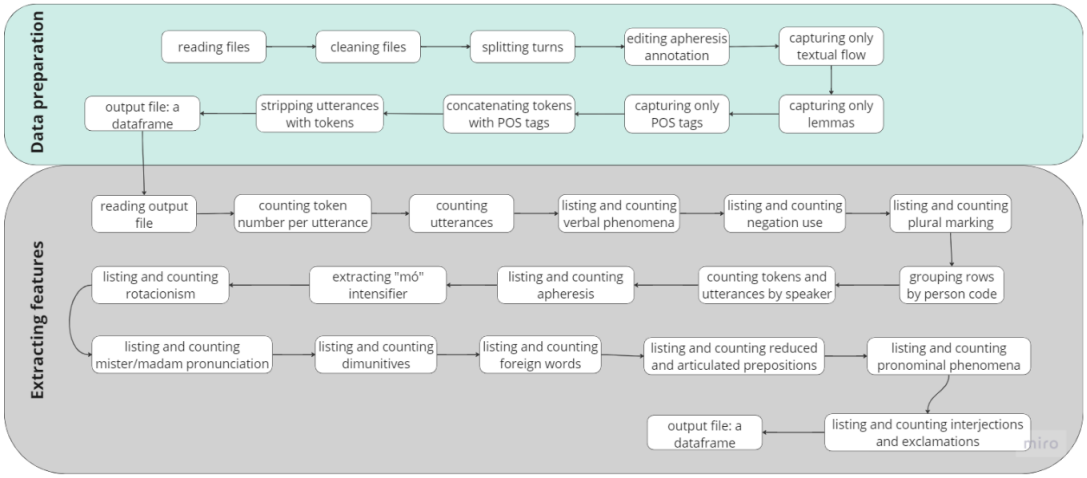
[8] Although the statistical model used here was done in R, the steps before applying the model were conducted in Python, because I am more proficient in Python than in R.

Figure 6.2: Extract from the file bfamcv01.cg.pos.txt

```
*LEO: o [o] <DET M S> Juninho [Juninho] <PROP M S> foi [ser] <V PS 3S IND VFIN> //
*GIL: ô [ô] <IN> / mas [mas] <KC> / voltando [voltar] <V GER> a [a] <PRP> a [o] <DET F S> questão [questão] <N F S> / falando [falar] <V GER> em [em] <PRP> e [e] <KC>
também [também] <ADV> falando [falar] <V GER> em [em] <PRP> povo [povo] <N M S> mascarado [mascaran] <V PCP M S> / esse [esse] <DET M S> povo [povo] <N M S> de [de]
<PRP> o [o] <DET M S> Galáticos [Galáticos] <PROP M P> é [ser] <V PR 3S IND VFIN> muito [muito] <ADV> palha [palha] <N F S> / eu [eu] <PERS M/F 1S NOM> acho [achar] <V
PR 1S IND VFIN> que [que] <KS> es OALT eles [eles] <PERS M 3P NOM> nũ OALT não [não] <ADV> deviam [dever] <V IMPF 3P IND VFIN> mais [mais] <ADV> participar
[participar] <V INF> / e [e] <KC> tal [tal] <DET M/F S> //
*LUI: não [não] <IN> //
*LEO: não [não] <IN> //
*LUI: eu [eu] <PERS M/F 1S NOM> acho [achar] <V PR 1S IND VFIN> não [não] <IN> //
*LEO: com [com] <PRP> certeza [certeza] <N F S> //
*LUI: com [com] <PRP> certeza [certeza] <N F S> es OALT eles [eles] <PERS M 3P NOM> nũ OALT não [não] <ADV> vão OALT vamos [ir] <V PR 1P IND VFIN> participar
[participar] <V INF> / uai [uai] <IN> //
*LEO: eles [eles] <PERS M 3P NOM> são [ser] <V PR 3P IND VFIN> piores [mau] <ADJ M P> do=que [do=que] <KS> o [o] <DET M S> Durepox [Durepox] <PROP> //
*EVN: é [ser] <V PR 3S IND VFIN> / pois OALT depois [depois] <ADV> é [ser] <V PR 3S IND VFIN> //
*LUI: agora [agora] <ADV> manda [mandar] <V PR 3S IND VFIN> uma [um] <DET F S> barrinha [barra] <N F S> minha [meu] <DET F S> //
```

Elaborated by the author

Figure 6.3: Flowchart of preprocessing and extracting features of the transcriptions



Elaborated by the author

One of the biggest conundrums when dealing with spontaneous speech is the definition of *inaccurate, noisy, and inconsistent data* (Hemalatha et al., 2012), mainly because everything that happens in speech is a phenomenon related to the context, our cognition, our history, or all of them combined. However, some elements were deleted or re-arranged to prevent errors or inconsistencies in the Data Science phase (described in section 6.3). List 6.2.1 specifies these elements in the transcription files.

Elements deleted or corrected in the transcriptions through coding

- Extra whitespaces (deleted)
- Speaker's code (deleted and saved in another variable)
- Utterance number (deleted)
- Hesitation markup (deleted)
- Parser annotation bugs (such as "doro [dolo]" changed to "doro OALT adoro [adorar]")

- Whitespace bugs (such as “três[três]” to “três [três]”)
- Empty turns (such as “*FLA: ” - deleted)
- Interrupted words (deleted)
- Extra angular brackets (replaced by a single symbol)
- The letter “ü” (replaced by “u”)

After cleaning the transcriptions, they were converted into a data frame, which is a two-dimensional structure organized in rows and columns. Then, we divided the rows (represented by each turn) by utterances, i.e., each row of the new data frame corresponds to a single utterance. By doing that, it was possible to extract information from each utterance.

Another concern was the non-standard forms because their annotation, “non-standard form OALT[9] standard form”, can impact the counting of tokens. For example, “ea OALT ela[10]” can be interpreted as having three words depending on the counting method. Thus, in order to avoid such a problem, they were united by a hyphen (-), which would make the previous example “ea-OALT-ela”. However, there are three types of non-standardization annotation. Examples of them are:

1. vão OALT vamos [ir] <V PR 1P IND VFIN> - *verb GO in the simple present, 1st person, plural*

   Note: non-standard form, OALT, standard form, lemma, and POS tag

2. pro OALT para [para] <PRP> pro-2 OALT o [o] - *preposition TO + article THE indicating singular, male*

   Note: non-standard form, OALT, standard form-1, lemma-1, POS tag-1, non-standard form-2, OALT, standard form-2, lemma-2, and POS tag-2

3. em [em] <PRP> numa-2 OALT uma [um] - *preposition IN + article A indicating singular, female*

   Note: standard form, lemma, POS tag-1, non-standard form-2, OALT, standard form-2, lemma-2, and POS tag-2

In the last two cases (with *pro* and *numa*), if we combine the non-standard forms with the standard forms, a new problem emerges: the number of lemmas and POS tags differ from the number of tokens. Therefore, a similar procedure of combination was also done with lemmas and POS tags to ensure the same number in all parts of the utterance. Taking *numa* as an example, the final string would be “em-numa-OALT-uma [em-um] ⟨PRP-DET F S⟩”. It is important to mention that this part was also manually reviewed. Different utterances were checked for the sake of monitoring the Python function created for this purpose.

After this operation, four new columns were created: (i) only textual information; (ii) only lemmas; (iii) only POS tags; and (iv) tokens and their POS tags united by an underscore. This type of division can be useful while searching for tokens that can have different tags depending on the context. Furthermore, this data frame was exported as both CSV and Excel files. Figure 6.4 shows an extract from the new data frame. The speaker bfamcv02JAE is represented twice because, in the same turn, they had more than one utterance. Furthermore, column “utterance_with_OALT” is the same as column “utterance” but with the non-standardization normalization previously done.

[9] This symbol indicates normalization.
[10] It means the personal pronoun “she”.

Figure 6.4: Extract from the data frame with the different configurations of each utterance

	acronym	file	utterance	person_code	utterance_with_OALT	clean_utterance	lemmas	pos_tagged	words_with_tags
328	JAE	bfamcv02	não [não] <IN>	bfamcv02JAE	não [não] <IN>	não	não	<IN>	não_<IN>
328	JAE	bfamcv02	nũ OALT não [não] <ADV> gostei [gostar] <V PS 1S IND VFIN>	bfamcv02JAE	nũ-OALT-não [não] <ADV> gostei [gostar] <V PS 1S IND VFIN>	nũ-OALT-não gostei	não gostar	<ADV> <V PS 1S IND VFIN>	nũ-OALT-não_<ADV> gostei_<V PS 1S IND VFIN>
329	TER	bfamcv02	oh [oh] <IN> adorei [adorar] <V PS 1S IND VFIN>	bfamcv02TER	oh [oh] <IN> adorei [adorar] <V PS 1S IND VFIN>	oh adorei	oh adorar	<IN> <V PS 1S IND VFIN>	oh_<IN> adorei_<V PS 1S IND VFIN>
330	RUT	bfamcv02	o' OALT olha [olhar] <V IMP 2S VFIN>	bfamcv02RUT	o'-OALT-olha [olhar] <V IMP 2S VFIN>	o'-OALT-olha	olhar	<V IMP 2S VFIN>	o'-OALT-olha_<V IMP 2S VFIN>

Elaborated by the author

After reading the exported data frame[11], the utterances were merged according to the "person code" column. This step was conducted to ensure that each data point concerns a speaker because this study intends to describe the speakers' language to identify sociolect patterns. Moreover, the general metrics were retrieved. These metrics included (i) the total number of speakers (consisting of known, unknown, and repeated speakers[12]); (ii) the total number of utterances; (iii) the total number of tokens; (iv) the average number of utterances per speaker; (v) the average number of tokens per speaker; and (vi) the average number of tokens per utterance.

Finally, the features in Table 6.1 were extracted. Each phenomenon received a course of treatment, which is displayed in Table 6.2. Since C-ORAL-BRASIL I is well-documented, it was possible to get the strings of interest from the book appendix (Raso and Mello, 2012). For example, there is an appendix section in which foreign words are listed. A language identification algorithm was not used because it could have presented some inconsistencies with, for instance, non-standard forms, which are plenty in the corpus. If the "source" column in Table 6.2 is "book appendix", this means that the textual information to retrieve a phenomenon was taken from the appendix in Raso and Mello (2012). If the linguistic element was retrieved from queries in the corpus, the "source" column is "transcription". In this case, the string was searched in the corpus to check how it was parsed and annotated due to the importance of its tags to the query. Each phenomenon underwent the treatment of a numerical procedure, "count forms per speaker".

The plural marking in nominal groups, verbal, and negation phenomena were extracted while the data frame data point was the utterances. This was conducted to ensure the regular expression query adhered to the utterance syntax and boundary. If it had been done after making the speaker a data point, the regular expression could have captured patterns that would surpass the utterance limits, thus, retrieving false results.

[11] In our case, we used the CSV file. If someone chooses to work with the Excel file, the function pd.read_csv() should be replaced by pd.read_excel().

[12] The total number of unique speakers was retrieved in a later phase.

Table 6.2: Source of each model variable

Phenomena	Source
apheresis	book appendix
rhotacism	book appendix
senhor/senhora pronunciation	book appendix
diminutives	transcription + book appendix
verb conjugation	book appendix
verbal agreement	transcription
plural marking	transcription
foreign words	book appendix
pronominal phenomena	book appendix
reduced and articulated prepositions	book appendix
non-standard negation particles	transcription + book appendix
interjections and exclamations	transcription

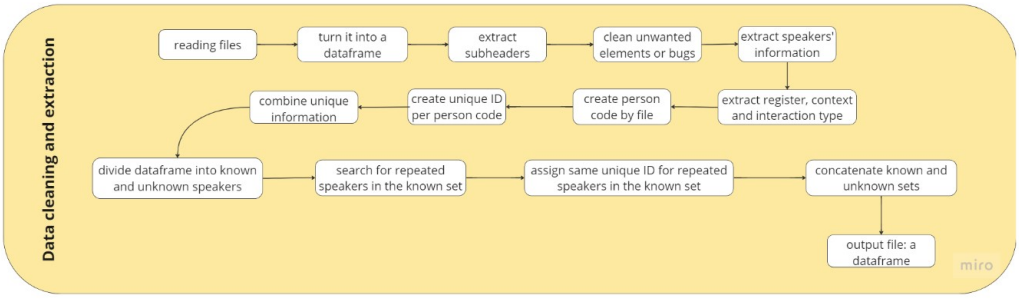
Elaborated by the author

The following subsection describes the procedures to preprocess and extract information from the header files.

6.2.2 Preprocessing and extracting features of the headers

Even though the metadata files are somewhat more structured than the transcriptions, their preprocessing was just as challenging. Some of the issues included: different file encodings, no consistency of alignment, and no regularity in the number of information entries from each speaker. For this reason, the preprocessing strategies used in the metadata files are discussed in this subsection. To visually understand the method, Figure 6.5 shows the operations done in this phase, and Figure 6.6 displays one of the header files.

Figure 6.5: Flowchart of preprocessing and extracting features of the header files



Elaborated by the author

Firstly, all information in the header was read using the simple Python function “open()”, which presents no problem in reading files with different encodings. Then, each piece of information was retrieved and

Figure 6.6: Metadata annotation

```
@Title: Soccer championship
@File: bfamcv01
@Participants: LEO, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
        GIL, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
        LUI, XXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
        EVN, XXXX (male, A, 2, undergraduate student, participant, Belo Horizonte/MG)
@Date: 31/05/2009
@Place: one of the participants' home, Belo Horizonte/MG
@Situation: chat among four students who have organized a soccer championship, four clip-on microphones
not hidden, non-participant researcher
@Topic: positive and negative aspects of the championship, plans for the next championship
@Source: C-ORAL-BRASIL
@Class: informal, family/private, conversation
@Length: 7' 00"
@Words: 1482
@Acoustic_quality: A
@Transcriber: Bruno Rocha
@Revisor: Heloísa Vale, Bruna Rocha, Tommaso Raso, Bruno Rocha
@Comments: The participants laugh sometimes. At certain points, the participants open plastic wrappings and
eat cereal bars. GIL, EVN and LUI pronounce the word "também" as "tamém". GIL pronounces the word
"mandar" as "meandar". Apheretic form: tendeu (entendeu).
```

Elaborated by the author - taken from the bfamcv01 file

placed in a data frame column through a series of regular expression queries. By doing so, most issues related to alignment and number of information entry were solved. For example, the text following "@Place:" until the line break is stored as the place of interaction. After that, texts were cleaned. Strings such as "[]" were deleted because they were a by-product of the operations previously performed in addition to not offering any information about the speakers or the interaction.

Speakers were stored as a list in the data frame, in which each item represented a speaker and their information. Because of that, each speaker's list was split to have one participant per row. Then, another round of cleaning was necessary. All the speaker's information was in one column, so it was imperative to divide each part into one column. In the beginning, there was a column value such as "LEO, (first-name)[13] male, A, 2, undergraduate student, participant, Belo Horizonte/MG". After the division, several columns were created: acronym, first name, sex, age, schooling, job, role, place of origin, and extra information[14]. Then, the "Class" column, from the "@Class:" subheader, was divided into three: register ([in]formal), context (private or public), and interaction type (monologue, dialog, or conversation).

Since this study is interested in how sociolects differ from each other, it was important to assign codes and, later, unique IDs to each speaker. The "person code" is a combination of the file name with the speaker's acronym, for example, "bfamcv02RUT". Also, a new column was created, which was a merge of first name, sex, age, schooling level, place of origin, and occupation. By doing so, each speaker would have a unique

[13] Intending to protect the speakers' privacy, we are not presenting their names.

[14] Only a few speakers have extra information to provide background information about them, especially the ones not born in Belo Horizonte/MG.

code. Nevertheless, such a procedure would need to be simpler and on the known speakers. Known speakers, those with all sociodemographic information in the header, were separated from the unknown speakers[15]. In the known set, speakers with the same merged column received the same unique ID, which was composed of the file name, the speaker's acronym, and the first occurrence index[16]. For example: if the merged column of "bfamcv02RUT" was repeated, all their repetitions would receive the ID "bfamcv02RUT4". After that, the known and unknown sets were concatenated into one data frame again.

Finally, after preprocessing the transcriptions and the headers, they were merged on the "person code" column, which was common in both. Then, the repeated speakers' information was concatenated in only one row per speaker. In this new larger data frame, a new data frame was created for the whole sample and each sociolect: two for sex; five for age; and three for schooling[17]. Having the necessary information for each sociolect, the simple metrics in List 6.1 (part 2) were extracted and we moved on to the next stream in the methodology. The next subsection reports the sample used in this research. Following that, the implementation of the computational model will be discussed.

6.2.3 Sample description

This thesis focuses on a sample of 248 **known unique** speakers portrayed in transcriptions of recorded speech. The term "known" here refers to people who have all their social information available in the corpus. It is important to note that this number differs from what is stated in the original documentation of the corpus because it was obtained through a laborious and intensive process of data extraction. For this, we investigated numerous methods to extract data from the metadata files to achieve the most accurate information possible. The sample size of 248 speakers was found to be the most trustworthy[18] representation of the known unique number of speakers in the corpus after the use of several different techniques. Table 6.3 displays a simple description of the sample used here.

Table 6.3: Sample numerical description

Description	Number	Remarks
total number of speakers	248	known unique speakers
total number of repeated speakers	55	appear at least in two different interactions
total number of utterances	31,336	
total number of tokens	186,325	
average number of utterances per speaker	≈ 126	
average number of tokens per speaker	≈ 751	
average number of tokens per utterance	≈ 6	

Elaborated by the author

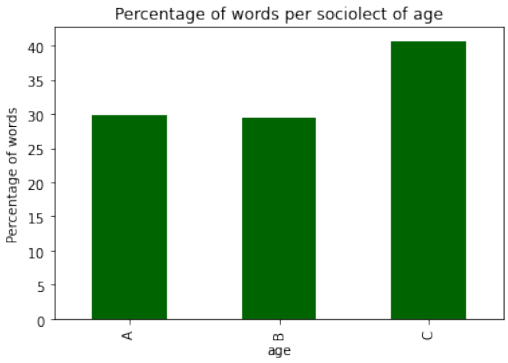
Now, Charts 6.7a, 6.7b and 6.7c show the ratio of words across all social variables. The ratio was done with the number of words uttered by the people of a sociolect by the total number of words in the sample. Not surprisingly, the ratio was similar to the one described in Raso and Mello (2012) (check Chapter 5),

[15] Most of the time, unknown speakers were by-passers who contributed little to the interaction.

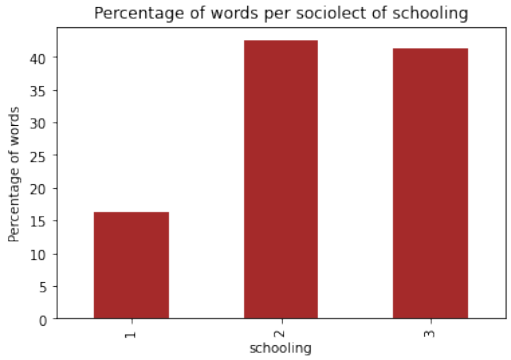
[16] The index from the data frame row.

[17] The script also created a data frame for the sociolect "x", which refers to the speakers we do not know any information. Such a data frame was not relevant for this research.

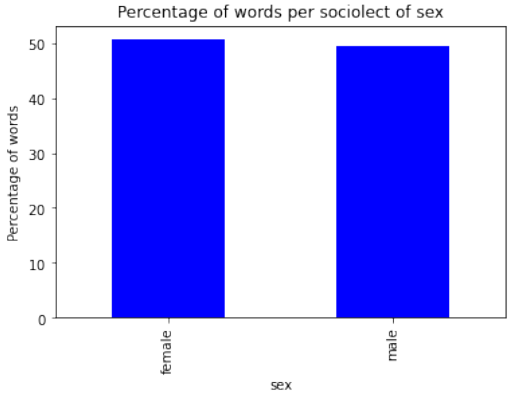
[18] We are not claiming that the number in the corpus documentation (check Chapter 5) is wrong, instead, we are defending that, considering the multiple configurations on the files caused by human errors, this number was the best we could get.



(a) Word percentage in age sociolects from the sample



(b) Word percentage in schooling sociolects from the sample



(c) Word percentage in sex sociolects from the sample

Elaborated by the author

which indicates that the corpus had noticeably solid planning to achieve a high level of representativeness. Note that the age categories are different from the ones discussed in Chapter 5. This is due to the fact that it was necessary to edit the age classes. This process will be further discussed in subsection 6.3.1. The table in https://bit.ly/3TwQGWx is an extract of the dataset used in the models.

Next, the steps to adapt the computational model to this research will be described.

6.3 Data science

As previously stated, one of the goals of this research is to test statistical approaches for the task of sociolect detection. In our search for references, we have not found the same methods we employed for sociolinguis-

tics profiling (outlined in the next subsections). This lack of precedent work emphasizes the urgency and originality of the investigation described in this thesis. Hence, the steps carried out were: a basic description of each linguistic variable, mainly extracting the minimum, maximum, average, and median values; then, different statistical methods were run.

6.3.1 Data visualization and sample testing

An easy and straightforward way of checking data distribution is through the use of visualization techniques. In our case, box plots[19] were generated in order to check how the data was arranged in relation to each linguistic variable. Also, box plots were used as they can give us the median, highest, and lowest values. Having the plots at hand, it is possible to see if there is any noticeable difference in the data distribution to further investigate in a statistical test.

A change in the data labels was necessary[20]. After doing some tests with regression models, the ages M and D were adjusted. The sub-datasets with ages M and D did not have enough distinctive data so that the model could compute, that is, there were not enough zeros and non-zeros in the dataset. For example, people in the age M group (less than 18 years old) have only one schooling level (level 1), if any. In order to overcome that and not lose more data, ages A and M were merged, likewise with ages C and D. Thus, in this case, age A would refer to people up to 25 years old, and age C would concern people equal or above 41 years old. The new age span is detailed in Table 6.4. After this preparation, the sub-datasets were ready for the next steps.

Table 6.4: New age spans after adjustments

Age	Description
A	$age \leq 25$
B	$26 \leq age \leq 40$
C	$age \geq 41$

Elaborated by the author

In the next subsection, the Variation-Based Distance and Similarity Modeling method (Szmrecsanyi et al., 2019) or VADIS will be explained.

6.3.2 VADIS — lines of evidence

VADIS is a series of tasks used to compare the grammar of different varieties, dialects, or languages based on empirical data. It borrows concepts and methods from Variationist Sociolinguistics, which examines how different forms can convey the same meaning (Labov, 1973), and Dialectometry, which studies dialects. Thus, this subsection describes the procedures to adapt this method to C-ORAL-BRASIL I, and our purposes. All the steps done in this part of the methodology were based on the R code[21] the VADIS authors

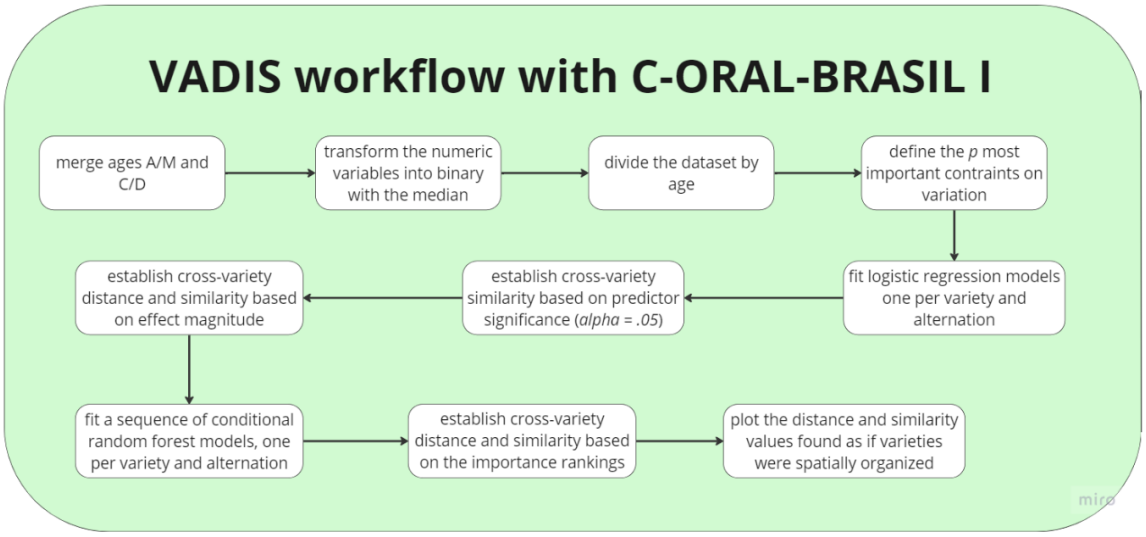
[19] "Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers." (Yi, 2021)

[20] We did not change directly on the data file, because we wanted to maintain the corpus labels. This procedure was done through the opening of new coding files.

[21] Our code adaptation can be found at https://github.com/joaoprvictor/sociolinguistic-profiling-pt-br-masters

kindly shared with the community[22]. Figure 6.8 displays the procedures performed in this stage.

Figure 6.8: VADIS workflow



Elaborated by the author

Because VADIS first step is a logistic regression model implementation, our data must be in a binary format (0 or 1) and be divided into smaller datasets. In the case of our dataset, the median[23] was chosen to be the numeric threshold to make the variables binary: if lower than the median, the variable equals zero; if higher, the variable equals one. The median was determined as a threshold because it is not influenced by extreme values as well and it guarantees that there is 50% of data below and above it. Moreover, VADIS takes a list of datasets to calculate the (dis)similarities. Regarding the dataset splitting, it was decided that "age" would be the dataset divider because it is a biological (i.e., the body changes) and a social (i.e., people group themselves by age) factor. The rationale behind this decision lies in the fact that linguistic patterns may vary significantly between individuals of different age groups. For example, a highly educated young male is likely to have distinct speech patterns compared to a similarly educated older male.

VADIS main phases are divided in order to find three lines of evidence: constraint significance; magnitude of effects; and relative importance ranking. The specificities related to the way the metrics and the code work are described in Szmrecsanyi et al. (2019). Instead, we will describe only the adaptation done for our data.

Line of evidence 1 is "*Are the same constraints significant across varieties?*"[24] (question taken directly from VADIS article), which was answered through the logistic regression models. Our *p* most important constraints were always the sociodemographic variables, schooling, and sex. Age was not put as a variable because the sub-datasets were already divided by it. The model for this line was always structured with the linguistic variable as the dependent or response variable, while the sociodemographic variables were the

[22] VADIS base code can be found at https://github.com/jasongraf1/VADIS

[23] Median is the center number in a sorted list of values. It reflects the midway of the data.

[24] Significance refers to the probability that an association or difference between variables is real and not merely a coincidence.

independent or predictor variables. It was done as such due to the purpose of verifying if and how much the linguistic phenomena listed in Table 6.1 in C-ORAL-BRASIL I could have been influenced by the speaker's social factors (age, sex, and schooling). An example would be $y = sex + age + schooling$.

As the model for each sub-dataset had been fitted, important metrics to VADIS were retrieved. They are:

- predicted correlation.
- Brier score.
- concordance index (or C value).
- log scores.
- Akaike information criterion (or AIC value).
- maximal variance inflation factor (or Max.VIF).
- kappa values.
- and Hosmer-Lemeshow test score (or HosLem.p).

The focus of this research will be on the "C values", "Max.VIFs", and "HosLem.p". C values indicate model accuracy; Max.VIF attests collinearity, if any; while HosLem.p tests the goodness of fit.

- **C values**: values below 0.8 indicate suboptimal model discrimination.
- **Max.VIF**: values greater than 10 point to possible problematic predictor collinearity.
- **HosLem.p**: values lower than 0.05 indicate possible issues.

Having these parameters in mind, each model was assessed concerning its accuracy rate. Some flexibility was needed, because, compared to what was done in Szmrecsanyi et al. (2019), this study had much less data. At first glance, no model was discarded until the three lines of evidence were verified.

With the metrics at hand, it was possible to calculate the cross-sociolect similarity based on predictor significance. The function `vadis_line1()` returned three objects: a table with predictor significance values, a distance matrix using the values in the predictor significance table, and a table of similarity scores[25].

Line of evidence 2 is "*Do the constraints have the same strength across varieties?*"[26]. In this part, based on the sociolect-specific regression models, the magnitude of effects is used to determine cross-variety distance and similarity. This was carried out by calculating a distance matrix based on the model estimates (using Euclidean distance), whether or not the effect sizes of the constraints are significant (Szmrecsanyi et al., 2019). All of this is accomplished through the `vadis_line2()` function.

Line of evidence 3 is "*Is the constraint hierarchy similar?*"[27]. Before doing line 3, it was important to fit a random forest model. Then, with the conditional random forest models, cross-sociolect distance and similarity based on the importance rankings of the predictors were determined. We used Spearman's rank

[25] Check Szmrecsanyi et al. (2019) to see how the similarity scores are calculated.

[26] Strength calculates the degree or intensity of a relationship or effect. It concerns how well a method can capture or detect a particular phenomenon.

[27] Relative importance is a concept used to comprehend the individual contributions of predictor variables to the variability or outcome of the dependent variable.

correlation between the relative variable importance rankings of the two sociolects to calculate the probabilistic distance between them. Again, the similarity computing was conducted through the `vadis_line3()` function.

Finally, with all metrics from all lines at hand, charts were created. Hence, sociolects were a matter of numbers, which made it feasible to display them in distance-based and clustering charts. Afterward, the charts were inspected in order to describe how the sociolects are grouped. The second approach is going to be detailed in the next subsection.

6.3.3 Non-parametric model

The type of data distribution is an important factor in determining which model will have a better fitting for the data. Yet, in some situations, the researcher is not able to make many assumptions about the data. In these cases, *"non-parametric tests are designed to have desirable statistical properties when few assumptions can be made about the underlying distribution of the data."* (Pappas and DePuy, 2004, p. 1). Therefore, for this study, the test used was the Mann-Whitney. One may argue that non-parametric testing is the last resource an analyst should use and that is correct. Our purpose in testing it first was to (i) check if it could retrieve reliable results, and (ii) to make it as a "preparation" for the parametric models[28].

The Mann-Whitney test was performed to check if the sociolects were from the same distribution[29], according to each social variable, and to try a simpler approach to describe our data after applying VADIS. The advantages of using this test are the non-parametric nature, applicability to continuous data, robustness to outliers, usefulness when dealing with small samples, easy interpretation, and observation independence. Due to being a non-parametric test, it does not presuppose any type of specific distribution. Besides, it does not depend on assumptions about the interval properties of the data, hence, the Mann-Whitney test is less sensitive to outliers. Furthermore, this test works well with small samples, where normality may not hold. The Mann-Whitney output is very straightforward to analyze. Finally, in this test, all the observations from the sample groups are independent of each other.

One of the scores the Mann-Whitney test returns is the p-value. For all statistical methods reported here, our significance threshold was 0.05 or 5%. This score was used to decide which social variable was more significant for a certain linguistic variable. Moreover, to avoid a Type 1 error (false positives), the Bonferroni correction in the p-value was necessary, making the threshold approximately 1.6% for age and schooling as they have three (3) levels each.

In the parts that follow, the tests with the Generalized Linear Model (GML) Poisson, and Negative Binomial will be explained. These methods are quite new in Linguistics, as VADIS was created in 2019, while the others have received many "followers" in recent years (Winter and Bürkner, 2021).

6.3.4 Count data models

In Linguistics, it is very common to have frequency data of discrete events, for example, how many instances of *uai*[30] appear in a Mineiro speech corpus. Poisson (PO) and Negative Binomial (NB) models are types of generalized linear models that are canonical methods to compute count data in other fields (Zuur et al., 2009; McElreath, 2018). Because of the nature of our data and the trend of incorporating these regressions in linguistics, these types of models were tested in our research as well. Here, we do not intend to present a

[28] However, parametric and non-parametric models are not complementary.

[29] The code used for this test is in the same file of plotting. You can find it at https://github.com/joaoprvictor/sociolinguistic-profiling-pt-br-masters/tree/main/R/non_parametric

[30] One of the most stereotypical expressions in the Mineiro dialect which can serve various purposes.

detailed full-on introduction to those regressions. Refer to Nicenboim and Vasishth (2016) and Winter and Bürkner (2021), instead.

PO and NB were tested to check if we could get plausible results with less effort compared to VADIS. In a PO model, the mean and the variance of the counts are equal. Moreover, the parameters are calculated using maximum likelihood estimation (MLE). NB models allow the variance of the counts to be greater than the mean and estimate parameters with MLE.

The two models were tested for each dependent variable (i.e., the linguistic variable) according to an R code. For each response variable, two formulas were created: one with no interaction and another with triple interaction. It is worth noting that interaction indicates that the independent variables influence each other. We started with the triple interaction model. Upon identifying whether there was any significant interaction, the model was updated and the test was carried out again. If there was no interaction between the variables, the model presented no interaction. The linguistic variables were modeled with and without predictor interaction as follows:

- triple: $\textit{response variable} = \textit{sex} \times \textit{age} \times \textit{schooling}$
- double: $\textit{response variable} = \textit{sex} + \textit{age} \times \textit{schooling}$[31]
- no interaction: $\textit{response variable} = \textit{sex} + \textit{age} + \textit{schooling}$

Each model was evaluated according to its homoscedasticity, normality, overdispersion, and zero inflation. Homoscedasticity describes a situation in which the "noise" in the data is the same across all values of the independent variables. From another perspective, normality concerns if the data is close to a normal distribution. Moreover, overdispersion is when the variance of the response is greater than what is estimated by the model. Finally, zero inflation happens when there are a lot of zero-valued observations.

Having these elements for each model, the final model for each linguistic variable was chosen. The final model for each linguistic variable was selected based on ANOVA with a 5% (or 0.05) significance threshold. This means that in the end, we ould have a mix of models of different types. For instance: a PO model for apheretic forms and an NB model for reduced prepositions.

After that, the Estimated Marginal Means (henceforth, EMM) were calculated. When a researcher desires to understand the differences between groups while taking into account the effects of relevant variables, they use EMM (Searle et al., 1980). This metric can account for intricate interactions while also providing group means that are accurate and interpretable. EMM achieves this by maintaining the influence of variables constant. Assume you have carried out a study in which you have measured a certain outcome (dependent variable) across several groups or conditions (independent variables). Taking into account the other variables in the study, EMM assists you in estimating the average value of the result for each particular condition or group. In other words, it lets you observe how the average outcome changes when you focus on one independent variable while keeping the others constant. This is very helpful for complicated statistical models.

In conclusion, considering the purpose of describing the main features of a sociolect, it is crucial to determine if sociolects can be grouped according to the linguistic elements analyzed here. For that reason, the Compact Letter Display (henceforth, CLD) technique (Piepho, 2004) was used. This method is given a multiple comparison test, the EMMs in our case, and retrieves the results of tests with multiple comparisons. As implied by its name, the outputs are displayed with letters, signaling which groups are or are not signif-

[31] There could have been other interaction combinations, but the age×schooling interaction was the only double interaction that was significant.

icantly different from each other. According to our search, the CLD technique[32] has not been previously applied in the field of Linguistics. This novel application underscores one more contribution of our research.

In the next section, the sociolinguistic stream of the methodology will be explained.

6.4 Sociolinguistics

With a quantitative approach in mind, Sociolinguistics aims at *"(...) identify[ing] the effect of the different effects/conditions in competition and determine the contribution of each category (structural and non-structural) associated with the occurrence of a certain form as opposed to another or others."*[33] (Gomes, 2012, p. 260). Using the metaphor of language variation as a war or battle, Sociolinguistics watches, describes, and analyzes the battle between linguistic forms and which elements give them strength to win the fight within a certain sociolect. With the help of statistical methods, it is possible to supervise each "army" division and detect which one has better weapons to overrule the other competing sides due to its potential to deal with a large number of soldiers. Thus, in this section, some traditional quantitative methods in Sociolinguistics are debriefed, and why the methods we selected are better at inspecting "our armies".

Some of the most famous quantitative tools in Sociolinguistics are Varbrul/GoldVarb[34], and Rbrul[35]. Varbrul/GoldVarb is a computer tool that was specially built to deal with sociolinguistic data and to run multivariate analyses. On the other hand, besides accommodating sociolinguistic data, Rbrul is a tool that runs in R codes, allows for more model parameter editing, and provides more information about the data (Gomes, 2012).

Even though these apparatuses were a great advancement for the field and some consider them more Humanities-friendly[36], they are not able to implement some fundamental operations. In Varbrul/GoldVarb, one issue is that all independent variables are considered fixed, which may not be the case for all sociolinguistic variables. According to Gomes (2012), this is strongly related to the theoretical and methodological trends back then, in which the speaker and the lexical item were not considered part of the variation model equation. While Rbrul overcame this problem by letting the user add random effects to the model, it only performs logistic regressions. As discussed in part 6.3.4, frequency data, which is very common in sociolinguistic work, is better computed with other types of models.

Therefore, although it may demand more effort from the researcher to learn to code, creating your own model from "scratch"[37] may have its advantages. Some of them are editing the parameters as the researcher desires; using the proper model type according to the data; and bringing innovation to the field as one tests new approaches and models. This is why regression models, deep learning, and other machine learning approaches are becoming famous in linguistics.

With that being said, our goal is to test statistical methods for sociolinguistic profiling. As previously explained, sociolinguistic profiling is a task to find patterns in one's sociolect that could reveal their potential social profiles (Perkins, 2021). Having that in mind, after running the models described in section 6.3, the results were analyzed considering what they can tell about the studied sociolects. If there was a significant

[32] It is very similar to the operation done in the Tukey test.

[33] Original text: (...) *identificar o efeito dos diversos efeitos/ condicionamentos em competição e determinar qual a contribuição de cada categoria (estrutural e não-estrutural) associada à ocorrência de determinada forma em oposição a outra ou outras.*

[34] Created by David Sankoff, Sali A. Tagliamonte, and Eric Smith, and available at: http://individual.utoronto.ca/tagliamonte/goldvarb.html

[35] Created by Daniel Ezra Johnson and available at http://www.danielezrajohnson.com/Rbrul_manual.html

[36] This term refers to the fact that these tools are easier to use, especially for people from the Humanities field who did not know how to code.

[37] There are tons of tutorials and codes available on the internet that can help in the creation of a model.

positive variable interaction between age and schooling in foreign word use, it could mean, for example, that the older and more educated this person is in Belo Horizonte, the more foreign words they use. This is a theoretically expected result and the data would be confirming it in this example. Another case would be if the model with interjections shows no significance in any of the predictors, which might indicate that interjections are not a linguistic clue of a specific sociolect in Belo Horizonte and, in fact, it is rather used across all sociolects, or we do not have enough data.

As for the non-parametric models (discussed in section 6.3.3), for the age and schooling variables, we applied the Mann-Whitney test, considering the Bonferroni correction at a significance level of approximately 1.6%. The results for the sex variable will be presented considering a 5% level of significance.

However, within the context of count data models (as elucidated in section 6.3.4), subsequent to implementation and examination of predictor significance, the initial focus was directed towards scrutinizing the groupings established by the CLD. When a group was different from the other, the EMMs were examined concerning which sociolects contained more and less of the linguistic element at hand.

Chapter 7

Sociolectal dynamics: mapping out variation

"In fact, it has been argued that corpora as such contain nothing but distributional frequency data (...)"
(Gries, 2010, p. 269)

The above statement is quite correct, even though it seems to curtail CL and CompSoc potential. Most corpus studies involve analyzing the frequency of occurrence or co-occurrence of linguistic units. Regardless of the approach the linguist takes, the data can go through a series of statistical operations in order to extract valuable information from the corpus. Because of that, this chapter appertains to the results of the statistical procedures described in chapter 6. First, we will share some important information about the data. Afterward, the results from the computational models will be presented.

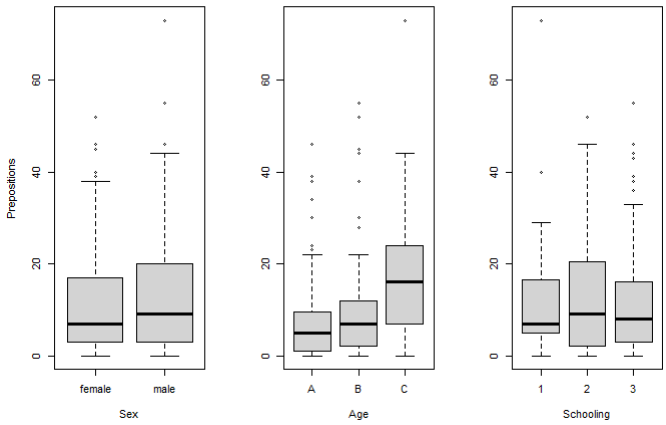
It is important to reiterate the sample used here, which was composed of 248 known unique speakers with approximately 126 utterances and 751 words on average per speaker. Moreover, in order to have a broad idea about the sociolects within the sample, some charts about the n-gram frequency ranking were generated. The focus was on 1-, 2-, and 3-grams. The 3-grams are presented in Appendices E, F, and G. All n-gram charts are at my GitHub repository[1] so as not to visually pollute this thesis. According to the charts, it seems that 3-grams are quite the same across sociolects. The word sequences *que é que*, *eu acho que*, and *como é que* appeared in all sociolects, which may suggest that high-frequency constructions are more related to dialect than to sociolects. Further research on that is needed. The 3-grams that may have appeared in a sociolect and are not common to the other sociolects are likely associated with the speech acts performed during the interaction. For instance, constructions like *a casa de* ('the house of') are probably related to acts that are related to informing location.

Furthermore, box plots were generated in order to check how the data was distributed. Figures 7.1 and 7.2 display the data distribution for the "prepositions" and "interjections" variables, respectively, according to each social factor. As shown in these plots, it seems that there is a similar behavior across social factors in these linguistic variables if we consider the median value (highlighted by the black line within the "box"). In the "preposition" box plot, the most prominent difference is in the `age` factor. It suggests an interesting difference among the sociolects, as the median age for sociolect `C` is higher than the median age for the other two sociolects. In the "interjection" box plot, there are no eye-catching differences. These assumptions were

[1] https://bit.ly/4aItj2y

later tested. Also, box plots allow us to identify outliers. When observing the medians, we may at first think that the distributions could be equal, but to attest it, the non-parametric test was applied.

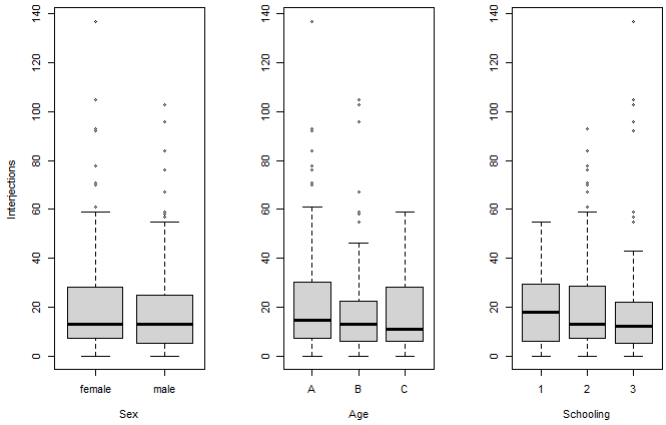
Figure 7.1: Preposition variable distribution



Elaborated by the author

Note: the y axis is the raw count of the prepositions

Figure 7.2: Interjection variable distribution



Elaborated by the author

Note: the y axis is the raw count of the interjections

The results of each method analysis will be presented in the following subsections.

7.1 VADIS assessment

VADIS was not successful in detecting the sociolects within the same geographical region. The model accuracy (marked by the C-values) was low probably due to small amount of data. Moreover, there was no way to get more data that could be compared to C-ORAL-BRASIL I, due to lack of time and resources nor to refine variables since they were well-established sociolinguistic variables. It is important to say that VADIS did not perform well with or without variable interactions. Possible explanations for that are the size of our sample and the conversion of numerical to categorical data. The original study was done with thousands of data points, while we worked with only 248 speakers. Furthermore, the count data needed to be converted to categorical data and by doing so, any type of graduality was lost. In the original paper, the smallest dataset had around 11.000 data points which was very far from the amount we had (248 data points).

On the GitHub repository page[2], it is possible to see an example of all scores extracted by VADIS using non-standard verb conjugation as a focus variable. The charts generated with this linguistic variable were also uploaded there. This was done just as a test to verify how it would work with our data from end to end. Also, it was an opportunity to have a ready-made pipeline to use VADIS with other sociolinguistic variables in future studies with C-ORAL-BRASIL I.

7.2 Non-parametric testing

Non-parametric tests are based on position or rank measurements, and make it possible to analyze a set of data, like parametric tests, but with fewer assumptions (or requirements to be met). As it was mentioned previously, VADIS was used to analyze the data set, but due to the sample size of the corpus, it was not possible to apply it. Thus, we opted, at first, for the Mann-Whitney test, which is the non-parametric version of the Student's t-test. This test is used to compare two independent samples. When analyzing the data set, we found that the variables age and schooling had three levels, so we adapted the test. The adaptation consisted of applying the tests considering only two levels of the three existing ones. However, in this case, we used the Bonferroni correction, which did the adjustment to the significance level depending on the number of comparisons.

For the variables age and schooling, the Mann-Whitney test was applied, considering the Bonferroni correction (significance level of approximately 1.6%). The results are displayed in Tables 7.1 and 7.2. The results for the gender variable are presented in Table 7.3, considering a 5% level of significance.

In Table 7.1, ages A and C demonstrated a greater amount of differences in our linguistic variables likely because they were the extremes in our age spectrum. This may indicate that one of them is either innovating, bringing new variants, or signal differences according to age grading.

As can be seen in Table 7.1, rhotacism, *senhor/senhora* pronunciation, foreign words, and interjections variables were not directly related to age. In other words, there is no statistical evidence for the age labels assessed to be significant in the mentioned variables using the Mann-Whitney test at a 5% level of significance. Possible explanations are:

- the level of dialectal acquisition: interjections, for example, could be a feature of the Mineiro dialect that does not have any stratification in the sociolects. That would mean that native speakers of the

[2] https://bit.ly/4aMFtb0

Table 7.1: Results of the Mann-Whitney test application considering the age variable

Linguistic variable	**A and B**	**A and C**	**B and C**
Non-standard negation forms	Equal (0.1377)	Different (0.0086)	Equal (0.2343)
Non-standard verb agreement	Equal (0.4756)	Different (0.0002)	Different (0.0013)
Non-standard verb conjugation	Equal (0.2952)	Different (0.0036)	Equal (0.0392)
Non-standard plural	Equal (0.5624)	Equal (0.0510)	Different (0.0136)
Apheresis	Equal (0.8714)	Different (0.0065)	Different (0.0041)
Rhotacism	Equal (0.9801)	Equal (0.9439)	Equal (0.9677)
Senhor/senhora pronunciation	Equal (0.1780)	Equal (0.1738)	Equal (0.9945)
Diminutives	Equal (0.7129)	Equal (0.0385)	Equal (0.0172)
Foreign words	Equal (0.1205)	Equal (0.1848)	Equal (0.8170)
Articulated and reduced prepositions	Equal (0.1871)	Different ($<$0.0001)	Different ($<$0.0001)
Pronominal phenomena	Equal (0.1177)	Different (0.0078)	Equal (0.2125)
Interjections and exclamations	Equal (0.7536)	Equal (0.4414)	Equal (0.6076)

Elaborated by the author

Mineiro dialect would learn and use interjections so similarly to the point that would be quite difficult to capture any difference in narrowed sociolects.

- the speaker's origin: most of the people in the corpus are from Belo Horizonte. Rhotacism is, in turn, a phenomenon more related to the rural and countryside areas (Freitag, 2011).

Table 7.2: Results of the Mann-Whitney test application considering the schooling variable

Linguistic variable	**1 and 2**	**1 and 3**	**2 and 3**
Non-standard negation forms	Equal (0.4362)	Equal (0.0556)	Equal (0.1290)
Non-standard verb agreement	Equal (0.2768)	Equal (0.1142)	Equal (0.6430)
Non-standard verb conjugation	Equal (0.6076)	Equal (0.0875)	Equal (0.2365)
Non-standard plural	Equal (0.3566)	Equal (0.2228)	Equal (0.6808)
Apheresis	Equal (0.8094)	Equal (0.9228)	Equal (0.5725)
Rhotacism	Equal (0.4887)	Equal (0.1591)	Equal (0.4036)
Senhor/senhora pronunciation	Equal (0.6210)	Equal (0.7090)	Equal (0.2728)
Diminutives	Equal (0.1014)	Equal (0.0582)	Equal (0.7793)
Foreign words	Different (0.0005)	Different (0.0002)	Equal (0.7568)
Articulated and reduced prepositions	Equal (0.7004)	Equal (0.7289)	Equal (0.4442)
Pronominal phenomena	Equal (0.7819)	Equal (0.4256)	Equal (0.5556)
Interjections and exclamations	Equal (0.7806)	Equal (0.1927)	Equal (0.1933)

Elaborated by the author

Not surprisingly, the only significant difference was found in the use of foreign words between people from the lowest and highest schooling levels. This result is expected since highly educated people tend to have more access to foreign languages.

Table 7.3: Results of the Mann-Whitney test application considering the sex variable

Linguistic variable	**Male and Female**
Non-standard negation forms	Equal (0.5569)
Non-standard verb agreement	Equal (0.6091)
Non-standard verb conjugation	Equal (0.1632)
Non-standard plural	Different (0.0244)
Apheresis	Equal (0.5703)
Rhotacism	Equal (0.5669)
Senhor/senhora pronunciation	Different (0.0354)
Diminutives	Equal (0.8142)
Foreign words	Equal (0.7827)
Articulated and reduced prepositions	Equal (0.2527)
Pronominal phenomena	Equal (0.6273)
Interjections and exclamations	Equal (0.6686)

Elaborated by the author

According to Table 7.3, non-standard plurals and *senhor/senhora* pronunciation show significant differences regarding sex. Values of the male sociolect had great variability (check Appendix I). Meanwhile, values of the female sociolect were concentrated below 3, with only one single value of 13, configuring a low use of non-standard plural marking by females. While the pronunciation of *senhor/senhora* had the opposite behavior, i.e., males tend to use the low-prestige forms less than females. These two linguistic variables are

mostly made of zeros. This is something to pay attention to while getting the outputs from the parametric models since this type of result was not expected[3].

Nevertheless, we must highlight that these results should only be considered in case the parametric models do not present a better fit. In the next subsection, the count data model results will be reported.

7.3 Count data models

As described in section 6.3.4, the models used here were the Poisson (PO) and the Negative Binominal (NB). Each linguistic variable went through the tests with each model with and without interaction between predictors. The final list of formulas and models is displayed in Table 7.4. A table with the intercept coefficients is presented in Appendix H.

Table 7.4: List of formulas and models per linguistic variable

	formula	**model**
non-standard verb conjugation	sex + age × schooling	NB
non-standard negation particle	sex + age × schooling	NB
non-standard verb agreement	sex + age + schooling	NB
apheresis	sex + age + schooling	NB
rhotacism	sex + age + schooling	NB
non-standard plural in noun phrases	sex + age + schooling	NB
senhor/senhora pronunciation	sex + age + schooling	NB
foreign words	sex + age + schooling	NB
reduced and articulated prepositions	sex + age + schooling	NB
pronominal phenomena	sex + age + schooling	NB
interjections and exclamations	sex + age + schooling	NB
non-standard diminutives	sex + age + schooling	PO

Elaborated by the author

From Table 7.4, we can see that most models were of NB type while only one was of PO. Moreover, surprisingly, what stands out in the table is that only two models had their formula with interaction[4] (symbolized by ×), whereas the rest had a no-interaction formula. We believed that there would be more interaction formulas because we, as humans, are many things at once (age, schooling, and sex[5]). Of course, it is possible to modulate them. For instance, at an academic conference, one would be expected to show the expertise they acquired through education, putting in the foreground their schooling experience. Nevertheless, our full identity is not so divisible, instead, we, as scientists, do that to capture nuances in the data.

Since we analyzed twelve variables, the reported results are going to refer only to the Estimated Marginal Means (EMM) to focus on what is important to the sociolinguistic analysis as well as not to tire the reader. What follows now is the description and discussion of the EMMs for each linguistic variable.

Firstly, there was no evidence that sex, schooling, or age can predict pronominal phenomena, apheresis, and rhotacism. This result is somewhat counterintuitive since previous works have shown that they could

[3]The non-parametric and parametric models are independent but we are using both of them to compare results.
[4]As explained in subsection 6.3.4, the variables were tested in triple interaction and then, the simplest formula was selected.
[5]Here it's more plausible to say gender but, in order to keep consistency with our data categories, we decided to maintain sex.

be influenced by such social factors. However, we had expected that interjections/exclamations would not be a significant feature of any of the sociolects because they are stereotypical marks of the Mineiro dialect. Considering that, our hypothesis was confirmed, as interjections/exclamations did not have distinctive groups with the EMMs. It could mean that this linguistic process is somewhat equally spread all over the studied sociolects, thus, the model could not find significant differences.

Concerning the groupings of the linguistic variables that distinguish sociolects, only the predictors with distinctive groupings are reported. For instance, If only the age groupings are displayed in the non-standard negation particle table, it means that this linguistic unit could solely be categorized by the `age` sociolects. In the next subsections, sociolects will be referred to in a `typewriter font`.

7.3.1 Reduced and articulated prepositions

Reduced and articulated prepositions are prepositions that suffered phonetic reduction or are concatenated with another word, such as *pra* (to/for) and *cumas* (with + some-FEM). Table 7.5, as our first result, shows groupings with the variable reduced and articulated prepositions.

Table 7.5: Grouping and EMMs of reduced and articulated prepositions

Age	**grouping**	**response scale EMM**	**EMM average**
A	a	8.38	2.13
B	a	9.89	2.29
C	b	16.89	2.83

Elaborated by the author

Table 7.5 can be read in the following way: looking at the "grouping" column, `age` sociolects A and B are significantly similar, while `age` sociolect C is significantly different from the other two. The "response scale EMM" is the means based on the numbers on the raw data; while the "EMM average" is a relative number based on the groupings. Now, departing from the EMM-related columns, it is possible to say that people from `age` C tend to utter more reduced and articulated prepositions. Our data partially confirms what was said in previous studies as it shows that age is a relevant index marker for the use of reduced and articulated prepositions (de Souza Santos and Silva, 2021), while schooling was not a significant predictor as found in Silva (2010). One conceivable rationale for this may be that older people who come from countryside tend to do more phonetic suppression since it is one of the main traits of the "Minas Gerais central region sociolect" (Pushchanka, 2017). Moreover, schooling may have not had significant distinctive groupings because reduced prepositions, such as *pra* (to/for) and *cum* (with a [masculine]), are well spread not only in the state of Minas Gerais but in Brazil as a whole.

7.3.2 *Senhor/senhora* pronunciation

An example of *senhor/senhora* pronunciation variation is "sior" instead of saying "senhor" (sir). Tables 7.6a and 7.6b display the results for *senhor/senhora* pronunciation for age and schooling sociolects. For this variable, there are two tables because two factors are significant and they are **not** in interaction. Furthermore, whenever a grouping receives "ab", it means that the model could not provide sufficient evidence to make a distinctive grouping with "a" or "b" (or other letters). The grouping "ab" implies that groups A and B are not statistically different from each other.

Table 7.6: Grouping and EMMs of *senhor/senhora* pronunciation

Age	grouping	response scale EMM	EMM average
A	a	0.02	-3.75
B	b	0.29	-1.13
C	b	0.32	-1.20

(a) Grouping and EMMs of *senhor/senhora* pronunciation in age sociolects

Schooling	grouping	response scale EMM	EMM average
1	ab	0.10	-2.24
2	b	0.41	-0.88
3	a	0.05	-2.95

(b) Grouping and EMMs of *senhor/senhora* pronunciation in schooling sociolects

Elaborated by the author

Older people tend to utter the non-standard forms more than younger people. Considering that between 1950 and 1970 there was a rural exodus in the state of Minas Gerais (Portes and dos Santos, 2012), this result might be explained by the way people who had lived in rural areas spoke[6]. Even though people in the older age intervals may not have been born or were very little when the exodus happened, their immediate interpersonal contact was with people who spoke a rural sociolect. Nonetheless, to assure this statement, further investigation is necessary on the subject. Moreover, saying *senhor* or *senhora* is also a matter of hierarchy and politeness, which may not be followed by young people when they are talking to people of their age or younger.

Regarding schooling, people with a higher education level tend not to use the low-prestige forms as previously predicted. However, people from the `schooling 1` sociolect were not distinctive enough from the other schooling groups, besides uttering it less than people in the `schooling 2` sociolect. That might have happened because of the merge we did with the ages A and M. People from age M (under 18) do not have a schooling level higher than 1 in our classification and they have little participation in the recorded interactions (thus, not saying *senhor* or *senhora* and its variants), which may have caused the EMM go lower.

Another potential justification is the contexts in which the people from `schooling 2` were recorded. Some of `schooling 2` people were recorded in public contexts, such as their working environment most of the time. In addition, they had jobs in which hierarchy and politeness were demanded, notably for those working in customer services. Some of the jobs were office attendant, secretary, store attendant, personal trainer, and saleswoman. Having that in mind, people with a certain type of job from higher schooling levels would say more *senhor* or *senhora*. However, that may not explain the non-standard forms. It should be noted that phonetic reduction, especially with pronouns (*senhor* $>$ *sô*), are great indicators of the Mineiro dialect. Therefore, the higher number of non-standard forms in `schooling 2` sociolect may have a bias because of the type of job and conversation context of its speakers.

7.3.3 Non-standard plural marking in NPs

In standard PT-BR, the plural should be marked in all parts of a noun phrase (NP). Non-standard plural marking in a NP happens when at least one of its parts does not obey the NP head number. To illustrate, the phrase *os presente*[7] has non-standard plural marking because "os" is in plural while *presente* (gift) is singular, but its meaning is "more than one gift". Tables 7.7a, 7.7b, and 7.7c show the groupings and EMMs regarding this variable.

[6] the rural areas of Minas Gerais are commonly called *caipiras* (similar to "redneck") and it has a pejorative meaning.

[7] TER speaker from the file bfamcv02 said it.

Table 7.7: Grouping and EMMs in non-standard plural marking

Sex	grouping	response scale EMM	EMM average
Female	a	0.06	-2.69
Male	b	0.26	-1.33

(a) Grouping and EMMs in non-standard plural marking in sex sociolects

Age	grouping	response scale EMM	EMM average
A	ab	0.10	-2.274
B	a	0.05	-2.871
C	b	0.41	-0.874

(b) Grouping and EMMs in non-standard plural marking in age sociolects

Schooling	grouping	response scale EMM	EMM average
1	a	0.03	-3.39
2	ab	0.20	-1.59
3	b	0.35	-1.04

(c) Grouping and EMMs in non-standard plural marking in schooling sociolects

Elaborated by the author

Yule (2010) claims that women tend to use highly prestigious units whenever there is a linguistic spectrum from low-prestige to high-prestige. Although this type of statement can and must be challenged in regard to who these women are, in which culture/society they are inserted, and which other social factors are considered in the analysis, this was confirmed in this linguistic variable. Therefore, females tend to use less non-standard plural marking than male speakers as presented in Table 7.7a.

Furthermore, dlder people tend to utter more low-prestige plural markings in NPs, which was also found in Almeida (2018) about a rural region in Rio de Janeiro[8]. Furthermore, non-standard plural marking is a mark of the "rural sociolect" in Minas Gerais, whose speakers would concentrate on the older age interval.

As striking as it may sound, people who speak the `schooling 3` sociolect are more prone to say low-prestige plural marking than `schooling 1` sociolect speakers. More research is needed to explain why this happened, but a viable interpretation is the level of monitoring speakers may have activated when recorded. Knowing that they were being taped, speakers of `schooling 1` sociolect became more self-conscious of their speech and monitored closely how they spoke. In opposition to this, speakers of sociolects `schooling 2` and `schooling 3` may have lowered their speech monitoring particularly if they were recorded in family/private contexts, uttering some non-standard forms[9].

[8] Another Brazilian state.

[9] This may indicate a multicollinearity with context. More research is needed.

7.3.4 Foreign words

In PT-BR, there are some loanwords from other languages as customary in most languages. Some native speakers have these words in their mental lexicon (Evans et al., 2007). From another perspective, bilingual (or polyglot) speakers may insert a foreign language word that is not part of the PT-BR loanword set. Words such as "anche" from Italian and "because" from English were counted in this variable. Table 7.8 shows the groupings and EMMs regarding this foreign word use.

Table 7.8: Grouping and EMMs in foreign words

Schooling	Group	response scale EMM	EMM average
1	a	0.424	-0.858
2	b	1.735	0.551
3	b	2.723	1.002

Elaborated by the author

As anticipated, speakers from the `schooling 2` and `schooling 3` sociolects are inclined to utter more foreign words. Assis-Peterson and Cox (2007) discuss that proper English language teaching[10], for instance, is only available for higher social classes, which are mostly occupied by people who have higher education levels. Not only that, but highly schooled people have access to services and goods that are displayed[11] or negotiated[12] in foreign languages, which makes them used to incorporating foreign words in their daily lives. Furthermore, the fact that the highest two levels in the data did not differ from each other in the groupings may mean that their actual job may not interfere with the foreign word use, since the difference between them is working in a job that (does not) requires a university degree.

7.3.5 Non-standard verb agreement

Non-standard verb agreement happens when the number and grammatical person marking in the clause subject is not the same as in the clause verb. *Nós vai no supermercado* 'we-1PL go-3S to the supermarket' (we're going to the supermarket) is a very common example of non-standard verbal agreement. Table 7.9 displays the results for this variable.

Table 7.9: Grouping and EMMs in non-standard verbal agreements

Age	Grouping	response scale EMM	EMM average
A	a	5,64	1,73
B	a	6,24	1,83
C	b	10,53	2,35

Elaborated by the author

As seen in Table 7.9, age was the significant predictor for non-standard verb agreement, which was not forecasted previously since we assumed that schooling would be a relevant factor. The available evidence

[10] Here we will not explore the reasons for it.
[11] Some Brazilian shopping malls have used the word "sale" instead of its Portuguese equivalent.
[12] For instance: business meetings or visa background check.

suggests that speakers from `age A` and `age B` sociolects are significantly different from people from the `age C` sociolect. Moreover, the EMM values revealed that `age C` sociolect people tend to utter more non-standard verb agreement, while `age B` sociolect has an intermediate amount and `age A` sociolect has it in an even lesser quantity. Also, non-standard verb agreement is an indicator of the "rural dialect" in Minas Gerais, which is represented by the speakers in older age intervals.

7.3.6 Non-standard negation particle

Negation particles have been going through a process of phonetic reduction in PT-BR (Silva, 2016). Hence, this variable concerns the different pronunciation forms (for example *n'*, *nũ*) and syntactic constructions (for example *nũ ... não* — double negation). Table 7.10 reveals the groupings and EMMs regarding this variable. Here, there is a predictor interaction so the reading must be done in two directions. It was quite expected that the interaction would happen between age and schooling due to the tendency that as someone gets older, they also get more educated. Table 7.10 can be read in the following way: age A-schooling 1 has the letter "a", as well as age A-schooling 2. From another perspective, if we look at the age intervals in the schooling levels, we would have: schooling 1-age A has the letter "A", whereas schooling 1-age B has the letters "AB".

Table 7.10: Grouping and EMMs in non-standard negation particles

	Schooling								
	1			2			3		
Age	grouping	response scale EMM	EMM average	grouping	response scale EMM	EMM average	grouping	response scale EMM	EMM average
A	Aa	1.15	0.139	Aa	3.09	1.128	Aa	3.07	1.120
B	ABa	3.62	1.286	Aa	3,64	1.292	Aa	3.28	1.187
C	Bb	5.63	1.728	Aab	4.08	1.407	Aa	2.18	0.779

Elaborated by the author

Note: Lowercase letters must be read horizontally (from age to schooling), while uppercase letters must be read vertically (from schooling to age).

Let us analyze the schooling levels within age A. There is no statistical evidence that indicates differences between schooling levels. Likewise, there is no statistical evidence pointing to distinctions among schooling levels at age B. This fact may be a result of dialect accommodation, meaning that negation forms as used in these sociolects are acquired via dialect. On the other hand, within the oldest people's (`age C`) sociolect, `schooling 1` sociolects differ statistically. The results indicate that the least educated people's sociolect (`schooling 1`) differs from the most educated speaker's sociolect (`schooling 3`). Note that here we cannot say anything about `schooling 2`, as it is no different from `schooling 1` or `schooling 3`. Our interest here is the difference in `schooling` sociolects within the oldest people's (`age C`) sociolect, especially considering their EMMs. This result is intriguing because the opposite levels have relatively extreme EMM values, and the middle level is a transition between them. The highest schooling level sociolect presented the lowest EMM, which would mean speakers of this sociolect are likely to say fewer non-standard negation particles, while people from `schooling 1` or low-educated people's sociolect tend to enunciate it more.

The next analysis corresponds to observing the ages within each level of schooling. Observing schooling 1, `age A`, and `age C` are statistically different. Our hypothesis of having age as a significant predictor was

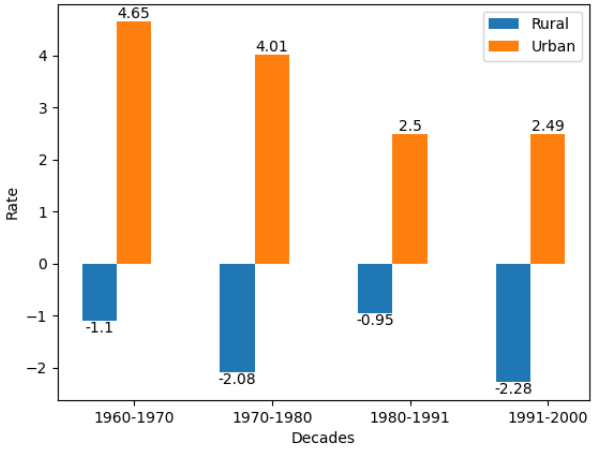
confirmed but not as expected. It was anticipated that younger people would utter more low-prestige forms, but the results show the contrary. The difference in ages when schooling is level 1 might be because of the speaker's education history. Someone from sociolect `age A` with schooling 1 is at the beginning of their educational experience so they are starting to have contact with the high prestige and standard variety of the language, whereas a person from sociolect `age C` with schooling level 1 is someone who stopped studying while they were supposed to be at school, so they might have lost contact with the standard variety and may speak only in the low-prestige variety.

Within level 2 of schooling, there is no statistical evidence that indicates a difference between ages. The same occurs at level 3 of schooling, as the results indicate that there is no significant difference. Thus, many groupings were the same, showing that there is significant similarity between the predictor groups. This may signal that this coherence between sociolects is because the non-standard forms of negation particles are somewhat spread out in Mineiro Portuguese due to the natural process of phonetic reduction that negation particles are going through. The same letter groupings across almost all sociolects in Table 7.10 indicate not only a coherence between sociolects but also might show the start of a change in progress. Only the older people differ from each other according to their schooling level, while younger people do not differ themselves based on negation particles. However, a more exhaustive examination of this subject matter is necessary.

Additionally, one should consider that the language difference between a highly educated senior and a lowly educated elderly person is bigger than between a highly schooled young individual and a less schooled young adult. The young adults in the corpus were born in the city, therefore, they acquired an "urban dialect", while the older people were citizens from the rural regions who migrated to the bigger cities, thus, they maintained their "rural dialect". As seen in Figure 7.3[13], in 1960-1970, the growth rate in the urban population was 4.65, while the rural population rate decreased 1.1. Looking at the other decades, the urban population growth rate remained higher than the rural population growth rate, which points out to the effects of the urbanization process in language as well.

[13] IBGE (2002) does not mention the scale in which the data is shown but it is the most credible Brazilian institution regarding the census. Hence, we are considering it anyway.

Figure 7.3: Geometric average population growth rate by household status in Minas Gerais



Adapted from IBGE (2002, v. 18, p. 14)

7.3.7 Non-standard verb conjugation

Examples of non-standard verb conjugation are *fazido 'feito'* (done) and *po' 'pode'* (can/be able to). Table 7.11 can be read in the following way: age A-schooling 1 has the letter "a", while age A-schooling 2 has the letter "b". Looking at the age intervals in the schooling levels, schooling 1-age A has the letter "A", whereas schooling 1-age B has the letters "AB".

Table 7.11: Grouping and EMMs in non-standard verb conjugations

	Schooling								
	1			2			3		
Age	grouping	response scale EMM	EMM average	grouping	response scale EMM	EMM average	grouping	response scale EMM	EMM average
A	Aa	1.79	0.584	Ab	5.20	1.649	Aab	5.20	1.649
B	ABa	5.57	1.717	Aa	4.84	1.576	Aa	3.78	1.331
C	Bb	6.72	1.905	Ab	9.01	2.199	Aa	2.71	0.99

Elaborated by the author

Note: Lowercase letters must be read horizontally (from age to schooling), while uppercase letters must be read vertically (from schooling to age).

Table 7.11 provides evidence that schooling and age were significant predictors for non-standard verb conjugation, which goes against our first hypothesis of it not being socially influenced. First, let us focus on the schooling levels within age A. `Schooling 1` is different from `schooling 2` but there is no significant

difference between `schooling 3` and the other sociolects.

Analyzing schooling at age B, there is no statistical evidence that indicates a difference between schooling levels. However, within age C, `schooling 1` is similar to `schooling 2`, and both are different from `schooling 3`. This result may be explained by the `age C-schooling 3` speaker's habit of using the standard variety of the language daily (since their work requires a higher education degree) making them use less low-prestige forms.

The following analysis relates to observing the ages within each schooling level. Within schooling 1, `age A` and `age C` are significantly different. It is also clear that we cannot say anything about level 2, as it does not differ from level 1 or 2. From this result, we may conclude that non-standard verb conjugations may differ in speakers of different ages with a lower education level. Again, this difference may be an aftermath of the "urbanization" as discussed in subsection 7.3.6. In other words, older people in schooling 1 may prefer the non-standard forms, in contrast, younger people with schooling 1 tend to use less low-prestige forms. Moreover, people in `schooling 2` and `schooling 3` do not differ concerning these forms regardless of age, showing that higher education has a great influence on non-standard verb conjugations.

Sociolinguistically speaking, the output in Table 7.11 can indicate that, as speakers get older and acquire more education, they are unlikely to say non-standard verb conjugations in their speech. Interestingly, people from `age A-schooling 3` utter more low-prestige conjugations than `age B-schooling 3` or `age C-schooling 3` speakers. One possible explanation is that younger speakers are likely to use certain speech patterns to fit in a group, regardless of their educational background. Moreover, among the 34 speakers from `age A-schooling 3` sociolect, 28 were recorded in family/private situations. The setting may have caused these speakers to have less speech monitoring and use more non-standard forms. It seems that `age A-schooling 3` sociolect has more low-prestige verbal forms because there are several occurrences of *o' 'olha'* (look) and reductions of *estar* (to be), such as *tá* (is) and *tava* (was)[14]. From another point of view, younger highly schooled individuals speaking more low-prestige elements could be attributed to free variation, indicating that predictors and models selected here cannot account for the whole of the observed variation (Weber and Kopf, 2023).

7.3.8 Apocopated diminutives

Examples of apocopated diminutives are *meninim 'menininho'* (little boy) and *filhotim 'filhotinho'* (little puppy). Tables 7.12a and 7.12b show the groupings and EMMs regarding this foreign word use.

Table 7.12: Grouping and EMMs in apocopated diminutives

Age	grouping	response scale EMM	EMM average
A	a	0.198	0.256
B	a	0.209	0.259
C	b	0.514	0.157

(a) Grouping and EMMs in apocopated diminutives in age sociolects

Schooling	grouping	response scale EMM	EMM average
1	b	0.508	-0.67
2	a	0.234	-1.45
3	a	0.179	-1.72

(b) Grouping and EMMs in apocopated diminutives in schooling sociolects

Elaborated by the author

[14] A closer look is needed to confirm it.

Table 7.13: Comparison between significant predictors in Mann-Whitney and Count Data Model results

	Age		**Sex**		**Schooling**	
Variables	Mann Whitney	Count data model	Mann Whitney	Count data model	Mann Whitney	Count data model
Reduced and articulated prepositions	yes	yes	-	-	-	-
Senhor/senhora pronunciation	-	yes	yes	-	-	yes
Non-standard plural marking in NPs	yes	yes	yes	yes	-	yes
Foreign words	-	-	-	-	yes	yes
Non-standard verb agreement	yes	yes	-	-	-	-
Non-standard negation particle	yes	yes	-	-	-	yes
Non-standard verb conjugation	yes	yes	-	-	-	yes
Pronominal phenomena	yes	-	-	-	-	-
Apheresis	yes	-	-	-	-	-
Apocopated diminutives	yes	yes	-	-	-	yes

Elaborated by the author

Firstly, our hypothesis that apocopated diminutives would be a dialect (not a sociolect) marker was denied by the results. The data presented in Table 7.12a indicate that older people are prone to use more apocopated diminutives, while younger people tend to use them less. Again, this might be the case for the "urbanization" of language, since this linguistic phenomenon is more related to rural area people.

Regarding schooling in Table 7.12b, less educated speakers tend to say more apocopated diminutives, whereas university degree holders are less likely to utter them. This may be explained by the fact that `schooling 1` speakers is composed of children (under 18 years old) and adults who do not have much schooling. Children are usually called by adults and elders by diminutives (*meninim* [little boy], *bonitim* [little beautiful boy], *homemzim* [little man], etc.), so they are just repeating what they hear from their immediate contact. On the other hand, less-educated adults tend to lean towards the low-prestige forms, because they have not had the school as a molder of language for the standard norm.

Having described the results from the count data models, the comparison between the non-parametric method and the count data model is discussed in the following section.

7.4 Comparison between count data and non-parametric models

In section 6.3.3, the Mann-Whitney test was explained. It is a pairwise comparison method that can check if there are significant differences between groups. Likewise, the CLD highlights the differences between groups, if any, considering the effects of significant variables. In this section, the results from both methods are contrasted. Table 7.13 displays such a comparison. Rhotacism and interjections/exclamations are not in the table because there was not a significant group difference in these linguistic variables in any of the tests.

As depicted in Table 7.13, the age predictor was significant for a great number of linguistic variables in

both tests. This finding is of particular interest because it may indicate that, depending on the linguistic unit, younger or older people are the ones dictating the use of the new forms in the variation scenario in Belo Horizonte. As an illustration, the non-standard negation particle results, presented in subsection 7.3.6, may likely suggest a change in progress due to the `age` sociolect differences.

Table 7.13 revealed that the CLD comparison had captured more complex groupings, while Mann-Whitney compared groups by single predictors. For instance, CLD was able to consider that age and schooling were predictors in interaction for non-standard negation particles. Moreover, CLD could show significant predictors not relevant in the Mann-Whitney testing, such as schooling in non-standard plural marking in NPs.

On the other hand, Mann-Whitney identified age as a significant group divider for pronominal phenomena and apheresis. It is plausible to hypothesize that age can influence the use of non-traditional pronominal and apheretic forms, especially considering that `age C` sociolect speakers are the most distinctive. However, we must remember that Mann-Whitney, EMM, and CLD use different statistics to calculate the group differences.

Finally, the combination of parametric models, EMM and CLD seem to be efficient methods to detect sociolectal differences. Not only the parameters on the parametric models can be edited but this sequence of methods can also account for interaction between predictors, allowing us to analyze more complex phenomena. Nevertheless, if a linguist does not have information on the data distribution or if parametric models do not work well, non-parametric testing is an adequate alternative. In the following section, a summary of the results will be provided.

7.5 Sociolinguistic profiling summary

Prior to delving into the outcomes outlined in the preceding sections and providing a concise overview of the findings, it is imperative to recollect the distribution of data across sociodemographic variables. Table 7.14 displays data distribution in addition to alphanumeric codes corresponding to each sociolect.

Table 7.14: Sample data distribution among the social variables in the percentage of words

Variable	**Code**	**Description**	**Percentage of tokens**
Age	A	$age \leq 25$	30
	B	$26 \leq age \leq 40$	30
	C	$age \geq 41$	40
Schooling	1	no education or up to 7 years	16
	2	with university major but job does not require it	43
	3	with university major and job does require it	41
Sex	Male	-	49
	Female	-	51

Elaborated by the author

Building upon the findings from the previous sections, Table 7.15 displays which social variables were significant and which sociolects had more or less of that linguistic element for each linguistic variable. In

the table, the symbol "—" means interaction between variables, so "age C — schooling 3" indicates "the sociolect of a person who is age C and schooling 3". Now, the symbol "+" indicates that more than one variable was significant but not in interaction; for instance, "male + age C" means the `male` and the `age C` sociolects separately.

Alternatively, the symbol "= ", as in age A=B, signifies that `age A` and `age B` lack distinctive features from each other; however, they represent sociolects characterized by the highest or lowest values of a specific variable. To illustrate, in reduced and articulated prepositions, `age A` and `age B` are not significantly different from each other and, compared to `age C`, they have the lowest amount of non-standard prepositions.

Table 7.15: Summary of results from CLD

Linguistic variable	**Significant social variable**	**Sociolect that has more**	**Sociolect that has less**
Reduced/articulated prepositions	age	age C	age A=B
Senhor/senhora pronunciation	age + schooling	age B=C + schooling 2	age A + schooling 3
Non-standard plural marking in NPs	sex + age + schooling	male + age C + schooling 3	female + age B + schooling 1
Foreign words	schooling	schooling 2=3	schooling 1
Non-standard verb agreement	age	age C	age A=B
Apocopated diminutives	age + schooling	age B + schooling 1	age C + schooling 3
Non-standard negation particle	age — schooling	-	-
Non-standard verb conjugation	age — schooling	-	-

Elaborated by the author

As seen in Table 7.15, most sociolects that have more or less of a certain linguistic element are the extremes, e.g., age A and age C, except for apocopated diminutives[15]. This result may indicate that people in the extremes tend to speak differently in a social variable category, whereas people in the middle categories sometimes behave like being from one extreme and sometimes like the other, which may suggest that middle categories are transition categories. A good example is `age B`. With apocopated diminutives, `age B` was grouped with `age A` but with *senhor/senhora* pronunciation, it was grouped with `age C`. Finally, the variables in which the interaction was significant do not have the "sociolect that has more/less" value because the two-way groupings (schooling-age and schooling-age) can have different letters. For example, `age C-schooling 1` and `schooling 1-age C` may have different groupings depending on the variable.

Our research findings unveil various linguistic trends within the analyzed sociolects. In terms of morphophonology, the prevalence of non-standard prepositions in age group C suggests a potential generational linguistic shift. From another perspective, the variation in the pronunciation of *senhor/senhora* reveals complex linguistic dynamics influenced by both age and schooling. The observed variation in age groups B=C and schooling 2 underscores the generational and educational factors in shaping pronunciation patterns, highlighting the intricate relationship between phonetic features and social variables.

In the realm of morphosyntactic, the consistent use of non-standard verb agreement in age C points to a possible age-related language movement, indicating that syntactic elements may likely undergo generational evolutions. Furthermore, with non-standard plural marking in NPs, the prevalence of the low-prestige forms

[15] However, age B's EMM was very close to age A's EMM

in `male`, `age C`, and `schooling 3` suggests a complex sociolinguistic scenario. Similarly, the frequency of apocopated diminutives in individuals from age group B and schooling 1 highlights potential sociolinguistic distinctions in morphosyntactic elements.

Furthermore, lexical features, including the incorporation of foreign words, show a clear association with education levels. The higher usage in sociolects with schooling 2=3 may imply that educational attainment plays a crucial role in lexical diversification.

Overall, the results indicate that different social groups are driving specific language variation processes. Age emerges as a key factor in morphosyntactic and phonetic variations, while sex and schooling influence morphosyntactic and lexical processes. The research provides a comprehensive insight into linguistic dynamics and the driving forces behind the observed variations in the studied sociolects. The next chapter moves on to consider the main conclusions that emerged during this research.

Chapter 8

Final words

The macro-social categories of class, gender, ethnicity, and age are abstractions over an infinite range of activities and conditions that constitute the lives in and for which people use variation.
Eckert (2016, p.3)

The previous assertion outlines what Sociolinguistics does in order to conduct studies. People do not experience social stratification whether in individual or social contexts, for instance, gender as binary or age as a biological marker (Eckert, 2016). However, sociolinguists do that to capture a piece of the linguistic "landscape" of a certain community and build speaker's sociolinguistic profiles. Grounded by such procedure, the present work constitutes quantitative and exploratory research on methods that can be used for sociolinguistic profiling. The main premise that guided us was lectal coherence, which is the main principle that leads to sociolect formation (Guy, 2013; Beaman and Guy, 2022). As argued by Guy (2013), the systematic co-variation of variable linguistic elements that share a social feature is known as lectal coherence. Having that in mind, a series of experiments were conducted so that a reasonable method of sociolinguistic profiling could be selected for the type of data we had.

Firstly, the foundations of Computational Sociolinguistics (CompSoc), discussed in Chapter 2, provided a satisfactory starting point given its interdisciplinary nature combining Sociolinguistics and Computing. One of these principles is data collection, which is paramount in any quantitative research. CompSoc is concerned with how data is available "in the wild" and how it will be compiled. Moreover, the implementation of computational techniques advances the field on the grounds that it systematizes a large amount of data and is quicker than manual work.

Chapter 3 reviewed some of the most commonly used terms in Sociolinguistics, e.g., speech community and community of practice, and assesses the feasibility of using the term "sociolect" instead. Sociolect is more accurate in the sense that it focuses on the language used by speakers, not on the membership itself. Supported by its precision, the notion of sociolects permits naming and analyzing more complex aspects of language variation, such as lectal contamination (Pijpops and Van de Velde, 2018) and lectal coherence (Guy, 2013; Beaman and Guy, 2022).

Chapter 4 listed the main basic principles that guide the Language into Act Theory (L-AcT) proposed by Cresti (2000). Besides being the theoretical groundwork for C-ORAL-BRASIL I compilation, the corpus-based pragmatic orientation in L-AcT serves as an effective resource for analyzing sociolinguistic patterns as individuals can **perform** their identity through speech acts and consequently make explicit, among many other things, their sociolects.

Chapter 5 described the corpus used in this study, C-ORAL-BRASIL I, a gold-standard speech corpus

made of informal texts. The present study was possible due to the great amount of information about the speakers and interactions in the metadata files. Furthermore, the transcription criteria, based on grammaticalization and lexicalization processes in PT-BR, offers a closer look into how speakers utter certain words, which was also important for this work.

Chapter 6 detailed the whole methodology used here: from getting to know the corpus and deciding the variables to preprocessing and modeling. The analyses were conducted with the Variation-Based Distance and Similarity Modeling (VADIS), Mann-Whitney test, and count data modeling (namely Poisson and Negative binomial), EMM, and CLD. It is important to highlight that the last method listed has not been used in Linguistics research as far as our search showed, which brings a methodological innovation to this study.

The results shown and explained in Chapter 7 helped us get the most suitable method for our data as well as an overview of Belo Horizonte sociolectal variation. VADIS did not perform well with our sample probably because of the amount of data and conversion from numerical to categorical values. On the other hand, the non-parametric testing was successful and provided a preview of what might happen in the parametric modeling experiment. Regarding parametric models, more complex relations between predictors were found in addition to outputting clearer and easy-to-read groupings.

Now going back to our research questions:

- What is the most suitable method for sociolect detection?
  The parametric models alongside the EMM and CLD calculations were the most convenient methods by reason of their parameter editing, consideration of complex relations between predictors, and user-friendly outputs. Of course, due to time and resource availability, several other methods were not accounted for in this research, which offers a new topic for future research.

- How can we distinguish Brazilian Portuguese speakers from different social groups through spontaneous speech transcriptions?
  In the last experiment, out of the twelve sociolinguistic variables closely related to lexicalization and grammaticalization processes in PT-BR, eight of them are distinctive enough between sociolects. This number indicates that an oral interaction, converted into written text through adequate transcription criteria, can showcase sociolinguistic patterns.

- How coherent are these groups' sociolects?
  The EMM and CLD outputs suggest that there is some degree of coherence in the inter-speaker level of a sociolect that differentiates one sociolect from another (for instance, age A and C groupings in apocopated diminutives – subsection 7.3.8) as well as in the inter-sociolect level that implies similarity between sociolects (for instance, age A and B groupings in apocopated diminutives - subsection 7.3.8).

According to the previous research questions and findings, this thesis has provided a deeper insight into the sociolectal landscape in Belo Horizonte. Prior to this study, it was difficult to make predictions about how different social groups spoke, because most papers and theses focused solely on one linguistic variable. This is the first study of substantial duration that examines associations between various linguistic variables and sociolects in PT-BR. Thus, the present study has gone some way toward enhancing our understanding of sociolectal variation and the method to analyze it.

After highlighting the importance of the results, it is critical to recognize the intrinsic limitations that define this thesis's scope and applicability. Accurately interpreting the results and offering a fair assessment of the research outcomes requires awareness of these limitations. Although C-ORAL-BRASIL I is a speech corpus, no complex phonetic variables (e.g., f0 contour) were incorporated into the analysis; only those that were marked in the transcriptions. Moreover, it was not possible to evaluate the speakers' occupations; therefore, it is unknown how their jobs have influenced their speech. Lastly, no other gold-standard speech

corpus of PT-BR is compiled and/or available, so there is an urgent need to first compile gold-standard speech corpora of other Brazilian dialects and then, repeat this study with the new corpora.

Furthermore, the use of the variable *sex* is quite questionable. As shown in the sociolinguistic tradition, most of the time, sex does not significantly influence a linguistic unit unless it is a phonetic phenomenon[1]. Considering that physiology is not the main force in variation, an option to overcome such an issue would be to have the speaker's gender (man, woman, non-binary, etc.). Since gender is how someone positions themselves psychologically, socially, culturally, and behaviorally, this social factor would be a better fit for sociolinguistic studies.

A natural progression of this work is to add variables taken from the audio files because phonetic variables would inevitably enhance the model power and would provide a more complete perspective on the sociolects in Belo Horizonte. Also, if the debate on lectal coherence is to be moved forward, a method that could consider all sociolects and linguistic variables at once needs to be developed.

In conclusion, this thesis holds the potential to contribute valuable insights to society. While our results require further testing and validation, they embody the initial groundwork for constructing a protocol for the sociolinguistic profiling of Brazilian Portuguese speakers. This endeavor, once fully developed, could offer practical applications, particularly in assisting law enforcement officers tasked with profiling suspects.

[1] Physiology is important, especially in phenomena that are affected by the formation of the vocal tract, such as the thickness of the vocal folds.

Bibliography

Almeida, E. M. (2018). A variação da concordância nominal num dialeto rural. Uma história de investigações sobre a língua portuguesa, page 77.

Alter, S. G. (2005). William Dwight Whitney and the science of language. Johns Hopkins University Press.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. Text & talk, 23(3):321–346.

Aristotle (2001). Política. Martin Claret, 6 edition.

Assis-Peterson, A. A. d. and Cox, M. I. P. (2007). Inglês em tempos de globalização: para além de bem e mal. Calidoscópio, 5(1):5–14.

Ataíde, C. A. d., da Silva, A. C. A., and Gomes, V. S. (2021). As estratégias acusativas de 2ª pessoa em cartas amorosas do sertão de pernambuco: um estudo pela via da sociolinguística histórica (the accusative strategies of 2nd person in romantic personal letters from the pernambuco: a study through the historical sociolinguistics). Estudos da Língua(gem), 19(4):157–182.

Austin, J. L. (1962). How to do things with words. Oxford University Press.

Avelar, L. L. M. R. N. d., Silva, M. R. d., and Almeida, T. P. d. (2013). As formas de negação com o item não no português falado em Santa Luzia: um estudo preliminar, page 27–36. Faculdade de Letras/UFMG.

Baker, P., Hardie, A., and McEnery, T. (2006). A glossary of Corpus Linguistics. Edinburgh University Press.

Barcala, M., Domínguez, E., Fernández, A., Rivas, R., Santalla, M. P., Vázquez, V., and Villapol, R. (2018). El corpus eslora de español oral: diseño, desarrollo y explotación. CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, 5(2):217–237.

Beaman, K. V. and Guy, G. R. (2022). The coherence of linguistic communities: Orderly heterogeneity and social meaning. Routledge.

Beers Fägersten, K. (2007). A sociolinguistic analysis of swear word offensiveness. Universität des Saarlands.

Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. Literary and linguistic computing, 8(4):243–257.

Biber, D. (2015). Corpus-Based and Corpus-Driven Analyses of Language Variation and Use. In The Oxford Handbook of Linguistic Analysis. Oxford University Press.

Biber, D. (2019). Text-linguistic approaches to register variation. Register Studies, 1(1):42–75.

Bick, E. (2000). "tagging speech data"-constraint grammar analysis of spoken portuguese. ODENSE WORKING PAPERS IN LANGUAGE AND COMMUNICATIONS, (1):11–30.

Bloom, D. E. (2004). GLOBALIZATION AND EDUCATION: An Economic Perspective, pages 56–77. University of California Press, 1 edition.

Bright, W. (1997). Notes. Language in Society, 26(3):469–470.

Brunato, D., Cimino, A., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2020). Profiling-UD: a tool for linguistic profiling of texts. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 7145–7151, Marseille, France. European Language Resources Association.

Bucholtz, M. (1999). "why be normal?": Language and identity practices in a community of nerd girls. Language in society, 28(2):203–223.

Burghoorn, M., de Boer, M. H. T., and Raaijmakers, S. (2020). Gender prediction using limited twitter data. CoRR, abs/2010.02005.

Bustan, F., Bire, J., Semiun, A., and Jehane, H. (2021). The function of plant metaphor as a symbol of unity for manggarai speech community. Academic Journal of Educational Sciences, 5(1):1–6.

Cabral, M. d. S. (2014). Um breve percurso sobre a história da linguística e suas influências na sociolinguística. Revista Acadêmica De Letras-Português, (2):85–93.

Camacho, R. G., Ceccantini, J., and Pereira, R. (2004). Norma culta e variedades linguísticas. Cadernos de formação: Língua portuguesa, pages 47–60.

Campbell-Kibler, K. (2010). The sociolinguistic variant as a carrier of social meaning. Language Variation and Change, 22(3):423–441.

Cantos Gómez, P. (2002). Do we need statistics when we have linguistics? DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 18:233–271.

Carlos, V. G. (2005). O comportamento linguístico na cidade de tarumã: um contraste entre pais e filhos. Entretextos, 5:31–57.

Carter, B. and Fenton, S. (2010). Not thinking ethnicity: A critique of the ethnicity paradigm in an over-ethnicised sociology. Journal for the Theory of Social Behaviour, 40(1):1–18.

Cavalcante, F. A. (2020). The information unit of topic: a crosslinguistic, statistical study based on spontaneous speech corpora. PhD thesis, Federal University of Minas Gerais.

Cavalcante, F. A. and Ramos, A. C. (2016). The american english spontaneous speech minicorpus. architecture and comparability. CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, 3(2):99–124.

Cheshire, J., Fox, S., Kerswill, P., and Torgersen, E. (2008). Ethnicity, friendship network and social practices as the motor of dialect change: Linguistic innovation in london. Sociolinguistica, pages 1–23.

Clyne, M. (2000). Lingua franca and ethnolects in europe and beyond. Sociolinguistica, 14(1):83–89.

Consortium, B. et al. (2007). British national corpus. Oxford Text Archive Core Collection.

Cooper, R. L. (1980). Sociolinguistic surveys: The state of the art. Applied Linguistics, 1(2):113–128.

Corder, S. P. (1971). Idiosyncratic dialects and error analysis. 9(2):147–160.

Correia, L. and Flores, C. (2021). Questionário sociolinguístico parental para famílias emigrantes bilingues (quesfeb): uma ferramenta de recolha de dados sociolinguísticos de crianças falantes de herança1. Revista de Estudos Linguísticos da Universidade do Porto-Vol, 75:102.

Coulthard, M., Johnson, A., Kredens, K., and Woolls, D. (2010). Plagiarism - Four forensic linguists' responses to suspected plagiarism, page 523–537. Routledge.

Cresti, E. (2000). Corpus di italiano parlato. Accademia della Crusca.

Cresti, E. (2018). The illocution-prosody relationship and the information pattern in spontaneous speech according to the language into act theory (l-act). Linguistik online, 88(1).

Cresti, E., Gregori, L., Moneglia, M., and Panunzi, A. (2018). The language into act theory: A pragmatic approach to speech in real-life. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), LB-ILR2018 and MMC2018 Joint Workshop: Language and Body in Real Life Multimodal Corpora, pages 20–25.

Cresti, E. and Moneglia, M. (2005). C-ORAL-ROM: Integrated reference corpora for spoken romance languages. John Benjamins Publishing Company.

David, P. D. (2005). O inglês no mundo: língua de prestígio. Trama, 1(2):209–215.

Davies, M. and Fuchs, R. (2015). Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). English World-Wide, 36(1):1–28.

de Souza Santos, E. and Silva, L. S. (2021). O efeito de variáveis sociais sobre o comportamento variável da preposição "para" em seabra-ba. UniLetras, 43:1–14.

Dutra, M. and Simioni, T. (2020). Análise da abordagem da variação na colocação pronominal em videoaulas do youtube. Falange Miúda - Revista de Estudos da Linguagem, 5(2):197–221.

Eckert, P. (2006). Communities of practice. Encyclopedia of language and linguistics, 2(2006):683–685.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. Annual Review of Anthropology, 41(1):87–100.

Eckert, P. (2016). Variation, meaning and social change. In Coupland, N., editor, Sociolinguistics: Theoretical Debates, page 68–85. Cambridge University Press.

Eckert, P. and McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. Annual Review of Anthropology, 21(1):461–488.

Eco, U. (1979). A theory of semiotics, volume 217. Indiana University Press.

Evans, V., Bergen, B. K., and Zinken, J. (2007). The cognitive linguistics enterprise: An overview. Equinox.

Evison, J. (2010). What are the basics of analysing a corpus? In The Routledge Handbook of Corpus Linguistics, pages 122–135. Routledge.

Fairclough, L. (2023). Towards methodological and theoretical synergies between forensic phonetics and third wave sociophonetics. Modern Languages Open.

FBI (2016). The unabomber.

Febriani, M. and Jufrizal, J. (2019). The comparative analysis of social dialect of Minangkabaunese used by employees and labourers in Padang. English Language and Literature, 8(3).

Finegan, E. and Biber, D. (1994). Register and social dialect variation: An integrated approach. volume 315, page 347. Oxford University Press New York.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Freitag, R. M. K. (2011). Entre norma e uso, fala e escrita: contribuições da sociolinguística à alfabetização. Nucleus, 8(1):1–10.

Gomes, C. A. (2012). Para além dos pacotes estatísticos varbrul/goldvarb e rbrul: qual a concepção de gramática? Revista do GELNE, 14(1/2):257–272.

Gomes, M. L. d. C., Dresch, A. A. G., and Almeida (2020). A Fonética Forense e a Comparação Forense de Locutor, page 27–54. Publicações IEL, 1 edition.

Gordon, M. J. (2006). Interview with william labov. Journal of English linguistics, 34(4):332–351.

Grafmiller, J. and Szmrecsanyi, B. (2018). Mapping out particle placement in englishes around the world: A study in comparative sociolinguistic analysis. Language Variation and Change, 30(3):385–412.

Grant, T. and MacLeod, N. (2020). Recursos e restrições na manutenção de identidades linguísticas: uma teoria de autoria. In de Almeida , D., Coulthard , M., and Sousa-Silva , R., editors, PERSPECTIVAS EM LINGUÍSTICA FORENSE, pages 76–94.

Gray, B. (2005). Informal learning in an online community of practice. International Journal of E-Learning amp; Distance Education / Revue internationale du e-learning et la formation à distance, 19(1).

Greenbaum, S. (1991). Ice: the international corpus of english. English Today, 7(4):3–7.

Gries, S. T. (2010). Useful statistics for corpus linguistics. A mosaic of corpus linguistics: Selected approaches, 66:269–291.

Gries, S. T. and Durrant, P. (2020). Analyzing co-occurrence data. In A Practical Handbook of Corpus Linguistics, pages 141–159. Springer, Switzerland.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping lexical dialect variation in british english using twitter. Frontiers in Artificial Intelligence, 2.

Guhin, J., Calarco, J. M., and Miller-Idriss, C. (2021). Whatever happened to socialization? Annual Review of Sociology, 47(1):109–129.

Gumperz, J. J. (2009). The speech community. Linguistic anthropology: A reader, 1(66):66–73.

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? Journal of Pragmatics, 52:63–71. Contexts of Use in Cognitive Sociolinguistics.

Guy, G. R. and Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. Lingua, 172(173):1–9.

Guy, G. R. and Zilles, A. (2007). Sociolingüística quantitativa: instrumental de análise.

Halliday, M. (1978). Language as Social Semiotic: The Social Interpretation of Language and Meaning. Arnold.

Haugen, E. (1966). Dialect, language, nation1. American Anthropologist, 68(4):922–935.

Hemalatha, I., Varma, G. S., and Govardhan, A. (2012). Preprocessing the informal text for efficient sentiment analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1(2):58–61.

Hernández-Campoy, J. M. (2014). Research methods in sociolinguistics. AILA review, 27(1):5–29.

Holmes, J. (2013). An introduction to sociolinguistics. Routledge, 4 edition.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hua, C., Qiufang, W., and Aijun, L. (2008). A learner corpus-esccl. In Proceedings of the Speech Prosody Conference, pages 155–158. Citeseer.

Hudson, R. A. (1996). Sociolinguistics. Cambridge university press, 2nd edition.

IBGE (2002). Tendências demográficas: Uma análise dos resultados do universo do censo demográfico 2000.

Ilbury, C. (2020). "sassy queens": Stylistic orthographic variation in twitter and the enregisterment of aave. Journal of Sociolinguistics, 24(2):245–264.

Izre'el, S., Mello, H., Panunzi, A., and Raso, T. (2020). In search of a basic unit of spoken language: Segmenting speech. In Izre'el, S., Mello, H., Panunzi, A., and Raso, T., editors, In Search of Basic Units of Spoken Language: A corpus-driven approach, volume 94 of Studies in Corpus Linguistics, page 1–32. John Benjamins Publishing Company, 1st edition.

Jacquemet, M. (2019). Beyond the speech community: On belonging to a multilingual, diasporic, and digital social network. Language Communication, 68:46–56.

Jessen, M. (2007). Speaker Classification in Forensic Phonetics and Acoustics, pages 180–204. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kendall, T. (2007). Enhancing sociolinguistic data collections: The north carolina sociolinguistic archive and analysis project. University of Pennsylvania Working Papers in Linguistics, 13(2):2.

Kendall, T. (2011). Corpora from a sociolinguistic perspective. Revista Brasileira de linguistica aplicada, 11(2):361–389.

Kendall, T. (2013). Data preservation and access. In Mallinson, C., Childs, B., and Herk, G. V., editors, Data Collection in Sociolinguistics, page 195–205. Routledge.

Kennedy, G. (1998). An introduction to corpus linguistics. Routledge.

Knight, D., Loizides, F., Neale, S., Anthony, L., and Spasić, I. (2020). Developing computational infrastructure for the corcencc corpus: The national corpus of contemporary welsh. Language Resources and Evaluation, 55(3):789–816.

Koerner, K. (1991). Toward a history of modern sociolinguistics. American Speech, 66(1):57–70.

Koven, M. (2011). Comparing stories told in sociolinguistic interviews and spontaneous conversation. Language in Society, 40(1):75–89.

Labov, W. (1963). The social motivation of a sound change. Word, 19(3):273–309.

Labov, W. (1973). Sociolinguistic patterns. Number 4. University of Pennsylvania Press.

Labov, W. (1986). The social stratification of (r) in new york city department stores. In Dialect and language variation, pages 304–329. Elsevier.

Labov, W. (2010). What is to be learned? LAUD.

Lacerda, M. L. (2021). Breve percurso histÓrico de abordagens linguÍsticas que antecedem e influenciam a constituiÇÃo da sociolinguÍstica variacionista. Revista do GEL, 18(1):68–100.

Lave, J. and Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge university press.

Leonard, R. A., Ford, J. E., and Christensen, T. K. (2017). Forensic linguistics: Applying the science of linguistics to issues of the law. Hofstra Law Review, 45(3):11.

Leuckert, S. (2020). Rethinking community in linguistics: Language and community in the digital age. In Jansen, B., editor, Rethinking Community through Transdisciplinary Research, pages 111–125. Springer International Publishing, Cham.

Levon, E. (2013). Ethnographic fieldwork. In Mallinson, C., Childs, B., and Van Herk, G., editors, Data collection in sociolinguistics: Methods and applications, pages 69–79. Routledge London and New York.

Lewandowski, M. (2010). Sociolects and registers–a contrastive analysis of two kinds of linguistic variation. Investigationes Linguisticae, 20:60–79.

Li, L. C., Grimshaw, J. M., Nielsen, C., Judd, M., Coyte, P. C., and Graham, I. D. (2009). Use of communities of practice in business and health care sectors: A systematic review. Implementation Science, 4(1):1–9.

Liberman, M. and Cieri, C. (1998). The creation, distribution and use of linguistic data: the case of the linguistic data consortium. In proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), pages 159–164.

Lieberson, S. (1980). Procedures for improving sociolinguistic surveys of language maintenance and language shift.

List, M. (2014). Sequence Comparison in Historical Linguistics. Dissertations in Language and Cognition. De Gruyter.

Maharana, K., Mondal, S., and Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1):91–99. International Conference on Intelligent Engineering Approach(ICIEA-2022).

Matus-Mendoza, M. (2002). The english lexical loan: A class marker. Journal of Hispanic Higher Education, 1(4):329–337.

Maxwell, A. (2018). When theory is a joke: The weinreich witticism in linguistics. Beiträge zur Geschichte der Sprachwissenschaft, 28(2):263–292.

Mayr, R., Roberts, L., and Morris, J. (2020). Can you tell by their english if they can speak welsh? accent perception in a language contact situation. International Journal of Bilingualism, 24(4):740–766.

McElreath, R. (2018). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

Mello, H. (2014). Methodological issues for spontaneous speech corpora compilation. The case of C-ORAL-BRASIL. Spoken Corpora and Linguistic Studies, pages 27–68.

Mello, H. (2023). Metodologia empírica de base computacional para a pesquisa sociolinguística: o encontro da sociolinguística e dos recursos computacionais, page 71–84. Pá de Palavra, 1 edition.

Mello, H., Raso, T., Mittmann, M. m., Vale, H. P., and Côrtes, P. O. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In Raso, T. and Mello, H., editors, C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informal, page 125–176. Editora UFMG.

Mesthrie, R. (2008). Sociolinguistics and sociology of language. In Spolsky, B. and Hult, F. M., editors, The Handbook of Educational Linguistics, page 66–82. Blackwell, 1 edition.

Mesthrie, R. (2011). Introduction: the sociolinguistic enterprise. In Mesthrie, R., editor, The Cambridge Handbook of Sociolinguistics, Cambridge Handbooks in Language and Linguistics, page 1–14. Cambridge University Press.

Meyerhoff, M. (2006). Linguistic change, sociohistorical context, and theory-building in variationist linguistics: new-dialect formation in new zealand. English Language and Linguistics, 10(1):173–194.

Miller, J. (2020). The bottom line: Are idioms used in english academic speech and writing? Journal of English for Academic Purposes, 43:100810.

Modiano, M. (1999). International english in the global village. English Today, 15(2):22–28.

Mollica, M. C., do Fundo, K. H., da Silva Gomes, L., Oliveira, M. d. S. P., and da Silva, R. F. (1998). Variação e função em aférese. Revista de Estudos da Linguagem, 7(2):71–87.

Moneglia, M. (2011). Spoken corpora and pragmatics. Revista Brasileira de Linguística Aplicada, 11:479–519.

Monte, A. (2019). A influência da escolaridade e do sexo/gênero no uso variável da concordância verbal de terceira pessoa do plural. Revista Diálogos, 7(1):89 – 104.

Monteiro, J. L. (2002). Para compreender Labov. Editora Vozes, Petropólis, Rio de Janeiro, 1 edition.

Morales Sánchez, D., Moreno, A., and Jiménez López, M. D. (2022). A white-box sociolinguistic model for gender detection. Applied Sciences, 12(5).

Morgan, M. (2004). Speech community. In Duranti, S., editor, A Companion to Linguistic Anthropology, pages 3–22. Basil Blackwell, Oxford.

Myers-Scotton, C. (2000). Code-switching as indexical of social negotiations. In Wei, L., editor, The Bilingualism Reader. Routledge, 1 edition.

Nance, C. (2022). Sound change or community change? the speech community in sound change studies: a case study of scottish gaelic. Linguistics Vanguard.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and De Jong, F. (2016). Computational sociolinguistics: A survey. Computational linguistics, 42(3):537–593.

Nguyen, D.-P. (2017). Text as social and cultural data: a computational perspective on variation in text. PhD thesis, University of Twente. SIKS dissertation series no. 2017-09.

Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part ii. Language and Linguistics Compass, 10(11):591–613.

Nycum, R. (2018). In defense of Valley Girl English. The Compass, 1(5).

Oliveira, A. A., dos Santos Valentim, M. A., and Santos, M. d. F. R. (2022). Rotacismo em alagoas: uma análise variacionista. Fórum Linguístico, 19(4):8671–8680.

Oliveira, A. J. d. and Santos, D. N. d. (2020). Concordância verbal no português brasileiro em maceió/al, brasil. Diversitas Journal, 5(4):3180–3195.

Oushiro, L. (2015). Dois pastel e um chopes: a concordância nominal e identidade(s) paulistana(s). Revista de Estudos da Linguagem, 23(2):389–424.

Pappas, P. A. and DePuy, V. (2004). An overview of non-parametric tests in sas: when, why, and how. Paper TU04. Duke Clinical Research Institute, Durham, pages 1–5.

Paquot, M. and Gries, S. T. (2021). A Practical Handbook of Corpus Linguistics. Springer Nature, Switzerland.

Pardede, J. A. and Ramadia, A. (2021). The ability to interact with schizophrenic patients through socialization group activity therapy. International Journal, 9(1):7.

Peres, E. P. (2006). O uso de Você, Ocê e Cê em Belo Horizonte: um estudo em tempo aparente e em tempo real. PhD thesis, Federal University of Minas Gerais, Belo Horizonte - Brazil. PhD in Linguistic Studies.

Perkins, R. C. (2021). The Application of Forensic Linguistics in Cybercrime Investigations. Policing: A Journal of Policy and Practice, 15(1):68–78.

Pezarino, M. X. V., do Prado da Silva, L. E., da Silva Rocha, E. P., and Luquetti, E. C. F. (2023). O dialeto caipira: uma análise sociolinguística da novela Pantanal. InterSciencePlace, 17(5).

Piepho, H.-P. (2004). An algorithm for a letter-based representation of all-pairwise comparisons. Journal of Computational and Graphical Statistics, 13(2):456–466.

Pijpops, D. and Van de Velde, F. (2018). Lectal contamination. how language-external variation becomes language-internal through language contact. In Variationist Linguistics meets Contact Linguistics.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. Journal of Sociolinguistics, 11(4):478–504.

Portes, É. A. and dos Santos, A. X. M. (2012). Aspectos da educação e do êxodo rural em minas gerais (1950-1970). Cadernos de Historia da Educacao, 11(2).

Pushchanka, A. (2017). The Portuguese of the state of Minas Gerais. Bachelor's thesis, Univerzita Karlova, Czech Republic. Bachelor's Degree in Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Ramos, J. M. (2002). A alternância entre "não" e "num" no dialeto mineiro: um caso de mudança linguística, page 155–167. Faculdade de Letras/UFMG.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. (2013). Overview of the author profiling task at pan 2013. In CLEF conference on multilingual and multimodal information access evaluation, pages 352–365. CELCT.

Raso, T. (2012a). O C-ORAL BRASIL e a Teoria da Língua em Ato. In Raso, T. and Mello, H., editors, C-ORAL-BRASIL-I: Corpus de referência do português brasileiro falado informal, page 91–123. Editora UFMG, Belo Horizonte.

Raso, T. (2012b). O corpus c-oral-brasil. In Raso, T. and Mello, H., editors, C-ORAL-BRASIL-I: Corpus de referência do português brasileiro falado informal, volume 1, page 91–123. Editora UFMG, Belo Horizonte, 1 edition.

Raso, T. (2013). Fala e escrita: meio, canal, consequências pragmáticas e linguísticas. Domínios de Lingu@gem, 7(2):12–46.

Raso, T. and Mello, H. (2010). The C-ORAL-BRASIL corpus. In Moneglia, M. and Panunzi, A., editors, Bootstrapping Information from Corpora in a Cross-Linguistic Perspective, page 193–213. Firenze University Press, Firenze.

Raso, T. and Mello, H. (2012). C-Oral Brasil I: Corpus de referencia do portugues brasileiro falado informal. Editora UFMG, Belo Horizonte, 1 edition.

Raso, T., Mello, H., Jesus, A. U., and de Deus, L. A. (2007). Uma aplicação da teoria da língua em ato ao português do brasil. Revista de Estudos da Linguagem, 15(2):147–166.

Raso, T., Vieira, M., et al. (2016). A description of dialogic units/discourse markers in spontaneous speech corpora based on phonetic parameters. Chimera: Romance corpora and linguistic studies, 3(2):221–49.

Reis, C., Antunes, L., and Pinha, V. (2011). Prosódia de declarativas e interrogativas totais no falar marianense e belorizontino no âmbito do projeto amper. In Anais do Congresso Brasileiro de Prosódia, volume 1.

Reppen, R. and Simpson-Vlach, R. (2019). Corpus linguistics. In An Introduction to Applied Linguistics, pages 91–108. Routledge, London, 3 edition.

Reynolds, W. N., Salter, W. J., Farber, R. M., Corley, C., Dowling, C. P., Beeman, W. O., Smith-Lovin, L., and Choi, J. N. (2013). Sociolect-based community detection. In 2013 IEEE International Conference on Intelligence and Security Informatics, pages 221–226. IEEE.

Rocha, B., Melo, E., Raso, T., and Mello, H. (2011). O pronome lembrete e a teoria da língua em ato: novas perspectivas de análise. In Anais do Congresso Brasileiro de Prosódia, volume 1.

Rocha, B. N. R. d. M. (2016). Uma metodologia empírica para a identificação e descrição de ilocuções e a sua aplicação para o estudo da Ordem em PB e Italiano. Master's dissertation, Federal University of Minas Gerais, Belo Horizonte - Brazil.

Rodrigues, G. F. d. S. (2019). Dimimutivo: um estereótipo linguístico do dialeto mineiro? In Anais do VII Simelp - VII Simpósio Mundial de Estudos de Língua Portuguesa, pages 1120–1127, São Paulo.

Sabaté-Dalmau, M. (2016). The englishisation of higher education in catalonia: a critical sociolinguistic ethnographic approach to the students' perspectives. Language, Culture and Curriculum, 29(3):263–285.

Sardinha, T. B. (2000). Corpus linguistics: history and problematization. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 16:323–367.

Schilling, N. and Marsters, A. (2015). Unmasking identity: Speaker profiling for forensic linguistic purposes. Annual Review of Applied Linguistics, 35:195–214.

Searle, S. R., Speed, F. M., and Milliken, G. A. (1980). Population marginal means in the linear model: an alternative to least squares means. The American Statistician, 34(4):216–221.

Shuy, R. W. (1990). A brief history of american sociolinguistics 1949-1989. Historiographia Linguistica, 17(1-2):183–209.

Shuy, R. W. (2001). Dare's role in linguistic profiling. DARE Newsletter, 1(3).

Silva, L. F. L. e. (2016). Negação verbal no Português Brasileiro:nbsp; Aspectos teórico-metodológicos em estudo baseado em corpus. Master's thesis. Master's Degree in Linguistics.

Silva, N. d. A. (2010). A preposição para e suas variantes no falar araguatinense. Master's thesis, Federal University of Paraíba, João Pessoa - Brazil. Master's in Linguistics.

Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology. Language Science Press, Berlin.

Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics: overlap and cross-pollination potential. Canadian Journal of Linguistics/Revue canadienne de linguistique, 62(4):685–701.

Szmrecsanyi, B., Grafmiller, J., and Rosseel, L. (2019). Variation-based distance and similarity modeling: A case study in world englishes. Frontiers in Artificial Intelligence, 2.

Tagnin, S. (2018). E a Linguística de Corpus vai desbravando novos horizontes... In Linguística de corpus : perspectivas, pages 11–15. Instituto de Letras - UFRGS, Porto Alegre, 1 edition.

Tavares, M. A. (2008). Conectores coordenativos: condicionamentos sociais em duas comunidades de fala brasileiras. Revista Linguítica, 4(1).

Teixeira, B. H. F., Barbosa, P. A., and Raso, T. (2018). Para a segmentação automática de fronteira na fala espontânea a partir de parâmetros prosódicos. In Finatto, M. J. B., Rebechi, R. R., Sarmento, S., and Bocorny, A. E. P., editors, Linguística de corpus : perspectivas, volume 1, page 425–446. Instituto de Letras - UFRGS, 1 edition.

Thomas, J. J. and Cook, K. A. (2005). Illuminating the path: the research and development agenda for visual analytics. IEEE Computer Society.

Timbane, A. A. (2014). Que português se fala em Moçambique? Uma análise sociolinguística da variedade em uso. Revista Vocábulo, 7.

Trudgill, P. (1999). The chaos before the order: New Zealand English and the second stage of new-dialect formation. In Jahr, E. H., editor, Language change: Advances in historical sociolinguistics, pages 197–207. Mouton de Gruyter Berlin, New York.

Trudgill, P. (2004). New-dialect formation: The inevitability of colonial Englishes. Oxford University Press, USA.

Vogelaer, G. D. (2013). Diasystem. De Gruyter. Available at: https://www.degruyter.com/database/WSK/entry/wsk_id_wsk_artikel_artikel_27339/html?lang=en.

Wagner, L., Clopper, C. G., and Pate, J. K. (2014). Children's perception of dialect variation. Journal of Child Language, 41(5):1062–1084.

Wang, M.-T., Henry, D. A., Smith, L. V., Huguley, J. P., and Guo, J. (2020). Parental ethnic-racial socialization practices and children of color's psychosocial and behavioral adjustment: A systematic review and meta-analysis. American Psychologist, 75(1):1.

Wardhaugh, R. (2006). An Introduction to Sociolinguistics. Blackwell, Oxford, 5 edition.

Wardhaugh, R. and Fuller, J. M. (2015). An introduction to sociolinguistics. John Wiley & Sons, Oxford, 7th edition.

Weber, T. and Kopf, K. (2023). Free variation, unexplained variation? Free Variation in Grammar: Empirical and theoretical approaches, 234:1.

Weinreich, U. (1954). Is a structural dialectology possible? Word, 10(2-3):388–400.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM, 9(1):36–45.

Wenger, E., McDermott, R., and Snyder, W. M. (2002). Seven principles for cultivating communities of practice. Cultivating Communities of Practice: a guide to managing knowledge, 4:1–19.

Wikipedia, the free encyclopedia (2019). Região metropolitana de belo horizonte. Available at: https://pt.wikipedia.org/wiki/Regi%C3%A3o_Metropolitana_de_Belo_Horizonte, accessed December 22, 2022.

Wikipedia, the free encyclopedia (2022). Minas gerais. Available at https://upload.wikimedia.org/wikipedia/commons/thumb/3/31/Minas_Gerais_in_Brazil.svg/300px-Minas_Gerais_in_Brazil.svg.png, accessed on August 18, 2022.

Winter, B. and Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. Language and Linguistics Compass, 15(11):e12439.

Wolfram, W. (2004). Social varieties of American English. In Finegan, E. and Rickford, J. R., editors, Language in the USA: Themes for the Twenty-first Century, page 58–75. Cambridge University Press, Cambridge.

Wolfram, W. (2017). Dialect in society. In Coulmas, F., editor, The Handbook of Sociolinguistics, chapter 7, pages 107–126. John Wiley Sons, Ltd, Oxford.

Yadlovska, O. S. (2022). Sociolect elements and genderlects in the modern Ukrainian language. Baltija Publishing, Riga.

Yi, M. (2021). A complete guide to box plots. Available at: https://chartio.com/learn/charts/box-plot-complete-guide/, accessed December 12, 2023.

Young, R. and Yandell, B. (1999). Top-down versus bottom-up analyses of interlanguage data: A reply to saito. Studies in Second Language Acquisition, 21(3):477–488.

Yule, G. (2010). The study of language. Cambridge University Press, Cambridge, 4th edition.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M., et al. (2009). Mixed effects models and extensions in ecology with R, volume 574. Springer, Laurel.

Zwicky, A. M. and Zwicky, A. D. (1982). Register as a dimension of linguistic variation. In Kittredge, R. and Lehrberger, J., editors, Sublanguage: Studies of Language in Restricted Semantic Domains. W. de Gruyter, New York.

Appendix A

Minas Gerais and Belo Horizonte maps

Figure A.1: Minas Gerais state map

Source: Wikipedia, the free encyclopedia (2022)

Figure A.2: Belo Horizonte metropolitan area map

Source: Wikipedia, the free encyclopedia (2019)

Appendix B

Parsed transcription extract

Extract from file bfamcv09.cg.pos

*GIL: não [não] <IN> // ela [ela] <PERS F 3S NOM> é [ser] <V PR 3S IND VFIN> Ornelas [Ornelas] <PROP M/F S> //
*CAM: Ornelas [Ornelas] <PROP M/F S> //
*ADR: ela [ela] <PERS F 3S NOM> vai [ir] <V PR 3S IND VFIN> me [eu] <PERS M/F 1S ACC> apresentar [apresentar] <V INF> a [a] <PRP> Danuza [Danuza] <PROP M S> // não [não] <ADV> importa [importar] <V PR 3S IND VFIN> como [como] <ADV> //
*GIL: credo [credo] <IN> //
*CAM: ô [ô] <IN> / é [ser] <V PR 3S IND VFIN> prima [primo] <ADJ F S> pobre [pobre] <ADJ M/F S> / fraga OALT flagra [flagrar] <V PR 3S IND VFIN> / aquelas [aquele] <DET F P> que [que] <SPEC M S> mandam [mandar] <V PR 3P IND VFIN> cartão [cartão] <N M S> em [em] <PRP> o [o] <DET M S> Natal [Natal] <PROP M S> e [e] <KC> nũ OALT não [não] <ADV> têm [ter] <V PR 3P IND VFIN> resposta [resposta] <N F S> // tipo [tipo] <N M S> isso [isso] <SPEC M S> //
*GIL: nada [nada] <SPEC M S> // a [o] <DET F S> família [família] <N F S> de [de] <PRP> ela [ela] <PERS F 3S NOM/PIV> é [ser] <V PR 3S IND VFIN> bem [bem] <ADV> estruturadinha [estruturar] <ADJ F S> aqui [aqui] <ADV> / viu [ver] <V PS 3S IND VFIN> //
*CAM: é [ser] <V PR 3S IND VFIN> //
*GIL: é [ser] <V PR 3S IND VFIN> //
*ADR: não [não] <IN> / só [só] <ADV> filma [filmar] <V IMP 2S VFIN> elas [elas] <PERS F 3P NOM> //

Appendix C

List of criteria not implemented

The phenomena below were not used in our analysis because the number of occurrences is not statistically significant, and; they are related to the topic discussed in the interaction. Moreover, the model would be very complex, which would not fit into a master's degree, and would have to incorporate variables related to vocal performance/acuity and interaction structure, which are not the purposes of this work.

1. Paralinguistic sounds without informational value
2. Paralinguistic sounds with informational value
3. Hesitations
4. Interrupted words
5. Onomatopoeias
6. Hesitations
7. Acronyms
8. Numbers
9. Non-transcribed and censored words
10. Retracting
11. Interruptions
12. Overlapping
13. "Mó" intensifier (meaning: "maior" in PT-BR and it is similar to "most" in English superlatives)

Appendix D

PALAVRAS tagset

Word class tagset

1. N - nouns
2. PROP - proper nouns (names)
3. SPEC - specifiers (non-inflecting pronouns that cannot be used as prenominals)
4. DET - determiners, can be used prenominals
5. PERS - personal pronouns
6. ADJ - adjectives
7. ADV - adverbs
8. V - verbs
9. NUM - numerals
10. PRP - prepositions
11. KS - subordinating conjunctions
12. KC - coordinating conjunctions
13. IN - interjections
14. EC - hyphen-separated prefix

Inflection tagset

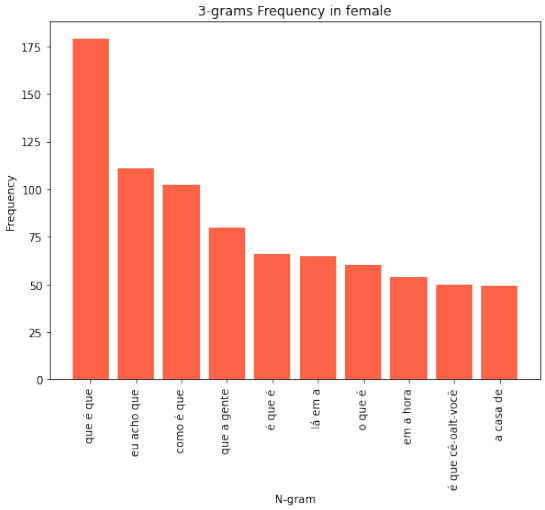
1. **Gender:** M (male), F (female), M/F [for: N', PROP', SPEC', DET, PERS, ADJ, V PCP, NUM]
2. **Number:** S (singular), P (plural), S/P [for: N, PROP', SPEC', DET, PERS, ADJ, V PCP, V VFIN, INF, NUM]

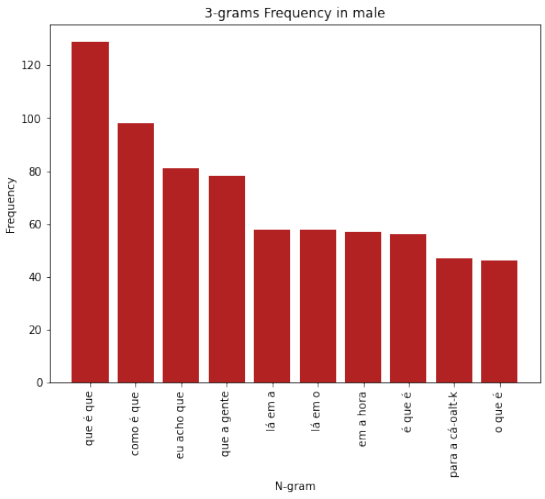
3. **Case:** NOM (nominative), ACC (accusative), DAT (dative), PIV (prepositive), ACC/DAT, NOM/PIV [for: PERS]
4. **Person:** 1 (first person), 2 (second person), 3 (third person), 1S, 1P, 2S, 2P, 3S, 3P, 1/3S, 0/1/3S [for: PERS, V VFIN, V INF]
5. **Tense:** PR (present tense), IMPF (imperfeito), PS (perfeito simples), MQP (mais-que-perfeito), FUT (futuro), COND (condicional) [for: V VFIN]
6. **Mood:** IND (indicative), SUBJ (subjunctive), IMP (imperative) [for: V VFIN]
7. **Finiteness:** VFIN (finite verb), INF (infinitive), PCP (participle), GER (gerund) [for: V]

There are syntactic rules tags, but they are not used in this research.

Appendix E

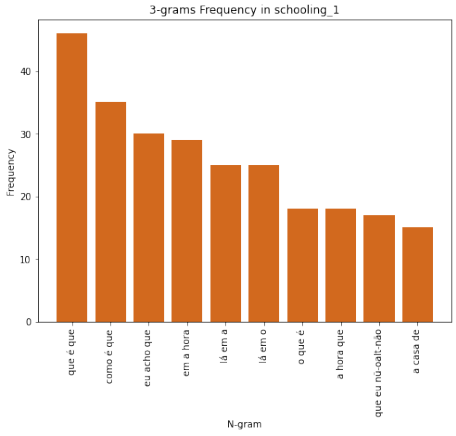
3-grams in sociolects divided by sex

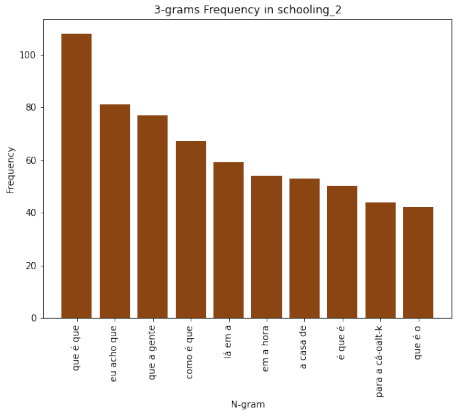


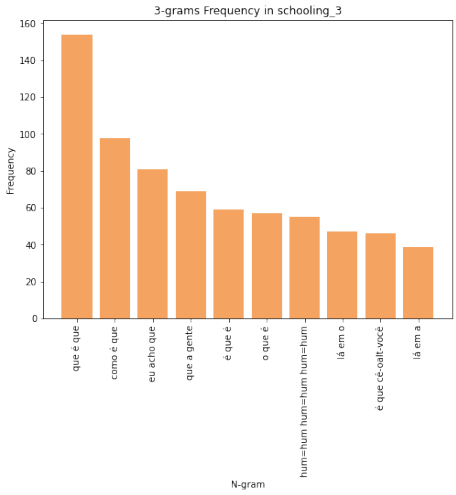


Appendix F

3-grams in sociolects divided by schooling

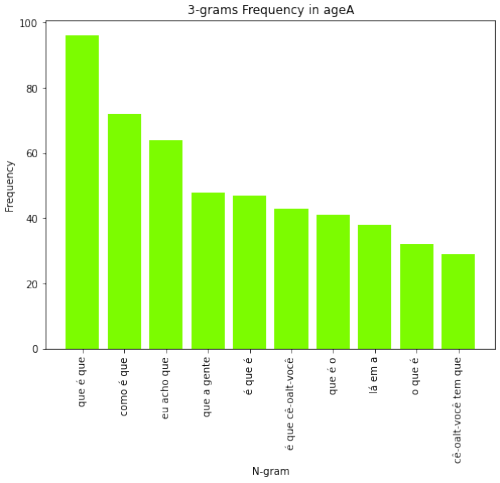


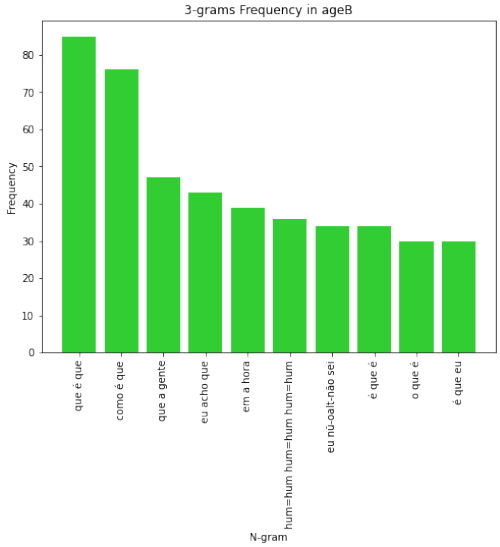


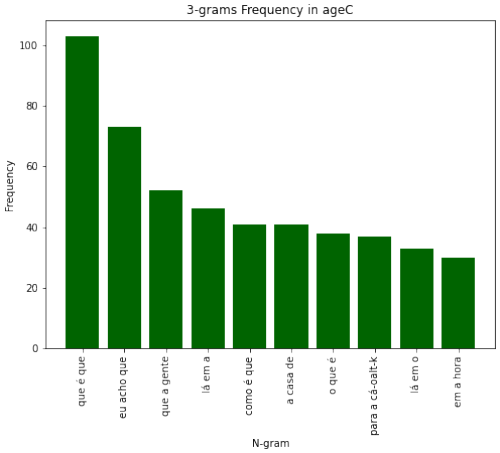


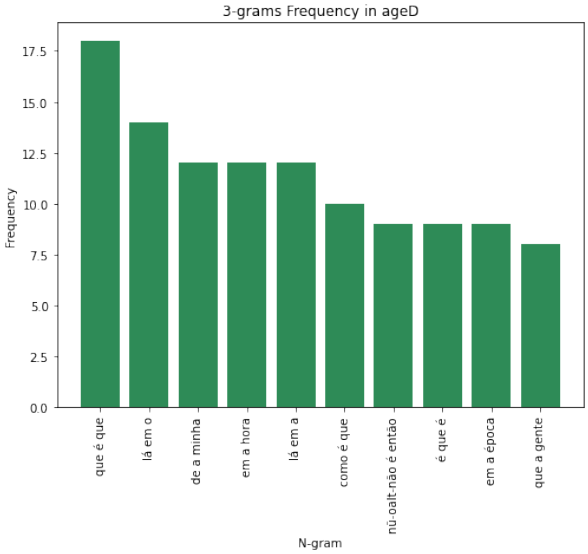
Appendix G

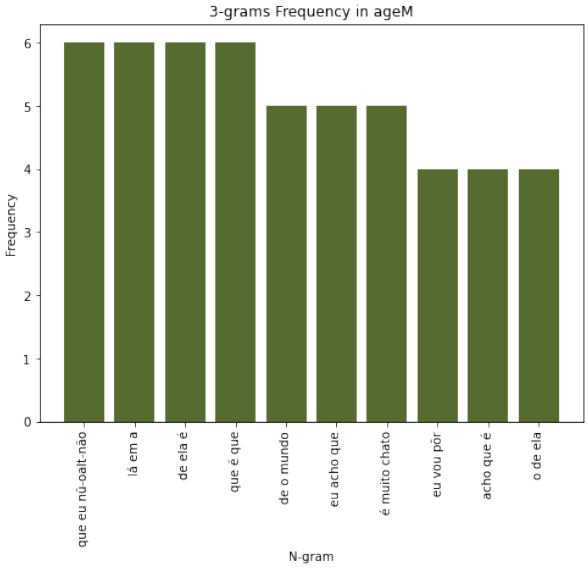
3-grams in sociolects divided by age








3-grams Frequency in ageD
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
que é que
lá em o
de a minha
em a hora
lá em a
como é que
nú·oalt·não é então
é que é
em a época
que a gente
N-gram


3-grams Frequency in ageM
Frequency
0
1
2
3
4
5
6
que eu nú·oalt·não
lá em a
de ela é
que é que
de o mundo
eu acho que
é muito chato
eu vou pôr
acho que é
o de ela
N-gram

Appendix H

Intercept coefficients

Table H.1: Intercept coefficients

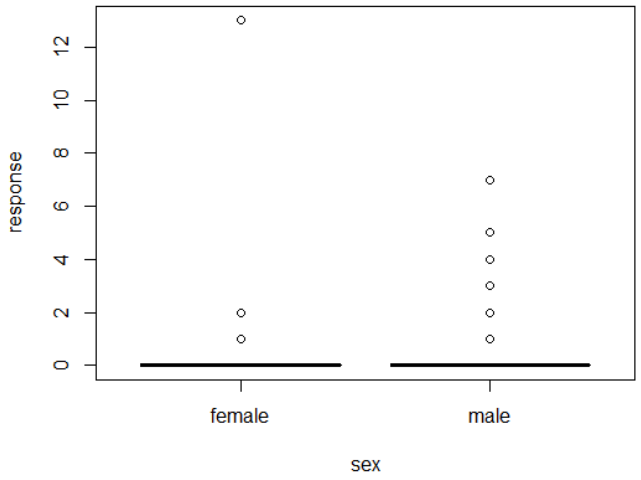
Model	Intercept
apheresis	2.5313
diminutives	-0.9906
foreign words	-0.6449
prepositions	1.9007
pronominal phenomena	1.0754
pronunciation of senhor/senhora	-4.8694
non-standard negation	0.2475
non-standard plurals	-4.3352
non-standard verb agreement	1.4773
non-standard verb conjugation	0.4226

Elaborated by the author

Appendix I

Box plot of non-standard plural in NPs according to sex

Figure I.1: Non-standard plural marking in NPs distribution considering sex



Elaborated by the author