CORPOR SYSTEM: *CORPORA* OF THE PORTUGUESE LANGUAGE AS SPOKEN IN SÃO PAULO

Zilda Maria Zapparoli

Universidade de São Paulo, CNPq, FAPESP, Brasil
zmz@usp.br

Abstract

This work briefly discusses the construction of the *Orthographic and Phonetic Information Databases of the Portuguese Language Spoken in the State of São Paulo (São Paulo City, Campinas, Itu) in a Relational Database System*. Informatics resources were used to store, process and analyze authentic oral language, and the Bases include orthographic and phonetic information about the Portuguese language as spoken in those areas of the state of São Paulo, organized, listed and stored taking into account linguistic and extralinguistic annotations. The results obtained can serve as a valuable aid, for example, in studies requiring automatic processing of the Portuguese language.

Index Terms: Linguistic Informatics, data processing technologies in Linguistic studies, CorPor project, relational database system, databanks of phonetic and orthographic information about the Portuguese language as spoken in São Paulo, electronic *corpora* of the Portuguese language as spoken in São Paulo

1. Introduction

This study is interdisciplinary *par excellence*, as it combines Linguistics and Informatics resources in the study of language in use, to store, process and analyze authentic oral language data. The work briefly discusses the construction of *Orthographic and Phonetic Information Databases (or DataBanks), Corpora and Lexicons of the Portuguese Language Spoken in the State of São Paulo (São Paulo City, Campinas and Itu*). The data were originally collected for a doctorate thesis (1980) and the bases generated at the time for mainframe computers, as in [1], have been made compatible with current operating systems.

The Bases are stored in the relational database format, which offers researchers the possibility of easy, reliable, rapid, and fully automatic access, for consultation, recovery and exploration of extensive and varied data, in the study of various aspects of language − phonetic, phonological, lexical, morphological, syntactic, textual and discursive.

This study, therefore, belongs in the field of *Linguistic Informatics*, drawing support from the various areas that share the belief in the positive results of the interaction between Linguistics and Informatics − it makes use of Informatics resources in Linguistics studies in order to build Information Bases that, in turn, can offer a contribution to the areas that use Linguistics in Computer Sciences, such as the automatic processing of the Portuguese language.

2. Methodological procedures

2.1. Structure of the oral language *corpus*

Speech samples produced by informants were collected between 1972 and 1973, totaling 54 hours of recordings that register dialogical interactions between documenter and 216 informants. Informants come from three cities in the state of São Paulo (São Paulo, Campinas and Itu), and are of both sexes, different ages and education level, and diverse socioeconomic backgrounds. In all, 432 dialogs were recorded, since there were two kinds of dialogic interaction with each informant: interviews and conversations.

The *Informants Distribution Diagram* presents the distribution of the informants in the categories (variables and their sub-levels), offering various possibilities for contrastive studies.

2.2. Constitution of the *corpus*: speech transcription for computational treatment

This is an annotated electronic *corpus* with the necessary information to identify linguistic variables (such as words, their position in the utterance as well as the position of the utterance in the discourse, orthographic and phonetic transcriptions, the kind of phonic juncture with the preceding and the subsequent words) and extralinguistic variables (such as region of origin, sex, education, age, socioeconomic background and the conditions in which the dialog was produced). There is an exclusive code for each lexical item, and about 180,000 occurrences.

The way in which information is codified and structured endows the Bases with the functionality that will permit the extraction of different *corpora* and lexicons.

2.3 Databank management system

The Information Bases are stored in a database system − *Firebird* − and the data structure follows the relational data model, so that the Bases contain linguistic and extralinguistic information about the various relations between the stored data, in this case a collection of orthographic and phonetic data of the Portuguese language as spoken in the State of São Paulo.

The environment used for programming was *Delphi*, produced by Borland Software Corporation, which uses Pascal Language with object-oriented extensions (Pascal Object), associated with resources of Structured Query Language (SQL) [2].

Besides research resources for access to the information on the Bases, the System includes resources for text production and for the edition of research results. For user access and research by means of SQL language commands, the *Orthographic-Phonetic Information Databases*, as well as *Corpora* and *Lexicons* (Dictionaries) generated from them, integrate the *CorPor System*, with each one of them constituting a module with its own records and fields.

3. The CorPor system: main components

3.1. Orthographic-Phonetic information databases of the Portuguese language as spoken in São Paulo

The *Orthographic-Phonetic Information Databases of the Portuguese Language as Spoken in São Paulo* bring information about each one of the 216 informants, organized according to the recording order and the annotation and structuring procedures adopted, i.e. the Bases bring lexical information organized according to the relations between linguistic and extralinguistic data. Table 1 brings an extract from the Databases.

3.2. Electronic *corpora* of the Portuguese language as spoken in São Paulo (textual databases)

Electronic Corpora of the Portuguese Language as Spoken in São Paulo (Textual DataBases) can be extracted from the *Orthographic-Phonetic Information Databases*, with various possibilities of exploration by linguistic analysis programs, as in [3], for use in different areas of language studies and related fields. It is possible to generate as many *corpora* as there are linguistic and extralinguistic variables annotated, with different combinatory possibilities. Below is an extract from the *corpus* of educated speakers of Portuguese from São Paulo (informants are from the city of São Paulo – *Paulistanos* − and have university degrees), with speech transcription. On Textual DataBases the punctuation codes were replaced by the corresponding marks.

Lexical Code: 1011111 – Informant from São Paulo (1), female (0), university degree (1), 25 to 29 years (11), upper class (1), stimulated response, dialogical interaction (1)

De profissional ou...

Nossa mãe! depende do dia —isso que é o problema, entende?— Eu optei um curso de complementação pedagógica e, agora, tem uns trabalhos, para apre/ apresentar, então, eu estou fazendo esses trabalhos: tem o de sociologia —para entregar— e um sobre o INCRA; tem uma tese que eu estou corrigindo a parte de português, toda parte de ortografia e construção —é de minha prima que tra/ trabalha no Butantã, sabe?; ela está fazendo uma tese sobre educação e saúde; também estou dando uma olhada na tese dela de manhã—. Tsu que mais que eu faço de manhã?... tempo de aulas, corrige-se provas; agora vai mudar —engano— vou mudar também; agora, de manhã, vou dar aula no Mackenzie; à tarde, venho para cá —varia—.

3.3. Orthographic-Phonetic frequency lexicon of the Portuguese language as spoken in São Paulo

The *Frequency Lexicon* was extracted from the complete version of the *corpus;* for each word, it presents the orthographic transcription (column 3), the corresponding phonetic transcriptions, with and without syllabic separation (columns 5 and 4 respectively), frequency of phonetic unit annotation (column 2) and cumulative frequency of orthographic unit (column 1), as in the sample presented in Table 2.

3.4. Inter-word coarticulation and phonetic liaison lexicon of the Portuguese language as spoken in São Paulo

The *Inter-word Coarticulation and Phonetic Liaison Lexicon*, also extracted from the *Orthographic-Phonetic Information Databases of the Portuguese Language as Spoken in São Paulo*, includes the phonetic liaison category (column 1), the accentual combination in inter-word coarticulation (column 2), the lexical-syllabic phonetic transcription of phonetic liaison occurrences, that is, phonic liaisons taking place between two or more words (columns 3, 4, 5 and 6), with the corresponding orthographic transcription (columns 7, 8, 9 and 10), as shown in the sample in Table 3.

4. Conclusions

In tune with the latest tendencies in language studies and cutting-edge technologies, this research can offer valuable contributions; (1) by meeting the demand, in Brazil, for electronic speech transcription *corpora* with phonetic transcriptions; (2) by permitting scientific and technological interchange and enriching the interaction between the exact sciences and language sciences; (3) within the scope of Linguistics, for research based on *corpora* and the utilization of computer technologies in studies of language in use; (4) at the interface between Linguistics and Informatics, by offering linguistic information knowledge for the development, testing and evaluation of speech processing systems for the Brazilian variety of the Portuguese language − recognition and synthesis −, one of the most complex areas in Natural Language Processing.

5. Acknowledgements

I would like to thank Manoel Vidal Castro Melo for his support and orientation in the analysis and programming for the development of the mainframe system and Edenis Gois Cavalcanti, for the creation of the system for use in PCs.

6. References

[1] Z. M. Zapparoli Castro Melo, "Análise do comportamento fonológico da juntura intervocabular no português do Brasil (variante paulista). Uma pesquisa linguística com tratamento computacional", Ph.D. dissertation, Universidade de São Paulo, São Paulo, SP, Brasil, 1980.

[2] C. Szyperski, *Component Software: Beyond Object-Oriented Programming*. Boston: Addison-Wesley, 1998.

[3] Z. M. Zapparoli, A. Camlong, *Do Léxico ao Discurso pela Informática*. São Paulo: EDUSP/FAPESP, 2002, 256 p. + CD-ROM.

[4] International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, 1999.

Figure and tables

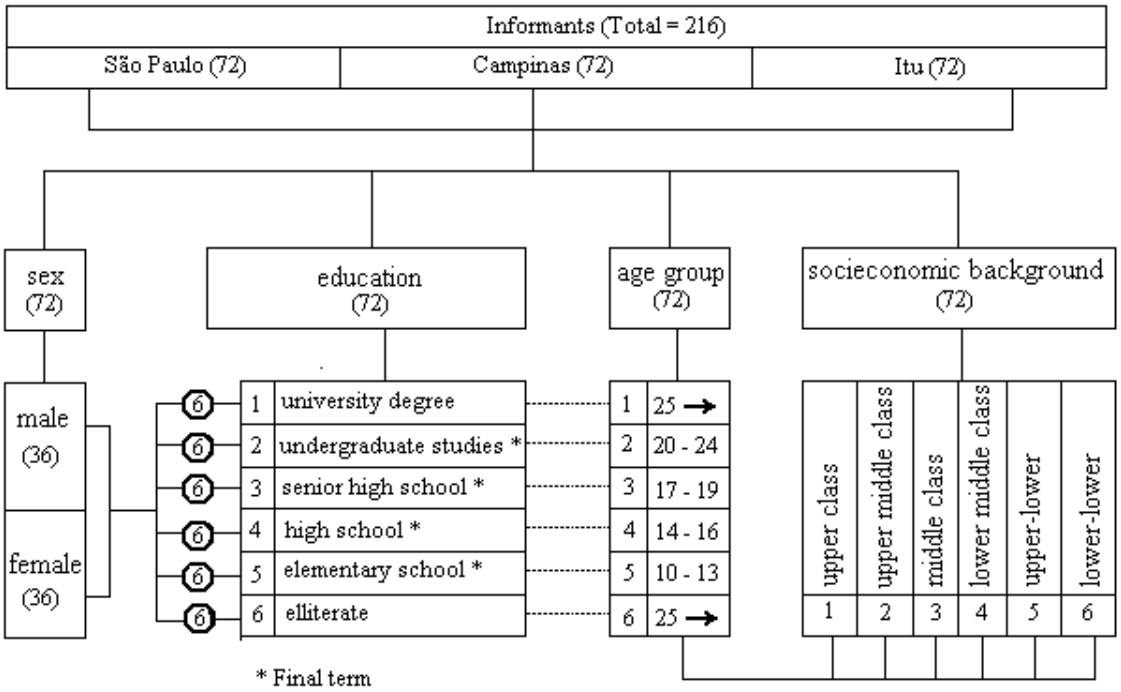

Figure 1. *Informants distribution diagram*

Table 1. *Orthographic-Phonetic information databases of the Portuguese language as spoken in São Paulo*

Key[1]	*Lexical Code*[2]	*Obs.*[3]	*Orthographic Transcription*[4]	*Punct.*[5]	*IS L*[6]	*Phonetic Transcription* [7]	*ES L/P*[8]
1	10111100101001		já			'ʒa	101
2	10111100101002		viajei		101	vi a 'ʒej	38
3	10111100101003		um		38	ʃũ	101
4	10111100101004		bocadinho	1	101	bo ka 'di ˌnu	1
5	10111100201001		eu			'ew	101
6	10111100201002		fui		101	'fuj	101
7	10111100201003		pela		101	pe l	5
8	10111100201004	6	Associação		5	a so sja 'sãw̃	101
9	10111100201005		dos		101	dus	100
10	10111100201006		Professores		100	pɾo fe 'so ɾiz	101
11	10111100201007		de		101	di	101
12	10111100201008	6	Francês	4	101	frã 'sej	32
13	10111100201009		sabe	7	32	'sa bi	1
14	10111100301001		olha	4		'ɔ ʎa	37
15	10111100301002		o		37	u	101
16	10111100301003		curso		101	'kuɾ sw̃	15
17	10111100301004		em		15	ĩ	101
18	10111100301005		si		101	'si	1
19	10111100301006		não	3		'nũ	1
20	10111100301007		não			'nũ	101

[1] Order.
[2] Lexical item identification code − informant, type of dialogue, discourse, utterance and word
[3] Code for morpho-syntactic deviations, acronyms, proper names, foreign words
[4] Orthographic transcription
[5] Punctuation code
[6] Initial syllable liaison code
[7] Phonetic transcription [4]
[8] End syllable liaison code / real pause

Tabela 2. *Orthographic-Phonetic frequency lexicon of the Portuguese language as spoken in São Paulo*

Accu. Freq.of Ortho. Trans. [1]	***Phone. Trans. Freq.*** [2]	***Orthographic Transcrption*** [3]	***Phonetic Transcrption*** [4]	***Phonetic Transcrption / Syllable*** [5]
2	2	abacate	aba'kaʦi	a ba 'ka ʦi
1	1	abacaxi	abaka'ʃi	a ba ka 'ʃi
1	1	abacaxis	jabaka'ʃiz	ja ba ka 'ʃiz
1	1	abaixo	a'baʝu	a 'ba ʝu
2	1	abaixo	a'bajʃu	a 'baj ʃu
1	1	abalado	aba'ladu	a ba 'la du
1	1	abandonar	abãdo'na	a bã do 'na
2	2	abandonei	abãdo'nej	a bã do 'nej
1	1	abandonou	abãdo'no	a bã do 'no
1	1	abatida	aba'ʦida	a ba 'ʦi da
3	3	aberta	a'bɛɾta	a 'bɛɾ ta
4	1	aberta	a'bɛrta	a 'bɛr ta
1	1	abertas	a'bɛɾtas	a 'bɛɾ tas
1	1	aberto	a'bɛʈtu	a 'bɛʈ tu
2	1	aberto	a'bɛʁtw	a 'bɛʁ tw
4	2	aberto	a'bɛɾtu	a 'bɛɾ tu
6	2	aberto	a'bɛrtu	a 'bɛr tu
7	1	aberto	ja'bɛɾt	ja 'bɛɾ t
8	1	aberto	ja'bɛɾtu	ja 'bɛɾ tu

[1] Orthographic transcription accumulated frequency
[2] Phonetic transcription frequency
[3] Lexical item orthographic transcription
[4] Lexical item phonetic transcription without syllabic division [4]
[5] Lexical-syllabic phonetic transcription [4]

Table 3. *Inter-word liaison lexicon of the Portuguese language as spoken in São Paulo*

Liaison [1]	***Stress*** [2]	***Phon 1*** [3]	***Phon2*** [4]	***Phon3*** [5]	***Phon4*** [6]	***Ortho.1*** [7]	***Ortho. 2*** [8]	***Ortho.3*** [9]	***Ortho.4*** [10]
101	TA	'ʒa	vi a 'ʒej			já	viajei		
101	AA	ʤũ	bo ka 'dᵢ ɲu			um	bocadinho		
101	TT	'ew	'fuj			eu	fui		
5	AA	pe l	a so sja 'sãw̃			pela	Associação		
100	AA	dus	pɾo fe 'so ɾiz			dos	Professores		
101	AA	di	fɾã 'sej			de	Francês		
37	AA	'ɔ ʎa	u			olha	o		
15	AA	'kuɾ sw̃	ĩ			curso	em		
101	TT	'nũ	'sej			não	sei		
33	ATA	sj	'ɛ	w		se	é	o	
15	AA	'kuɾ sw̃	ĩ			curso	em		
101	TA	'si	si			si	se		
2	AA	'va lj	a			vale	a		
17	AA	'pᵉ n	ĩ 'tẽʃ ʤi			pena	entende		
27	AT	maj z	'ew			mas	eu		
101	AA	'wa ʃu	ki			acho	que		

[1] Inter-word coarticulatory category
[2] Inter-word syllable stress – combinatorial stress in inter-word context (T = stressed syllable; A = unstressed syllable)
[3] Phonetic transcription of word 1 in word sequency [4]
[4] Phonetic transcription of word 2 in word sequency [4]
[5] Phonetic transcription of word 3 in word sequency [4]
[6] Phonetic transcription of word 4 in word sequency [4]
[7] Orthographic transcription of word 1 in word sequency
[8] Orthographic transcription of word 2 in word sequency
[9] Orthographic transcription of word 3 in word sequency
[10] Orthographic transcription of word 4 in word sequency