Gramateca: corpus-based grammar of Portuguese

Diana Santos

Linguateca/University of Oslo
d.s.m.santos@ilos.uio.no

Abstract. This paper has two aims: to present Gramateca, an initiative for corpus-based grammar of Portuguese recently launched by Linguateca, discussing its rationale and aims, and to present a small pilot of an advanced corpus description for Portuguese, which we believe to be a pre-requisite to do sound grammar work, "know thy corpus".

Keywords: Portuguese, grammar, replication, corpus characterization

1 Introduction

There are fortunately many corpus-based studies of Portuguese, from the initial studies of *Português Fundamental* in Portugal and the NURC project in Brazil, see [1, 2]. We can also cite [3, 4], two extensive high-quality corpus-based grammars that have recently been published. However, there is no way to directly check the data used by those studies, or, better, its interpretation.

Also, we believe that time is ripe for providing the possibility to do statistically informed larger scale grammar studies, or grammar (as a massive noun), which allow furthermore consulting the material/data used. This is why we announced Gramateca in January 2014, `http://www.linguateca.pt/Gramateca/`.

We believe that both NLP and linguistics have a lot to contribute to Portuguese grammar – see the recent discussion in the corpora-list on this matter – and we hope to have contributions and discussion from all interested. There are (and have been) several computational grammars of Portuguese, such as the one embedded in PALAVRAS [5] or the one used in the Tycho Brahe project [6]. Although their deployment and evaluation is naturally based on corpora, they are built to be used by computers, and their documentation is not necessarily corpus-justified.

The way we see it, there are two possible models for proceeding with a corpus-based grammar, both inspired by Biber's work: (i) Finding out, through multidimensional analysis, what are the relevant dimensions that allow one to characterize genres or text types in Portuguese, as in [7]; and (ii) Using a balanced annotated corpus of (four?) kinds of text and providing a quantitative characterization of several areas of grammar, as in [8]. However, in both cases the authors created (and annotated) a specific corpus to serve as basis for the grammar – and in the second case the corpus was not even made available for inspection. This is something we intend to improve upon, both by using available corpora, and by not restricting to one unique corpus the grammar work. We believe, the more corpora, the better, and that corpora of old language and of novel genres[9] are essential as well.

So, we supply access to all these corpora and their annotation, not requiring that all be used. Rather, one should be able to pick exactly the genres and subsets that concern her or him. Since the particular subset will be made clear in the results and conclusions, and in the resulting annotated data (more on this below), nothing should hinder different grammarians to use different subsets. What we intend with Gramateca, by giving the community access to the annotated material, i.e., to the interpretation of the raw material, is to ensure replication, and more fine-grained experimentation from our peers.

In this initial paper, we describe the corpora and their (automatic) annotation. As a full fledged parser, PALAVRAS [10] must be acknowledged, since it is the main responsible for AC/DC annotation and even the first take of quantitative description could not be done without it. But we also intend to pursue the comparison of different automatic frameworks to analyse the same data (as already present in a few corpora).

2 Quantitative description of the AC/DC corpora, first take

In order to be able to know what kind of (annotated) material we make available to the community – 1,300 million tokens, roughly 882 millions from Brazil and 244 millions from Portugal – it is important to provide a bird's eye view of all corpora, and also some quantitative features about each (all numbers are dated 14 May 2014).

It turns out that this is easier said than done, because the corpora in AC/DC have been compiled by different people in different continents and with widely differing purposes, see [11, 12]. So, a common denominator often lacks in the names associated with the different parts – when the corpora have this external information at all! Also, most descriptions conflate genre, register, dissemination mode, and the oral/written dichotomy. The integrity of each corpus is not changed, anyway. The bird eye's view corresponds to the corpus `todosjuntos` (all toghether). What is shown in Figure 1 is therefore just a very rough approximation of the text sizes involved.

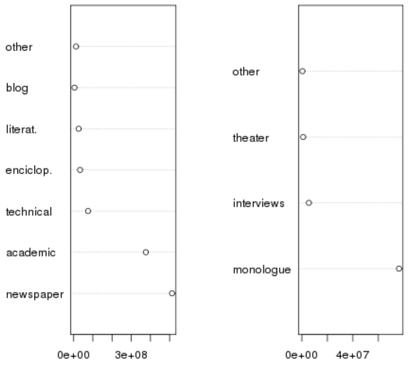

Fig. 1. Distribution of genres in the AC/DC material: written and oral

The most basic quantitative information is available from the AC/DC pages, namely the number of all tokens (“unidades”), running word tokens (“palavras”), word types (“tipos”), sentences (“frases”), and the number of PALAVRAS-attributed parts of speech: nouns (S), verbs (V), adjectives (Adj), and proper nouns (Prop),[1] ilustrated for some corpora in Table 1. Using all sorts of information associated with the corpora, one could also compute e.g. number of clauses (main verbs), tense forms, passives, colours, emotions, body parts, as illustrated in Table 2. For these cases the uncertainty is considerably higher, so a measure of confidence or a confidence interval for each should be provided. As to comparing corpora with radically different sizes, blindly normalizing the different figures is dangerous, for two distinct reasons: First, one should not lightly cross-over different texts in one corpus; second, ideally one should use the smallest corpus (or corpus subdivision) as the unit, and divide the larger corpora into several of the same size, and then take the average of the resulting values.

Table 1. Basic numbers for AC/DC corpora: Museu da Pessoa, NILC/São Carlos, CHAVE and ENPCpub

Corpus	Unidades	Palavras	Tipos	Frases	S	V	Adj	Prop
MP	1,836,371	1,421,142	42,795	93,525	237,010	263,957	50,815	35,146
NSC	42,914,402	32,461,799	399,763	1,988,621	7,113,650	4,298,530	1,842,594	323,412
CHV	124,093,624	97,884,763	696,775	4,682,363	20,699,744	12,745,355	5,911,581	5,430,724
ENPC	93,160	72,374	12,874	4,371	13,273	12,774	3,853	2,542

Table 2. Other numbers for AC/DC corpora: types correspond to lemma types

Corpus	Clauses	Perfeitos	Passives	Colours	(types)	Emotions	(types)	Body	(types)
MP	229,055	46,956	6,390	651	64	12,075	437	12,075	437
NSC	3,528,855	610,856	257,849	26,845	410	231,704	1,681	88,745	336
CHV	10,429,598	1,715,006	699,868	79,659	587	711,525	2,353	249,564	372
ENPC	10,809	2,076	444	252	63	1,197	321	1158	314

3 Quantitative description of the AC/DC corpora, advanced

Since corpora are not homogeneous, and often have well defined parts, corresponding to different authors, or texts, or varieties, or domains, one should be careful before amalgamating data, lest one got uninteresting and uninterpretable numbers.

So, all the statistics one might be interested in should be obtained in a form that does not illegally (or better, against linguistic sense) conflate reasons for diversity. We pursue here two examples to give concrete data relevant for making this point. For other pitfalls of quantitative characterization in Portuguese corpora, see [13].

[1] We also provide numbers for adverbs, personal pronouns, prepositions, conjunctions, determiners, specifiers and numerals, ommitted from the table for lack of space.

3.1 Sentence length

One simple issue to illustrate the dangers of amalgamation is measuring sentence length. Trying to get the average sentence length from all Portuguese (written) sentences will not give us very much. Rather, it is possible and probable that different genres, different texts, different authors, etc. will have different sentence lengths. So in Table 3 we present a set of sentence lenghts derivable from our corpora (in words, not tokens, that is, disregarding punctuation):

Table 3. Sentence length of several corpora (sections): MP stands for Museu da Pessoa, and p and b for the variety

MP	MPp	MPb	CONDIVb	CONDIVp	VercialJD	VercialEQ	ObrasMA
14.0	13.1	14.3	14.2	18.4	12.2	17.2	20.2

The distribution of sentence length in each corpus or author could also be more relevant (and characteristic) than its average, especially since averages are well-known to be prone to deformation due to outliers. In the previous table we have therefore measured average sentence length of Júlio Dinis (JD), Eça de Queirós (EQ) and Machado de Assis (MA). Back to the corpora as a whole, in figure 2 one can see the sentence lengths for different corpora in a boxplot.

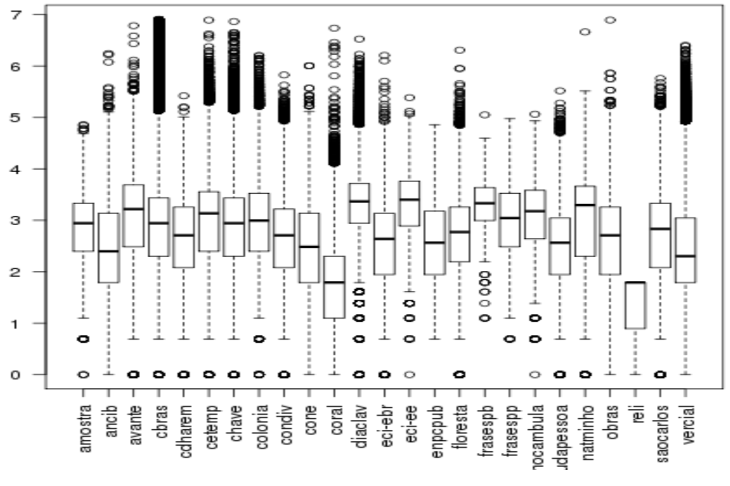

Fig. 2. Distribution of (the logarithm of) sentence lengths in several corpora

3.2 Lexical density

Another feature that is often used as characterizing for example written vs. oral speech is lexical density, in the sense of the distribution of content vs. grammatical words in X words, or per clause or utterance. [8] find that for English the lexical density of oral speech (using also what they call "inserts") is significantly lower than for written genres.

One could even be more specific and measure the density of personal pronouns (number of personal pronouns per 100 words), or of emotion adjectives (per 1,000 words), in different genres/corpora/etc. To give a flavour of the numbers obtained in those cases, and of some surprises, see Table 4. Further data and discussion on these matters can be found in [14].

Table 4. Other indicators for several corpora (sections)

Feature	MP	MPp	MPb	CONDIVb	CONDIVp	VercialJD	VercialEQ	ObrasMA
Pers. pron.	5.9	6.1	5.0	1.7	2.2	6.7	4.2	3.6
Emotion adj.	0.9	1.1	0.8	1.4	1.8	4.7	5.3	4.0

While this is obviously a very tiny subset of all possible quantitative features, it allows us to stress both (i) the large number of statistics we can get from the corpora, and (ii) the care one has to exercise to get meaningful numbers for one's own analysis. So, while a higher density of emotion adjectives in dedicated newspaper than in interviews should be surprising at first, digging into the fashion and football subparts may explain it. Conversely, the fact that there are more personal pronouns in oral interviews from Portugal than from Brazil may be explained by more clitics in Portugal vs null objects in Brazil (a purely linguistic explanation that could be conteracted by more null subjects in Portugal), but also by a corpus characteristic, that of significant more interaction in the interviews. See a similar puzzle on different colour words per variety solved in [15].

4 Concluding

Perhaps the most important reason for the present article is to increase the Gramateca community so that it reaches a wider audience, and that more people interested in Portuguese grammar can benefit from our efforts.

There is nothing required to join the mailing list as an observer, or to use whatever results we have been able to get. If you want your work to be considered as part of Gramateca, the only thing we request is that you publish it/make it available through our site, together with the data and a detailed enough description so that it can be replicated (and hopefully improved).

At the date of writing, (independent) work on conditional connectives [16], body language [17] and emotions [18] has been launched, and the annotated data have been made available for inspection, while tools for revising annotation and for facilitating comparisons among subcorpora [19] are under development.

References

1. Santos, D.: Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. In Johannessen, J., ed.: Language Variation Infrastructure: Papers on selected projects. Volume 3. (2011) 113–128
2. Varejão, F.d.O.A.: O português do Brasil: Revisitando a História. In: Cadernos de Letras da UFF Dossiê: Difusão da língua portuguesa. Volume 39. (2009) 119–137
3. Ilari, R., ed.: Palavras de classe aberta. Vol 3. da Gramática do Português Culto Falado no Brasil. Editora Contexto (2014)
4. Raposo, E.B.P., do Nascimento, M.F.B., da Mota, M.A.C., Segura, L., Mendes, A., Vicente, G., Veloso, R., eds.: Gramática do Português I e II. Gulbenkian (2013)
5. Bick, E.: Portuguese syntax: Teaching manual (2000)
6. Galves, C.: Tycho Brahe Parsed Corpus of Historical Portuguese: SYNTACTIC ANNOTATION SYSTEM (2007-2008) http://www.tycho.iel.unicamp.br/ tycho/corpus/manual/synfrm.html.
7. Biber, D.: Variation across speech and writing. Cambridge University Press (1988)
8. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: The Longman grammar of spoken and written English. Longman (1999)
9. Crystal, D.: Internet Linguistics: A Student Guide. Routledge (2011)
10. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University, Aarhus, Denmark (November 2000)
11. Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G., eds.: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000). (31 May-2 June 2000) 205–210
12. Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço AC/DC. In: The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009). (8-11 de Setembro 2009)
13. Santos, D.: Podemos contar com as contas? In Aluísio, S., Tagnin, S., eds.: New language technologies and linguistic research: a two-way road, Cambridge Scholars Publishing (2014)
14. Santos, D.: Comparing oral (transcribed) and written corpora in portuguese (2014) Presentation at GSCP2014, Stockholm, http://www.linguateca.pt/Diana/download/CompOralEstoc.pdf.
15. Santos, D., Silva, R., Freitas, C.: Pluralidades na cor: contrastando a língua do Brasil e de Portugal. In da Silva, A.S., Torres, A., Gonçalves, M., eds.: Línguas Pluricêntricas: Variação Linguística e Dimensões Sociocognitivas, Braga, Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa (2011) 555–572
16. Marques, R.: Modalidade e condicionais em português. submitted (2014)
17. Freitas, C., Santos, D., Sousa, R., Jansen, H., Mota, C.: Esqueleto: Body language in Portuguese. submitted (2014)
18. Santos, D., Mota, C.: Opinions and sentiment analysis in Gramateca. submitted (2014)
19. Simões, A.: Comparador: forma de auscultar corpos no AC/DC (2014) http://www.linguateca.pt/documentos/Comparador.pdf.