# First steps of Gramateca: a corpus-based grammar initiative for Portuguese, driven by Linguateca

Diana Santos

ILOS

d.s.m.santos@ilos.uio.no

20 February 2014

## Gramateca

- Using the AC/DC corpora to do grammar(s) for Portuguese
- A long wished for development
- A framework, a community, and results
- An initiative under Linguateca's philosophy

## Why?

- Linguists keep analysing data: but where do these analyses end?
- "Data which have been analysed, and/or used as examples, are more interesting than just data": is this true?
- Quantify and sort: this has not been done in a broad-coverage basis for Portuguese, although there is some corpus-based grammar work around



## Sampson's quote

Sampson (2003) claimed that he considered his book *English in the computer* (Sampson, 1995), describing the problems and solutions in the annotation of English of the SUSANNE corpus, his greatest contribution to linguistics, no matter the fact that people kept using and reusing and citing the corpus while almost completely disregarding the underlying linguistic work:

> *From our point of view, the explicit annotation scheme is the central output of our research effort, and the corpora that we develop in the process of debugging the annotation scheme should be seen as secondary by-products (though in practice it seems that this scale of priorities is not one which others can easily be persuaded to share).*

## This is a moot point... is it?

There are in fact two kinds of people

- People want white material to "enterrar o dente", not material already "swallowed"
- People want things already done, "off the shelf", to compare with

It all depends

- most categories are not straightforward, and people do not like to read manuals/documentation/other people's papers
- people do not want to document their annotation choices



## Original plan

Absolutely irrelevant? At least, initial dissemination: discussion list with open archives

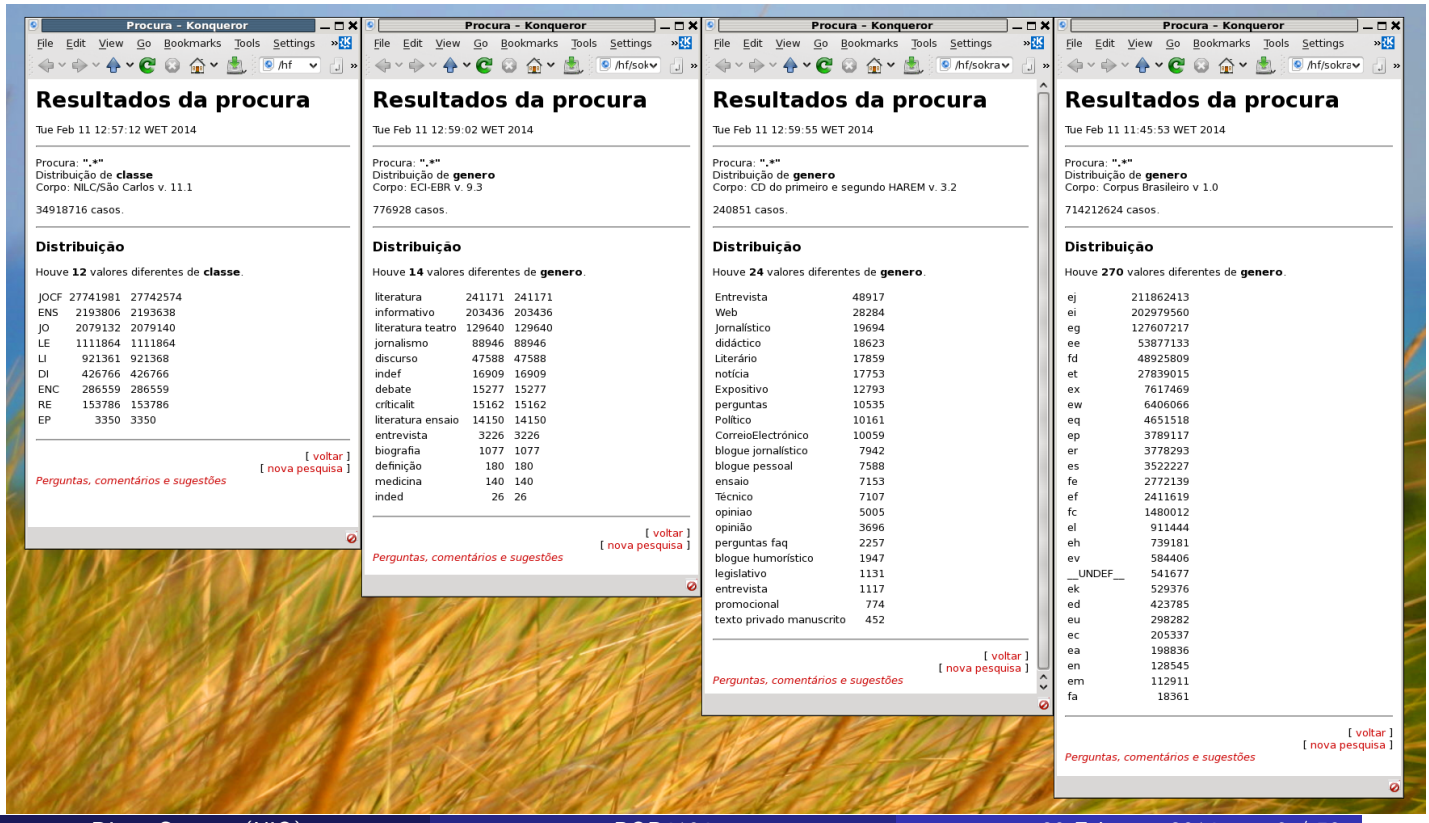
## Current plan

- Quantitatively documenting each corpus (qualitative descriptions aready exist)
- Organizing the site (e.g. how to keep many versions of annotation in a form which can be used by linguists?)
- Doing some grammar work
- Publication (Creative commons plus papers and presentations)
- Lines of action
- What kinds of statistical tools to implement?
- Organization
  - Top down list of subjects, bottom-up lines of action
  - Evaluation of the material



## "Genre/register/mode" assigned: written

# Different corpora, different genres






# Genre: oral

This is the kind of "genre/register/mode" that people (or we) have assigned to their corpora: oral material.

## Different corpora, different genres

Gramateca or AC/DC solution: leave the corpora with their respective genres, but conflate, in the "Together corpus" into a common grid. This way one can use the original material, and one can also try to make sense of the complete material.

This is still not easy, as you can see by the following decisions:

- Magazines and newspapers: two different genres?
- Horoscopes, TV program: under newspaper?
- Textbooks: under technical?
- Clinical literature from medicine products: under technical?
- Literary criticism: under essay?
- Plays: under oral?
- Parliamentary proceedings: under written, legal?
- Computer manuals: under technical?
- Biographies: under enciclopedic?
- Poetry: under literature?
- Reviews written by students: a different genre?
- Letters and email: two different genres?



## A little more on the oral corpora

First of all, they are not speech corpora, they have all been interpreted and transcribed by (different kinds of) linguists or other people.

- Corpus brasileiro: freely available material taken from the web
- Museu da Pessoa: ordinary people (students?) have heard the records and written down what they heard. The purpose of this museum is to keep alive the memories of common people... not linguistic research. Interestingly, there are different styles and genres of the interviews, which have been done by (again) non linguist reserachers. Different conditions in Brazil and Portugal.
- Diaspora TL-PT: interviews conducted by (ordinary people) members of the East Timorese community to other members, with the implicit goal of (also) learning their atitudes towards language etc. The interviews were then transcribed by syntacticians/semanticists.
- C-ORAL-BRASIL: spontaneous conversation in Minas Gerais, with the specific purpose of studying the dialect of non-educated people. Transcribed by phonologists, and conversation analysts.

## Examples of the oral material

Corpus Brasileiro, TV Debate:

> *De que adianta ter dinheiro, ter bens materiais?*
> **Lula** *Eu sei o que é enchente, porque morei na Vila Carioca, em São Paulo, no Bairro do Ipiranga, porque morei na Vila São José, em São Caetano, porque morei na Ponte Preta, em São Paulo, e todas as casas que eu morei, até um metro e meio de água entrava dentro de casa. Por isso eu sei o que é enchente .*



## Examples of the oral material

Museu da Pessoa, from Portugal:

> *Havia alguns que fugiam e quando voltavam ainda levavam mais. Eu casei no dia 15 de Janeiro e já trabalhava na Câmara Municipal de Gaia, eu comecei a trabalhar na Câmara em 1951, e fui dar-lhe o dinheiro referente aos quinze dias de vencimento do mês de Janeiro e ele, mesmo sabendo que eu já estava casado e que precisava do dinheiro, ficou-me com ele, enquanto existiam outros filhos que ganhavam e não davam dinheiro nenhum aos pais. Só depois do 25 de Abril é que eu me apercebi do ódio encapotado que havia em Portugal.*

# Examples of the oral material

Museu da Pessoa, from Brazil:

> *– Eu entrei no Aché em 03 de agosto de 1992, há dez anos atrás. Até então, eu morei uma época em Vitória, no Espírito Santo, onde até pleiteei uma vaga no Aché, mas na época eu era solteiro e tinha uma certa exigência, você tinha que ser casado, eu não consegui. Voltando ao Nordeste, já casado, de situações assim, surgiu uma vaga, surgiu no setor, no interior.*



# Examples of the oral material

Diaspora TL-PT:

> *A: – E conseguiu ter esse passaporte e viajou cá em Portugal ou passou...*
> *B: Não, com esse passaporte nós não poderíamos ter o visto para entrar aqui em Portugal, porque não havia embaixada portuguesa na Indonésia .*

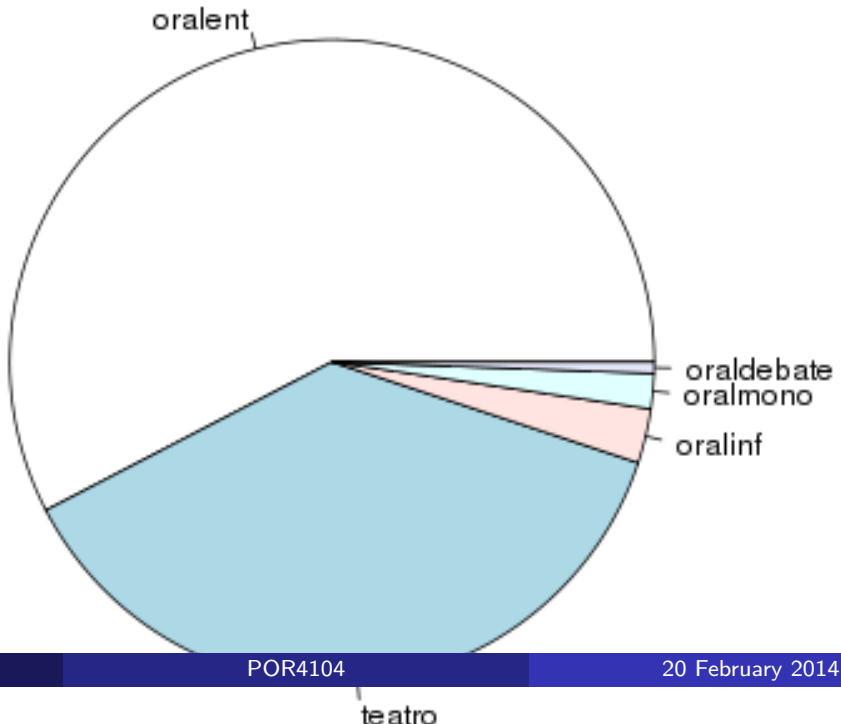
# Examples of the oral material

C-ORAL-BRASIL, simplified for AC/DC:

*Eu tava lá em & ca + meia-noite e meia, rolando de rir, acordando a vizinhança +*
**DUD** *Primeiro você põe o dedinho, aí cê põe o dedinho e vai forçando, não sei o quê, até ficar nu sei o quê, e aí, vai, depois de, mete bronca, nu sei o quê. Aí no segundo dia cê vai, põe dois dedinhos, nu sei o quê, só um pedacinho, nu sei o quê, depois até a metade, e aí depois, mete bronca.*

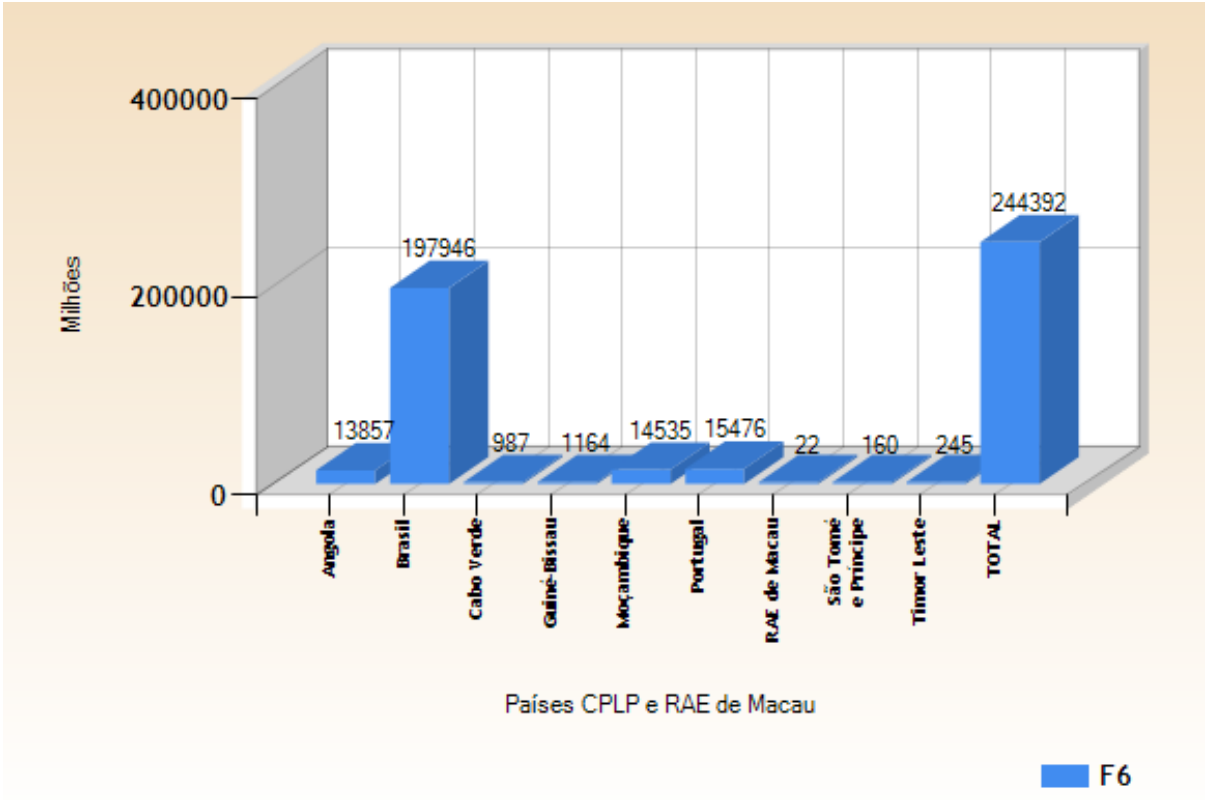


# Distribution of oral material



# Varieties of Portuguese

There are two "national" varieties and several emerging varieties.






(from `http://observatorio-lp.sapo.pt/pt/dados-estatisticos/`

# Portuguese as an international language

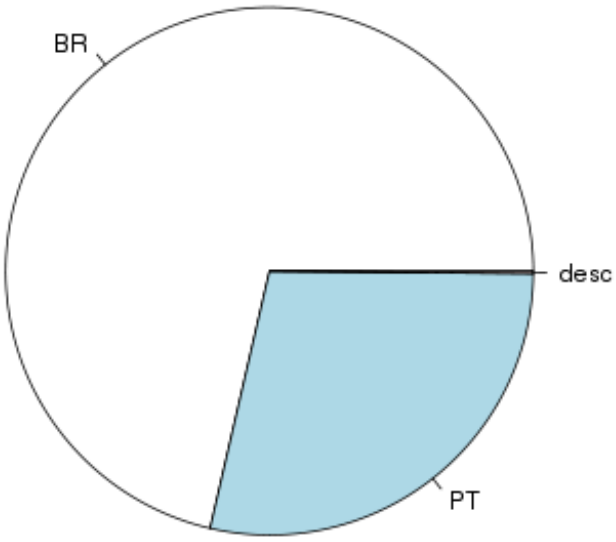
- Portuguese has been identified bu the UK as one of the ten languages for the future, `http://observatorio-lp.sapo.pt/Content/Files/languages-for-the-future-report.pdf`
- Portuguese is the most widespread language in the southern hemisphere
- Portuguese is crucially connected with oil, in Brazil, Angola, and Timor

Historically

- Portuguese discoverers and missionaries have been crucial for Western-East relations (Japan, China, India)
- The Portuguese empire has had to solve/deal with language and cultural diversity in a global way: there is much good and bad to be learned from us

# Varieties of Portuguese in the corpora

Our corpora have, in addition, quite different ortographic conventions. In fact, the last *Acordo Ortográfico* made things even worse! (Veiga et al. 2011)






# Dating

Different periods are covered by different materials.

- From 1500 to 1920: Vercial, Portuguese literary texts in updated ortography
- From 1820 to 1950: OBras, Brazilian literary texts in updated ortography
- From 1852 to 1998: COMPARA, originals in Portuguese
- From 1972 to 2002: COMPARA, translations in Portuguese
- Three decades: 50s, 70s and 90s: ConDiv
- CETEMPúblico, CHAVE and NILC/São Carlos: 90s
- New corpora of new genres: 2000s

## Annotation

There is a lot you can get in the AC/DC corpora

- emotions
- syntax
- colours
- body words
- clothing

and, specifically, for specific corpora,

- author
- title
- variety
- subject/topic/theme
- date
- neologisms



## What is (a) grammar?

The scientific description of a language? Social conventions so that we are able to communicate?

> *Labov's statement that "[t]he central axiom of sociolinguistics is that community is prior to the individual" [...] nothing could be further from the truth. [...] sociolinguistically relevant if, at some point of time, it has passed through the filter of the human mind, where it was either readied for production or comprehended and interpreted. (Gries, 2013:7)*

See John M. Ellis: Every language is a particular system of classification

> *[...] linguistic categories are primarily the reflection of the collective purposes of the speakers of a language rather than direct reflections of the structure of the world. (Ellis, 1993:34)*

## What is (a) grammar? (contd.)

Through interaction we learn to think, and language is the most important bit of that interaction.
Could we get concepts (i.e., feelings, ethics, facts/individuals/events) without having heard / read them from others?
Grammar and lexicon are the limits that are required to communicate.
Grammar is learned in a specific order (the most frequent things first?) and this may be relevant to understanding basic categories and – less basic.
And then there is the last part, that has to do with (linguistic) power: Some forms are selected as high-brow, others as incorrect. For this there are style manuals and normative/prescriptive grammars, to separate the educated from the non-educated. Often this is also what makes varieties: you say this is this variety, but that in the other. "You shall not govern over my variety"...



## This is what mainly grammars have been written to

- teach children (or less sure adults)
- teach foreigners
- teach the computer (NLP)

In other words, there is some received knowledge (adults, native speakers, humans...), and there's X who/which needs it.
But this is not the whole story, because the goals of these three activities are different

- improve their own language / explain their best tool
- teach them a NEW language / correct their misconceptions
- be able to do things with (human) language (really a interested activity)

## Teaching and grammar?

Of course all this can be problematized:

- Language or language(s)? Teaching academic English can be useful for both natives and foreigners
- Native speaker's errors do not necessarily are good examples for foreigners
- Immersion and not grammar is how adults should be taught

but I won't do it.



## Corpus-based grammars

- Do not trust the corpus blindly! Ask it and build it with care, but keep always the upper hand. It is the grammarian who knows what is interesting (or what he wants to study). (There are no data without theory!)
- In our case (Gramateca), one can use different corpora for different things, and/or use the "altogether" corpus for rough quantitative overviews.
- "Corpus-driven" is an idiotic concept. There is always a lot in the texts that you are not interested in. Corpus-driven can be OK for machines (or if we are facing an unknown language), but not for people!

## Concrete Gramateca

If one goes bottom up, we have currently three areas of attack

1. Conditional connectives
2. Oral vs. written clues
3. The grammar of the human body

If one goes top down, the wisest is to use the traditional organization(s) and see how each of them can be illuminated by corpus work.



## Comparing with Johansson et al. (1999)

In 1999, after many years of joint work, the Longman grammar of English was published, based on a small corpus, automatically annotated and revised by the team. I don't know how the sections were chosen, but I suppose it was to be traditional in that sense. The authors mention the word revolutionary, anyway.
The statistical processing is rather straightforward:

- distribution per part-of-speech
- distribution by four genres/text type

Statistic significance is dismissed by saying that only the important/relevant issues are mentioned.

## What is a (traditional) grammar organization?

- Well, it depends on the grammarian. It would be a very interesting subject, comparative grammar writing, but we will not attempt it. Rather I will be very centered in my own teaching, and provide/produce a chapter/study for each thematic class: around 30-36.
- Note that for publication it is possible that people use different parts and/or give them a different bias, so the idea that a grammar is a straight jacket is not necessary after hypertext was invented. See e.g. Kristine Eide and Kåre Nilsson's grammar for UiO students in the 1990s.



## Publication

- A grammar for a language is work on progress, and by definition it will never be finished because language evolves as well as linguistic theories. So we are launching (better means for) a never-ending research area...
- In addition, publishers are one of the best (and unethical) businesses in town! I have no intention whatsoever to make them richer, given that (so far) the Internet is still free and the people who would benefit most from a Portuguese grammar is excluded to have access to the publishers' business.
- Finally, a grammar of a language is for practical purposes, not for the advancement of personal egos, so it should not be copyrighted, but copylefted. Anyone who wants to get it can improve it but all get to benefit from it (except most publishers).

## Quantitative characterization

What is so good about numbers? They allow us to generalize...

- Words, sentences, clauses
- Proper names, themes, emotions, parts of speech, lexical density, subbordination

and distribute

- authors, dates, varieties, genres

So the first infrastructural service we will implement for Gramateca is

- distribution on demand
- correlation on demand



## Examples of quantitative characterization

- Distribution of passive in relative clauses
- Distribution of relative clauses per genre
- Distribution of emotions per genre/theme
- Distribution of colours per variety
- Distribution of emotions per author

All this is already possible to do, but (relatively) cumbersome.

## Problems of quantitative characterization

- All these classes require many linguistic decisions
- The units over which one distributes and counts require thought:
  - what is the biggest part, the Brazilian or the Portuguese in the Museu da pessoa?
  - who uses more colours? (percentage of colour adjectives in all adjectives, or percentage of colour words in all words, or percentage of sentences with at least a colour word in all sentences)
- Without documentation, or possibility of replication, it is totally useless: "There are truthful people, there are liars, and there are statisticians"
- another problem is summing up everything (smoothing): again, you can come to the opposite conclusion if you are not careful (Simpson's paradox)



## Simpsons paradoks (from SPR4104)

Til sammen ("pooled") finnes det en assosiasjonsretning (marginal, i det det leses i margen). Men, om vi ser hvert tilfelle, finner vi kondisjonale assosiasjoner i motsatt retning.
Data[1]: Dødsdømte i Florida (1976-1987) og rasen til advokatene deres (svart/hvit). 2x2 krysstabell.
Når man ikke tar i betraktning rasen til den dødsdømte, er svarte advokater bedre. Hvis man tar i betraktning rasen til den dødsdømte ("hvis man kontrollerer for rasen, og lar den være konstant"), kommer man til den motsatte konklusjonen.
Hva gikk galt? Det finnes en sterk avhengighet mellom rasene til dødsdømte og advokatene deres. Og i tillegg fikk hvite kriminelle dødsdom mye oftere.

[1] Eksempel fra Agresti 1996. Selv om det heter Simpsons paradoks, ble den først oppdaget av Yule.

# Simpsons paradoks (from SPR4104)

Utforsk Simpsons paradoks ved å sammenlikne de forskellige tabellene.
Studieopptak (fra Devore & Berk 2007)

| | ja | nei |
|---|---|---|
| menn | 233 | 324 |
| kvinner | 88 | 194 |

| | ja | nei |
|---|---|---|
| menn | 41.83124 | 58.16876 |
| kvinner | 31.20567 | 68.79433 |



# Simpsons paradoks (from SPR4104)

Mekkingopptak

| | ja | nei |
|---|---|---|
| menn | 151 | 35 |
| kvinner | 16 | 2 |

| | ja | nei |
|---|---|---|
| menn | 81.18280 | 18.81720 |
| kvinner | 88.88889 | 11.11111 |

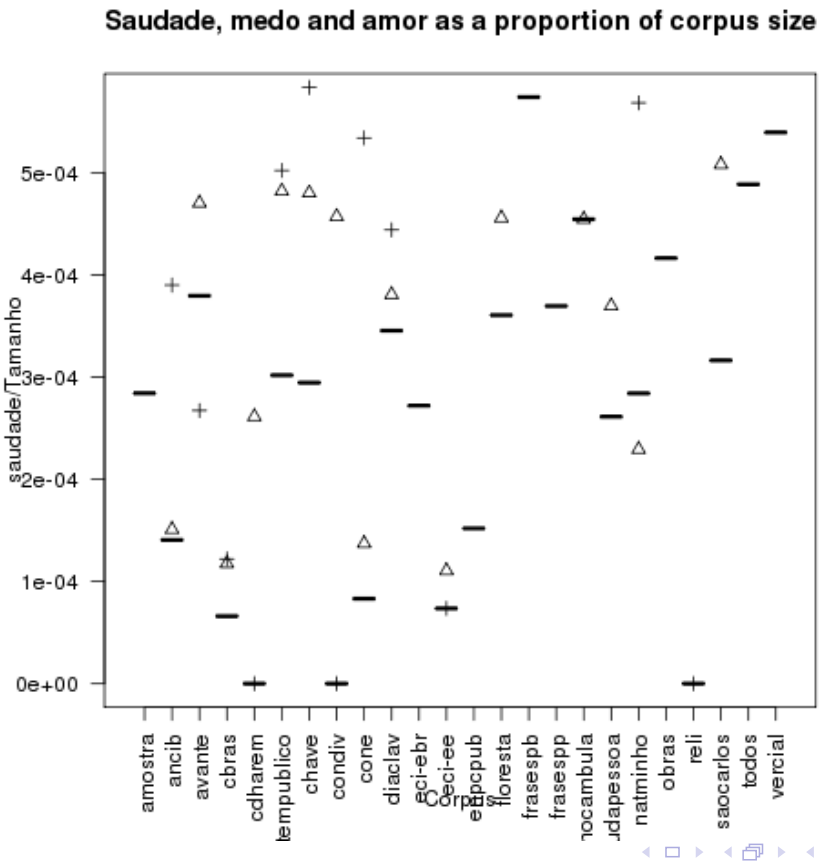
Litteraturopptak

| | ja | nei |
|---|---|---|
| menn | 82 | 289 |
| kvinner | 72 | 192 |

| | ja | nei |
|---|---|---|
| menn | 22.10243 | 77.89757 |
| kvinner | 27.27273 | 72.72727 |

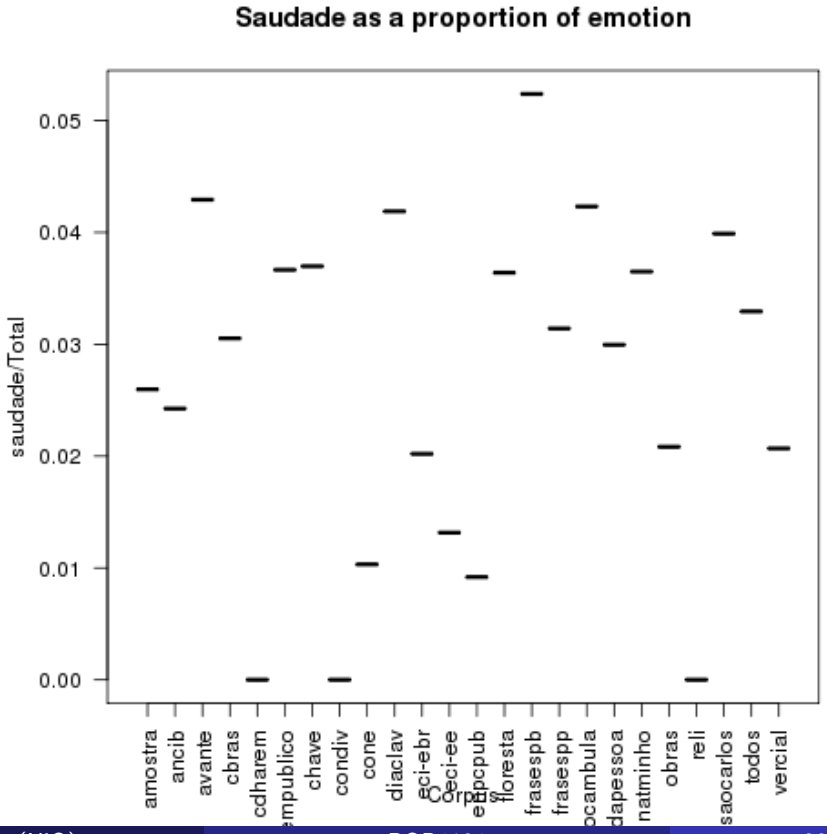
I begge har kvinner storre andel enn menn!

# Exploring emotions

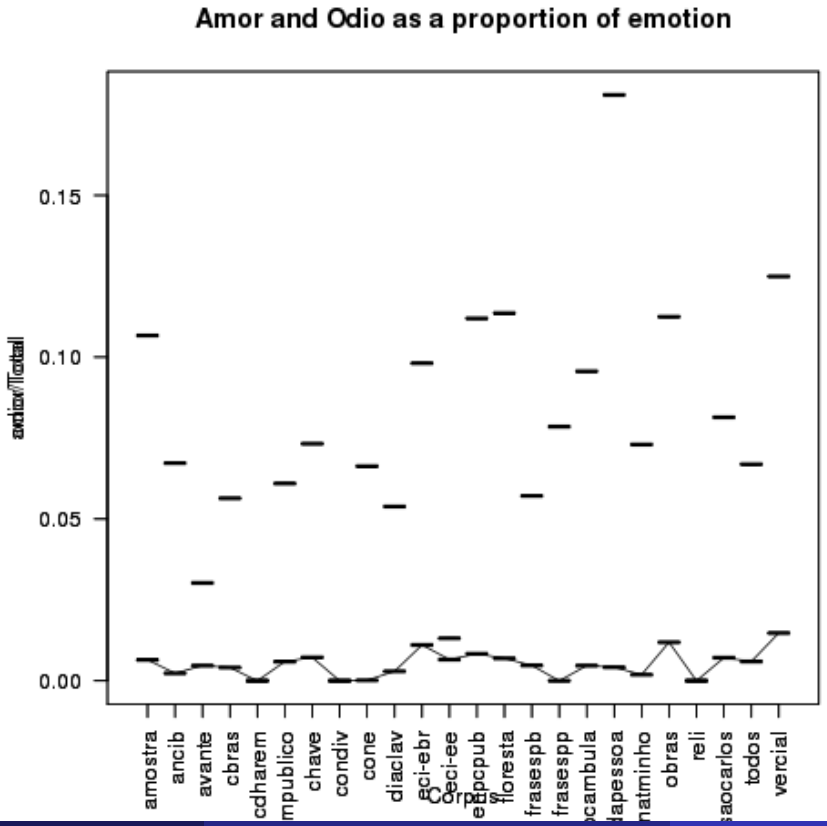





# Exploring emotions: saudade



# Exploring emotions: amor and odio (loving and hating)





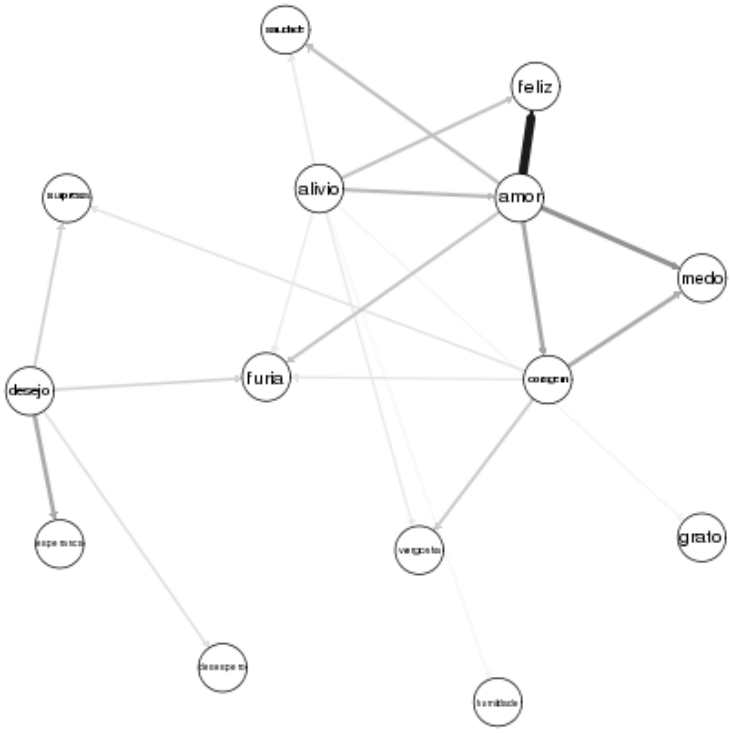

# Emotion network

A novel suggestion, that is on the making: Count co-occurrence of emotion (words).

- For each sentence (pair of sentences?), count emotion co-occurrence.
- Create an emotion network that displays the relationships among emotions: the strength of the link between two emotions is proportional to their co-occurrence.

So far a test only with one of the corpora, Vercial (fiction).

# Emotion network, example from Vercial






# Musings on comparing speech and writing

Inspired by Biber's book *Variation across speech and writing*

- Biber suggests that variation in English can be described by seven factors, which cut across speech and writing
- "No dimension of variation [...] correlates with a simple spoken/written contrast"
- Still, he claims that writing allows more variation than speech.

He uses English and Tukulaelae Tuvaluan as examples of the need to

> *considerable research into the range of speech situations and the functions of linguistic features before attempting a macroscopic analysis. (p. 205)*

It is high time that variation across speech and writing in Portuguese is studied.

## Comparing speech and writing in Gramateca: preliminary ideas

For starters, use these three "indicators"

- vocative and second person use (semantic second person!)
- lexical bundles
- passive (interesting because of three or more forms)

Lexical density, defined as the percentage of open/closed words or inserts, or lexical diversity, defined as the number of different lexical items per text/corpus.
One inspiring work was Biber & Gray (2010) comparing speech and academic prose.



## Concluding remarks

- Still very much in the beginning
- Hope to be able to provide a good service to the community
- Hope to find out some interesting knowledge about Portuguese grammar
- Thank you for your comments!

Keep yourselves posted, by joining the mailing list!

# References


- Agresti, Alan. *An Introduction to Categorical Data Analysis*. John Wiley and Sons, 1996.
- Biber, Douglas. *Variation across speech and writing*, Cambridge University Press, 1988.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & E. Finegan. *The Longman grammar of spoken and written English*. 1999, London: Longman.
- Biber, Douglas & Bethany Gray. "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness". *Journal of English for Academic Purposes*, 9, 1, 2010, pp. 1-82.
- Devore, Jay & Kenneth N. Berk. *Modern Mathematical Statistics with Applications*. Thomson Brooks/Cole. 2007.
- Ellis, John M. *Language, Thought and Logic*. Evanston, IL: Northwestern University Press, 1993.




# References


- Gries, Stefan Th. "Sources of variability relevant to the cognitive sociolinguist, and corpus- as well as psycholinguistic methods and notions to handle them". *Journal of Pragmatics* 52, 2013, pp. 5-16.
- Sampson, Geoffrey. "Thoughts on two decades of drawing tree", in Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, Kluwer Academic Publishers, 2003, pp. 23-41.
- Veiga, Arlindo, Sara Candeias & Fernando Perdigão. "Conversão de Grafemas para Fonemas em Português Europeu – Abordagem Híbrida com Modelos Probabilísticos e Regras Fonológicas". *LinguaMÁTICA* 3, 2, 2011, pp. 39-51.