Eckhard Bick (Odense)
Marcelo Módolo (São Paulo)

Letters and Editorials: An annotated corpus of 19th century Brazilian Portuguese

This paper describes the structural design, philological editing and grammatical annotation of a 175.000 word corpus of 19th century Brazilian Portuguese (PB), consisting of newspaper leaders and letters to the editor (Barbosa / Lopes 2002). The corpus was compiled as part of a planned diachronic *megacorpus,* within the *Projeto Para a História do Português Brasileiro,* a joint effort of researchers in eight Brazilian states, and has already been used in ongoing efforts to describe the social history of PB, as well as the evolution of grammatical changes in the language. The material was grammatically annotated with the PALAVRAS parser (Bick 2000), a robust Constraint Grammar system using word based tags for word class, lemma, inflexion and syntactic function, as well as directional dependency markers at both group and clause level. Historical spelling variations were treated by systematic orthographical filtering, lexicon additions and morphological heuristics. Preliminary test runs on small parts of the corpus indicate error rates of 2.5% for word class and 7–8% for syntactic function. Though the system performs twice as well on current texts, these numbers suggest that a "modern" CG-parser can achieve results on 19th century data with only minor alterations.

1. The corpus project

The *Projeto Para a História do Português Brasileiro* (PHPB) was created through the initiative of Prof. Dr. Ataliba Teixeira de Castilho (University of São Paulo / USP). On October 8th, 1996, Dr. Castilho issued a call for researchers oriented towards historic-diachronic studies of Brazilian Portuguese to promote a seminar with the purpose of (i) introducing research activities in that area and (ii) verifying the possibility of integration in a national collective project. The *1st Seminar on the History of Brazilian Portuguese* was held at USP in April, 1997. By that time, the outline and basic principles of a mega-project had been established, which would eventually consist of three stages, to be developed in parallel: 1.) Compilation of the Diachronic corpus of Brazilian Portuguese, 2.) Social history of Brazilian Portuguese and 3.) Grammatical change of Brazilian Portuguese. So far researchers have opted for one of the following theoretical models: Variation and Change, Generativism (Principles and Parameters), Functionalism (grammaticalization processes).

Newspaper leaders and letters to the editor, one of the corpora of the first phase of this project, consists of ca. 175.000 words, totaling 568 letters published in Brazilian periodicals from the 19th century. This corpus was compiled by the regional teams of Pernambuco (Coordinator: Marlos de Barros Pessoa), Bahia (Coordinator: Rosa Virginia Mattos e Silva), Rio de Janeiro (Coordinator: Dinah Callou), Minas Gerais (Coordinator: Jânia Ramos), São Paulo (Coordinator: Ataliba Teixeira de Castilho) and Paraná (Coordinator: Sônia Cyrino), using documents from the respective states. Finally, the material was organized and edited by Afrânio Barbosa and Célia Lopes, with the title "*Críticas, queixumes e bajulações na imprensa brasileira do séc. XIX: cartas de leitores e de redatores*".

All documents carry the following ID-data: state / city, type of text (newspaper leaders or letters to the editor), title of periodical, date / edition and source / quota.

The transcriptions of the texts were conservative in the sense of philological transcription, following the directions of *2nd Seminar on the History of Brazilian Portuguese*, compiled in Megale (2001: 553–555). Thus, some graphic signals have been used:

> [ ] indicates the absence of one letter / syllable in the word or a word of a statement. E.g. a[c]eita-se pedidos; para poder continua [ ] vender; para o verão e arti[ ]s de modas;
>
> [[ ]] indicates that letter / syllable / word are repeated. E.g dirigi[[gi]]ram; dinheiro [[a dinheiro]];
>
> [illegible] and similar marks were used in cases like the following: assim ao modo de [illegible] que há tempos; faz [omission] sciente ao Público; vende-se huma propriedade [corroded] de tres andares; de profição agronomo. [space] com bôas referências;
>
> | in the majority of cases, the simple bar indicates a line break;
>
> || indicates a paragraph break;
>
> *italics* indicate reconstruction of abbreviations. E.g. S*enho*r, r*éi*s, n*úmer*o, Exc*elentíssi*mo, N*ossa* S*enhora*, *Re*V*erendíssi*mm*a.*

The corpus also contains texts from the areas of public administration (17th, 18th and 19th centuries), private administration (16th, 18th and 19th centuries), private documents (18th, 19th and 20th centuries), literary texts (18th and 19th centuries) and newspaper advertisements, already compiled by Guedes / Berlinck (2000; 19th century). However, the material has to be largely extended and diversified, so that the results of our analyses will become representative for the history the Brazilian Portuguese. Meanwhile, we would like to quote a metaphor used by Prof. Castilho, when referring to the project: "at the same time that we are constructing a boat, we are sailing on it".

The next step towards an extension of the corpus will be the compilation of a collection of private letters, which are only scarcely represented in public archives.

2. Automatic grammatical annotation

Grammatical annotation allows quantitative or qualitative linguistic research not easily undertaken on raw-text corpora. For instance, lemmatization allows a historical corpus to be quantified in terms of lexical distribution and orthographical tendencies. Likewise, annotation with syntactic categories will allow to search for structural / word order patterns, valency patterns or selection restrictions.

The corpus was annotated sequentially at two levels: a) word based form and function tags and b) syntactic trees. Out of a total of 175.000 words, about 4% (15 texts from Bahia, 1830's) were manually revised at level (a).

2.1 Methodological framework

The methodological framework used for annotating the corpus was a multilevel Constraint Grammar parser, PALAVRAS (Bick 2000 and <http://beta.visl.sdu.dk>). This robust parser primarily targets modern written Brazilian and European Portuguese, but has been successfully used for non-standard input before. Apart from dialectal and speech data, one earlier experiment with historical data has been performed in connection with the *Tycho Brahe* project (<http://www.ime.usp.br/~tycho/corpus/>; cf. Britto / Finger / Galves 2002). For the current project, the parser was enhanced with a data-driven orthographical filtering module and pattern matching heuristics (cf. fig. 1 on the following page).

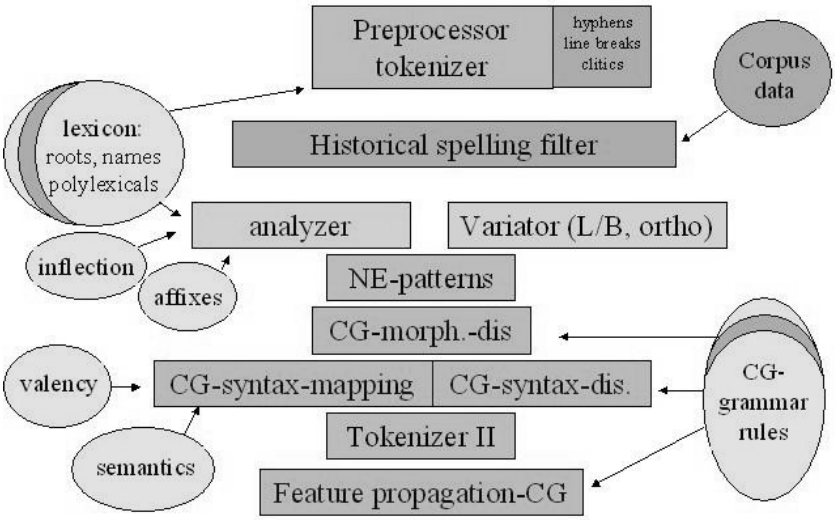

Fig. 1. System architecture of PALAVRAS

In word based CG annotation, each token receives modular tags for PoS, inflexion and syntactic function (@-tags). The latter include directional dependency arrows facilitating the construction of syntactic trees. Thus, @SUBJ> is a subject to the left of its verb, @N< is a postnominal dependent:

```
A                [o] <artd> DET F S               @>N
expedição        [expedição] N F S              @SUBJ>
contra           [contra] PRP                     @N<
o                [o] <artd> DET M S                   @>N
Mexico           [México] PROP M S                  @P<
sahio ALT saiu   [sair] <fmc> V PS 3S IND       @FMV
a                [a] PRP                        @<ADVL
5                [5] <card> NUM M/F P             @P<
de               [de] PRP                           @N<
Julho            [julho] N M S                        @P<
```

In tree-format, each node receives an upper case function tag and a lower case form tag, separated by a colon. Constituent-daughterhood is – in source format – expressed by '=' indentation. Thus, the three daughters of the finite relative clause *'que a compõe'* are indented one level deeper than the mother node, N<:fcl.

```
STA:fcl
SUBJ:np
=>N:art('o' <artd> M P)   os
=H:n('soldado' M P) soldados
```

```
=N<:fcl
==SUBJ:pron-indp('que' <rel> M S)       que
==ACC:pron-pers('ela' F 3S ACC)  a
==P:v-fin('compor' PR 3S IND)    compõe
P:v-fin('ser' PR 3P IND)  são
SC:np
=>N:pron-det('vosso' <poss 2P> M P)     vossos
=H:n('irmão' M P)   irmãos
```

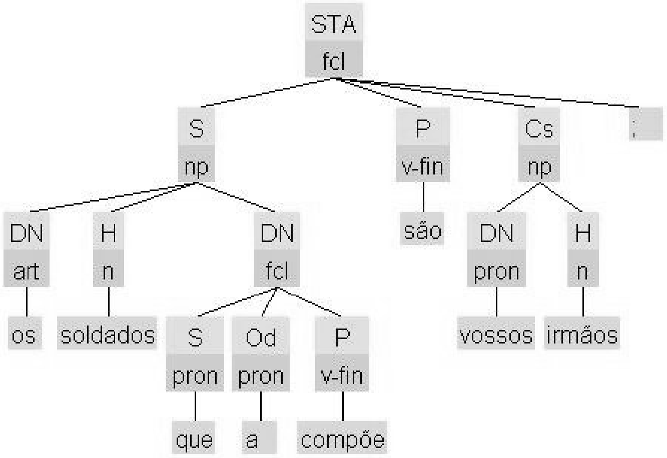

Fig. 2. Analysis tree

2.2 Adapting the parser

Tokenization and normalization: The first hurdle in the *Cartas* corpus was tokenization, a preprocessor task, where words had to be (re)created from line-separated parts ('|'-marked: *Victo- | ria, dis- | seme*), often in the presence of faulty hyphenation (*acompanhal | os, em | prega-las*). At the same time, apostrophes had to be automatically recognized as either accent markers *(a' = à),* hyphens *(mostra'se)* or ellision markers *(minh'alma).*

The second, and lexically very relevant, problem was that of recognizing word forms to be matched in the parsing lexicon, since spelling conventions were different, and – worse – unstable in our historical data. The following is a list of orthographical topics that were treated by a pattern-matching preprocessor.

- *geminated and triple consonants*: dd > d, ff > f, cc[aou] > c, sss > ss etc. *(occupão, soffra, deffeza)* cave: *vosa* > *vossa*
- *fusion*: *heide, hade > hei de, há de*

- *"greek" spelling*: ph > f, th > t, y > i *(enthusiasmo, authoridades, systema)*
- *nasals*: emt > ent (*semtirá*), om[df] > on (*comfiada*), aon > ão (*orgaons*) chaotic -ão: *áo, ào, âo, aõ, aò, àm,"ao, [r]ao*, *ôes, óes, oens* ans?$ > ã (*amanhan*, *irmans*), un$ > um (*algun*)
- *extra hiatus*-h: *sahiu, coherência, surprehender, adhesivo, anhelar*
- z/s-*dimorphism*: isa > iza, [aeu]z > s, [óú]s$ > z (*civilisadas, acuza, uzo, brazileiros, cazo*), cave: *crus* (> *z*), *gosa* (> *z*), *cafuzo, avós, após*
- s/c-*elision*: sci > ci, cqu > qu: s*ciência, acquiescência*
- *lack of tonic accents*: *di**a**rio, not**i**cia, admir**a**vel, excell**e**ncia,* cave: *faria*
- *superfluous accents*: *s**á**e, d**ó**e, cor**ô**a, contin**ú**a,* clitics: *imp**ol**-os*
- *fluctuating accents*: *n**ò**s, ter**à**, pa**ï**s, **à**lias, **á**s*

The dynamic lexicon: Inflexion and ambiguity: The morphological analyzer generates (ambiguous) lists of readings for input word forms. In order to match its paradigmatical endings rules, it needs to be told that historical Portuguese graphically expressed phonetic variation, as in the 3rd person singular *passado simplês* in *-io* and *-eo* (*consentio, envadio, attrahio, commetteo*), the nominal plural *-aes* (*officiaes, quaes*) or vowel variants in tonal verb roots (*podesse, surge, trouce* instead of *pudesse, surge and trouxe*). Also, additional accents would mark a vowel as closed or open, respectively (*recebêram, levára, chóra*). In some cases, variants would, of course, create new ambiguity (*mais* = *más* / *mas*), affecting the disambiguation grammar.

A final, less heuristic, but labour-intensive, solution was adding new variants of base forms or even inflected forms to the parsing lexicon. This path was chosen in the case of multiple or non-systematic (but frequent) variation (*agoa, anarchia, athé*) or, of course, in the case of lexemes not present in the existing modern Portuguese parsing lexicon (*bemfeitor, buçal, comtanto que*). A general lexicon-adaptation was to activate, for the handling of lacking accents, so-called R-forms (root-only forms) already present in the lexicon for derivation analysis. Note that in all cases the original orthography was maintained alongside the normalized annotation form.

The parsing grammar: Since the strategy chosen was not to write a new grammar from scratch, but to make rules targeting modern Portuguese work with historical data, problematic input had to be marked as such manually (e.g. Latin and French quotes) or automatically, so rules would be able to either ignore it or adapt to it. Due to the high incidence of upper case usage for ordinary nouns, many name chain fusion rules had to be inactivated in order to avoid over-generating of multi-word names. A special focus was on structurally important features (e.g. "superfluous" comma before *e*, "new" conjunctions like *comquanto*).

For some grammatical features, rules had to be changed or added, like in the case of direct objects marked with the preposition *a*, which – for human NPs – proved to be much more frequent in historical text. Another example were *que*-clauses attached to nouns without a mediating preposition (*o motivo porque, notícia que, o dizer que*). Interestingly, some grammatical peculiarities could be classified as "cross-Romance", like French-style superlatives (a), inflected participles after *ter* (b) and infinitive-markers (c), or Spanish-style clitic-fusion (d) and the above-mentioned "prepositional accusatives" (e):

(a) documentos os mais honrosos
(b) os que alguns votantes já tinham feitos
(c) era bem de dezejar ...
(d) a bondade de ouvirnos
(e) trate a todos com urbanidade

In rare cases it was necessary not only to accommodate for new syntactic structure, but to introduce *new* syntactic *functions,* not present in modern Portuguese, as for formal / provisional subjects and objects, a "germanic" feature only shared in the Romance family by French:

(f) ainda quando suas qualidades maraes o não fizessem merecedor desse e maiores cargos; *elle* está certo, que he indigno delles quem os procura
(g) Ninguém ouvio, nem ousará affirmal-*o* que Luiz Fernandes requeresse protesto algum

3. Evaluation

On a small test chunk from the *Cartas* corpus, the adapted CG-parser achieved F-scores (equally weighted recall / precision) of 97.7% for part of speech, and 91.4% for syntactic function, respectively. Constituent structure is partially implied by CG-tags, but was not evaluated in isolation. Though these numbers suggest error-rates are twice as high as for modern Portuguese, they are still good enough to allow semi-automatic corpus annotation, with a reasonable pay-off against manual revision.

	F-score (modern Port.)	F-score (*Cartas*)
Word class (PoS)	> 99 %	97.7 %
Syntactic Function	> 96 %	91.4 %

Table 1. F-scores of the PALAVRAS parser

In order to quantify parsing error tendencies on the one hand, and grammatical characteristics of the historical corpus on the other, the distribution of a number of categories was evaluated contrastively in revised and automatically annotated text chunks.

	of all @		N/PROP	PERS	*se*	*a*-PRP
@SUBJ>	4.6 (5.1)	of these:	48.4 (49.2)	16.1 (16.9)	6.3 (5.6)	
@<SUBJ	**1.8** (1.1)	of these:	79.5 (70.3)	**14.3** (20.6)	**6.3*** (9.6)	
@ACC>	1.9 (2.0)	of these:	**1.7*** (0.2)	59.5 (54.9)		0.1* (0.0)
@<ACC	4.9 (5.2)	of these:	77.5 (76.3)	12.8 (14.5)		**3.9** (1.4)

Table 2. Subject / object percentages in the revised corpus
(raw in parenthesis)[1]

The *Cartas* data show a strong preference for right-position of subjects (28% of subjects) and an increased prevalence of direct objects with “a” (4%). A comparison with the unrevised annotation (in parenthesis) indicates that these areas also are particularly error-prone – most likely because the parsing grammar (still) has a strong bias towards a more “modern” SVO subject/object distribution, a conjection corroborated by the diachronic comparison in the following table:

	of all @	of these			
		N/PROP	PERS	*se*	*a*-PRP
@SUBJ>	4.6 (5.5 – 6.8)	**48.4** (69.0 – 74.1)	**16.1** (6.0 – 8.3)	6.3 (3.4 – 1.4)	
@<SUBJ	1.8 (0.8 – 0.7)	79.5 (80.2 – 86.2)	14.3 (12.9 – 7.2)	6.3 (**11.8** – 3.3)	
@ACC>	1.9 (0.9 – 0.8)	1.7 (0.3 – 0.6)	59.5 (**50.5** – 57.3)		0.1 (0.0 – 0.0)
@<ACC	4.9 (4.5 – 4.9)	77.5 (83.9 – 90.8)	12.8 (9.7 – **2.7**)		3.9 (0.4 – 0.5)

Table 3. *Cartas* historical data vs. modern data (PE – PB)[2]

The above comparison with the two main variants of modern Portuguese shows a clear decrease of *a*-accusatives in modern Portuguese and higher incidence of pronoun-usage in the *Cartas* corpus, which, however, might also be due to a difference in genre (the modern data were news text). More interestingly, the distribution of object clitics shows that historical Brazilian Portuguese has the same percentage of post-positioned clitics as modern European Portuguese (ca. 1/3 of all clitics), while modern Brazilian Portuguese avoids this construction (12%).

A special trait of the CG annotation used here is the form and function tagging not only of words, but also of subclauses. This permitted us to evaluate the distribution of complex categories, like non-finite adverbial clauses:

[1] * = very few instances, < 10.

[2] PE = European Portuguese (*CetemPúblico* newspaper corpus), PB = Brazilian Portuguese (*Folha de São Paulo* newspaper corpus).

	of all @		INF	GER	PCP
@ICL-ADVL>	0.2 (0.2)	of these:	0.0* (0.0)	84.6 (95.6)	15.4* (4.4)
@ICL-<ADVL	0.9 (0.6)	of these:	7.4* (6.6)	92.6 (92.0)	0.0* (0.1)

Tabel 4. Non-finite adverbial clauses (corrected vs. raw)

	of all @		INF	GER	PCP
@ICL-ADVL>	**0.2** (0.05 – 0.04)	of these:	0.0* (0.0 – 0.0)	84.6 (87.2 – 91.7)	15.4 (12.8 – 8.3)
@ICL-<ADVL	0.9 (0.4 – 0.2)	of these:	7.4* (7.2 – **3.2**)	92.6 (90.4 – 95.4)	**0.0*** (2.5 – 1.3)

Table 5. Non-finite adverbial clauses (corrected historical vs. PE – PB)

Non-finite adverbial subclauses were 2–5 times as common in the *Cartas* corpus than in either modern Portuguese variant. Even in relative terms, the use of fronted participle clauses and adverbial infinitive clauses was highest in *Cartas.* Again, modern Brazilian data moved further away from the historical antecedent than modern *European* Portuguese.

4. Conclusions and outlook

We have described the research context and scope of a 19th century historical corpus of Brazilian Portuguese, as well as the transcription and annotation principles used for its grammatical annotation. A combination of token filtering techniques and lexicon additions, as well as rules for orthographical variation, historical inflexion paradigms and context sensitive disambiguation rules allowed us to address this task using a modified, pre-existing Constraint Grammar for modern Portuguese with reasonable results (2.3% PoS errors and less than 8% syntactic function errors).

Future work should focus on manual revision of larger sections of the corpus, as well as a broader text type base, in particular by balancing the public letters to the editor by private letters from corresponding periods. Also, a revised CG annotation could be fed into a function-based phrase structure grammar in order to construct a syntactic treebank covering historical language data, a method previously employed for modern newspaper texts in the *Floresta Sintá(c)tica* project.

References

Afonso, Susana / Bick, Eckhard / Haber, Renato / Santos, Diana 2002: 'Floresta sintá(c)tica': a treebank for Portuguese; in: González Rodríguez, Manuel / Suárez Araujo, Carmen Paz (eds.): *Proceedings of LREC 2002, 3rd International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29–31 May 2002). Paris: ELRA, 1698–1703.

Barbosa, Afrânio / Lopes, Célia (eds.) 2002: *Críticas, queixumes e bajulações na Imprensa Brasileira do séc. XIX: cartas de leitores e cartas de redatores*. Rio de Janeiro: UFRJ.

Bick, Eckhard 2000: *The Parsing System 'Palavras' – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus: Aarhus University Press.

Britto, Helena / Finger, Marcelo / Galves, Charlotte 2002: Computational and linguistic aspects of the construction of the *Tycho Brahe Parsed Corpus of Historical Portuguese*; in: Pusch, Claus D. / Raible, Wolfgang (eds.): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache · Romance corpus linguistics: Corpora and spoken language*. Tübingen: Narr, 137–146.

Guedes, Marymarcia / Berlinck, Rosane de Andrade (eds.) 2000: *E os preços eram commodos: anúncios de jornais brasileiros do século XIX*. São Paulo: Humanitas.

Megale, Heitor *et al.* 2001: Normas para a transcrição de documentos manuscritos para a história do português brasileiro; in: Mattos e Silva, Rosa Virgínia (eds.): *Para a história do português brasileiro.* Vol. II, t. II. São Paulo: Humanitas / Fapesp, 553–555.