Comparing oral (transcribed) and written corpora in Portuguese

Diana Santos

ILOS

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

10 April 2014

Introduction

- One first study under Gramateca
- The need to develop a quantitative infrastructure
- The research questions
- Some problems

Gramateca

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

AC/DC: a repository of different corpora

- full syntactic annotation with PALAVRAS (Bick, 2000)
- semantic annotation (colours, emotions, body parts) by Linguateca
- publicly searchable on the web, no strings attached
- currently ca 1,000 million words available

Research questions and initial problems

- Are there non-obvious differences between oral and written speech?
- Or the differences due to communicative function and purpose are more important?
- Is the procedure to address these issues sound and useful?

What are we actually comparing? Written data, anyway.

1. How much is captured by transcription?
2. How much is translated/added into written form?

If we wanted a really unbiased study, we should also have read out aloud some of the written texts and compare factors like intonation, pauses, speed, and so on.

Particular grammar areas

- vocative and second person use (extending Freitas & Santos 2010),
- lexical bundles – turned into bodily multiword expressions,
- passive, extending Santos (2014) – one inspiring work was Biber & Gray (2010) comparing speech and academic prose.

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. (see Biber et al., 1999).

The beauty of oral Portuguese

Written conventions of Portuguese are rather different from those of English, as beautifully pointed out by Bennett (2010).
So it is conceivable, in fact very probable, that oral conventions are also special for Portuguese

- not only what you talk about,
- but how do you talk

... if these things can be separated at all.

"Genre/register/mode" assigned: written

Genre: oral

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

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:

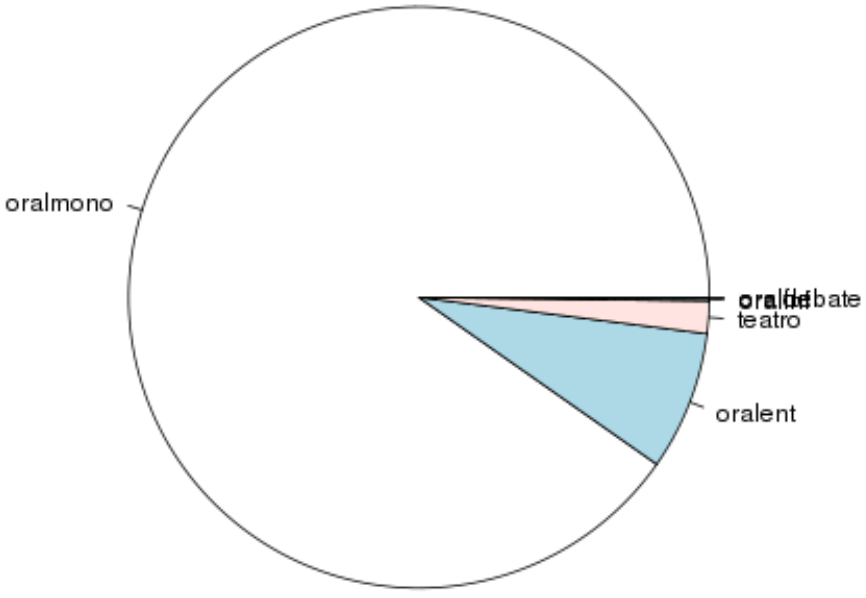
Eu tava lá em & ca + meia-noite e meia, rolando de rir, acordando a vizinhança +
DUD *Primeiro você põe o dedinho, aí cê põe o dedinho e vai forçando, não sei o quê, até ficar nu sei o quê, e aí, vai, depois de, mete bronca, nu sei o quê. Aí no segundo dia cê vai, põe dois dedinhos, nu sei o quê, só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

Examples of the oral material

VERCIAL, theater:

Telmo *Romeiro, quem és tu?*
Romeiro (Tirando o chapéu e levantando o cabelo dos olhos.) – *Ninguém, Telmo, ninguém: se nem tu já me conheces!*
Telmo (Deitando-se-lhe as mãos para lhas beijar) – *Meu amo !*

Distribution of oral material



Dates in the material

Different periods are covered by different materials.

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

Our corpora have, in addition, quite different ortographic conventions.

Annotation

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

- syntax
- emotions
- colours
- body words
- clothing

and, specifically, for specific corpora,

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

One talks and writes for someone

How much does one mention or refer to the other? And how? Is it inversely proportional to how much one mentions oneself? Or do they go hand in hand? Is this a habit of language? Like in *don't you think?*, *kjønner du?* (no-bo), *estás a ver* (pt-pt), or a real mention/connection? How should one measure this? Per number of words? Per turn? Per change of subject? Frequency alone, or distribution along the talk? Operationalization: second semantic person, and first (singular and plural). Problem: reported speech or free indirect speech.

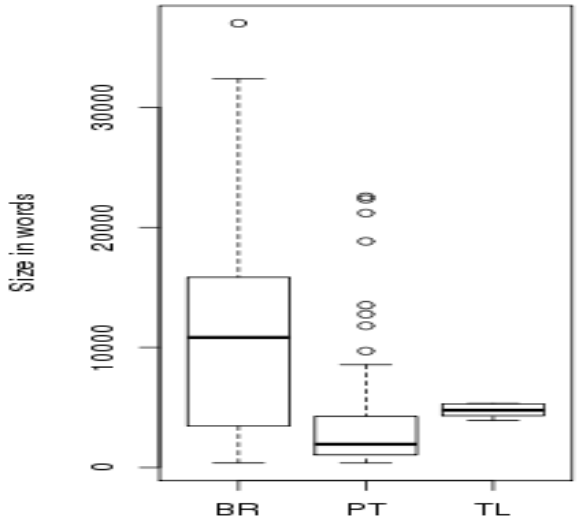
Measuring second person in Portuguese

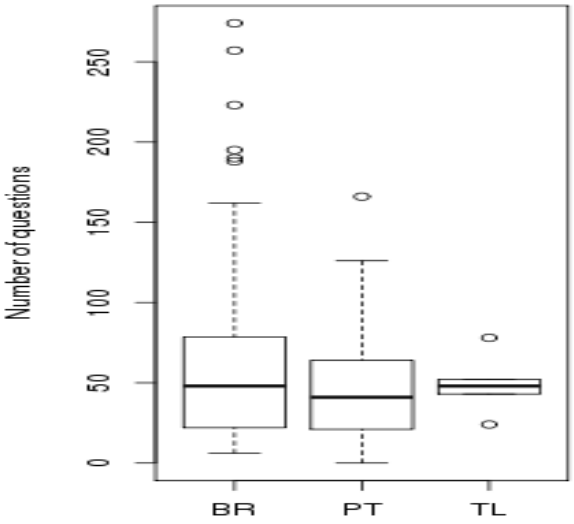
In addition to the respect forms *o senhor, a senhora* in both varieties, it is actually easier in Brazilian Portuguese, because of widespread use of *você*, while it is almost impossible to detect polite second person in grammatical third person in Portugal.
First numbers in Museu da Pessoa:

Pronoun	Total	BR	PT
tu	952	769	267
você	9473	9206	183
vós	51	1	50

Second numbers: Measuring second person in Portuguese

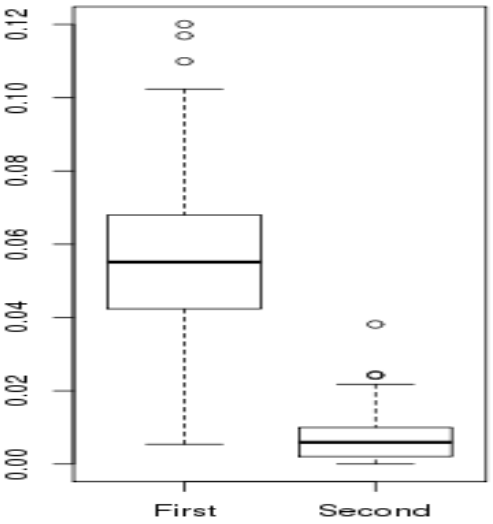
Corpora of interviews, but quite different interviewees, and interviewers.

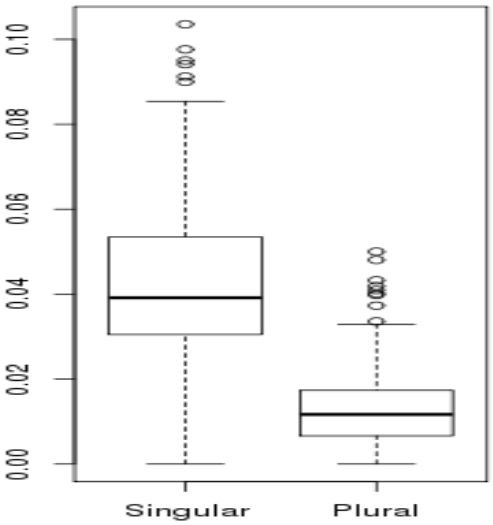




Second numbers: Measuring person in interviews

Relative numbers, per words. First person (including *a gente*), second person (including *o senhor*). Singular and plural compared for first person.





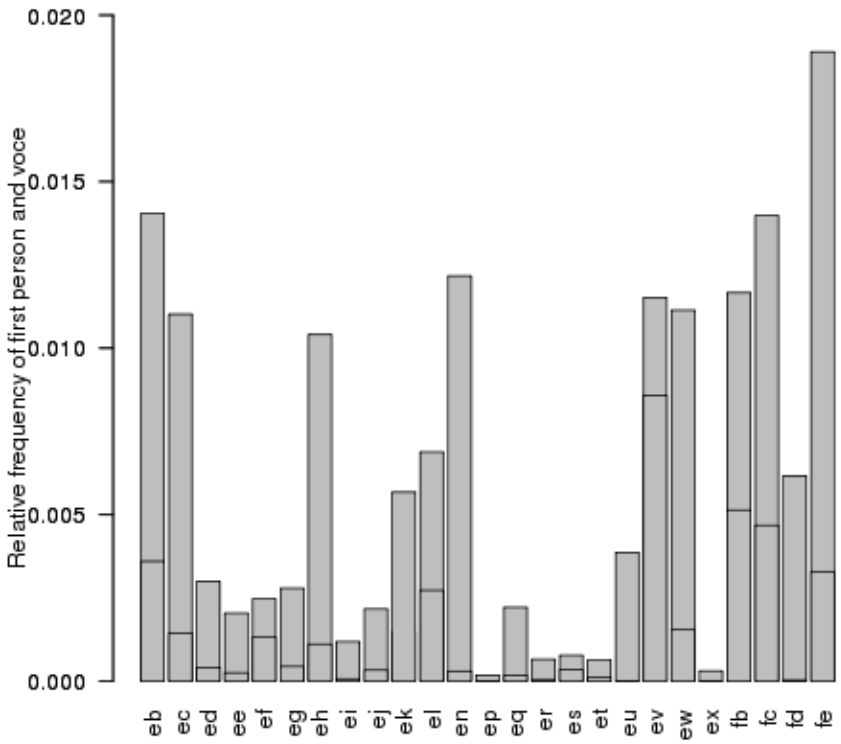
Measuring person in Portuguese in other oral corpora

Other comparisons: In *Corpus Brasileiro*, with four different kinds of oral speech, the percentage of *você* and of first person singular

	Total	você	%	1st person	%
fb	17311	89	0.00514	202	0.0117
fc	1,442,787	6751	0.00468	20169	0.0140
fd	48,963,032	1412	2.88e-05	301761	0.0062
fe	2,772,139	9115	0.00329	52387	0.0189
fa	1050	6	0.00571	-	-

fe: interviews; fd: parlament debates; fb: TV debates; fc: presidential speeches; fa: soccer reports.

Measuring person in Portuguese in other corpora



Comparison among all genres in *Corpus Brasileiro*. (F – falado, E – written.) Short stories (eb) have significantly higher cases of first person than most oral genres, and the highest relative frequency of *você*.

Musings on comparing speech and writing

We all know this... Biber's book *Variation across speech and writing* (1988)

- 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)*

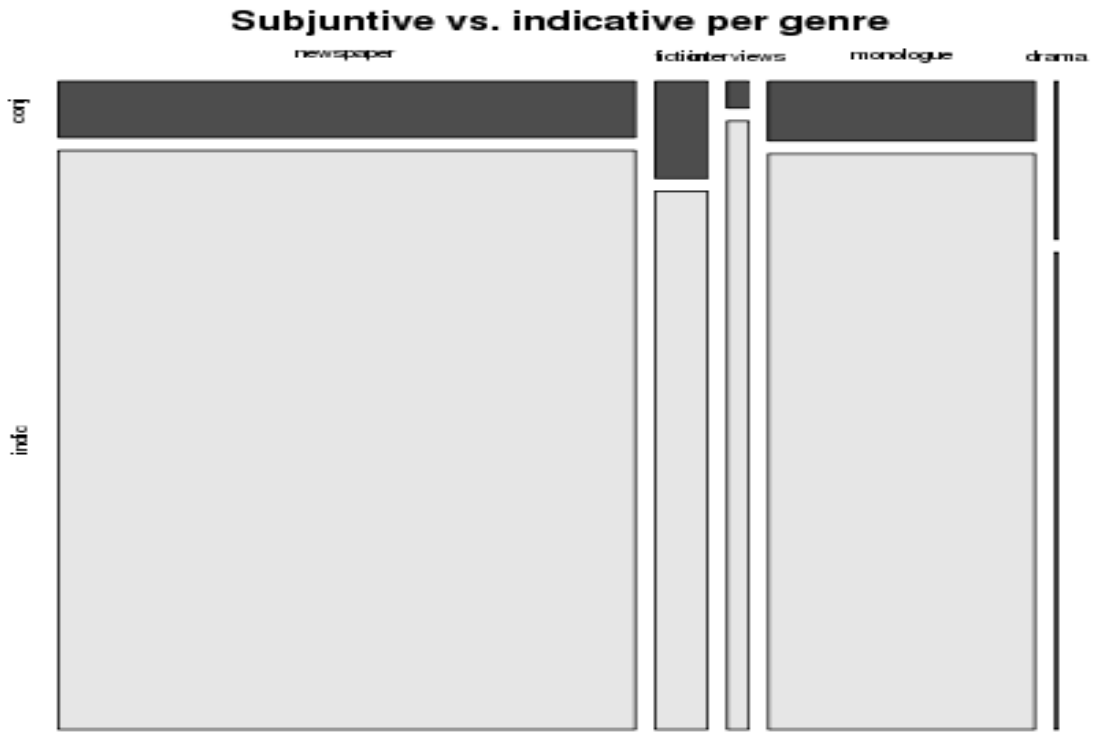
It is high time that variation across speech and writing in Portuguese is studied.

Considering forms of addressing others

Address, the recipients of our speeches or debates or monologues, are obviosly of keen importance! But, if there is an area in Portuguese fraught with difficulties and subtleties, it is precisely how to address others – and it is not easily measurable by second person occurrence.

- Would a president refer to his/her countrymen as *vocês*?
- Would a member of Congress addressing the audience or the president of Congress use *você* or *tu*?
- When using first person plural, would s/he mainly refer to her/his group/party (exclusive we), or the whole country (inclusive we)?
- When using inclusive we, would it be mostly descriptive, or imperative/exortative? This is possible to measure grammatically by the mood (indicative or subjunctive).

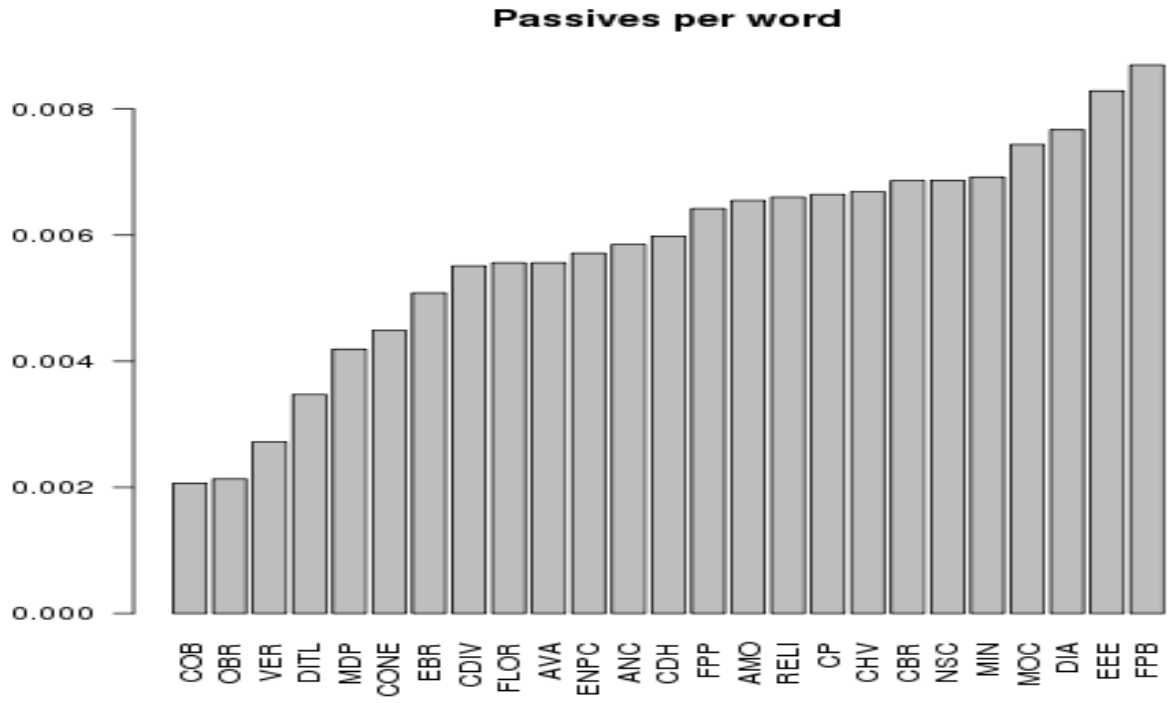
Considering forms of addressing others, first person plural



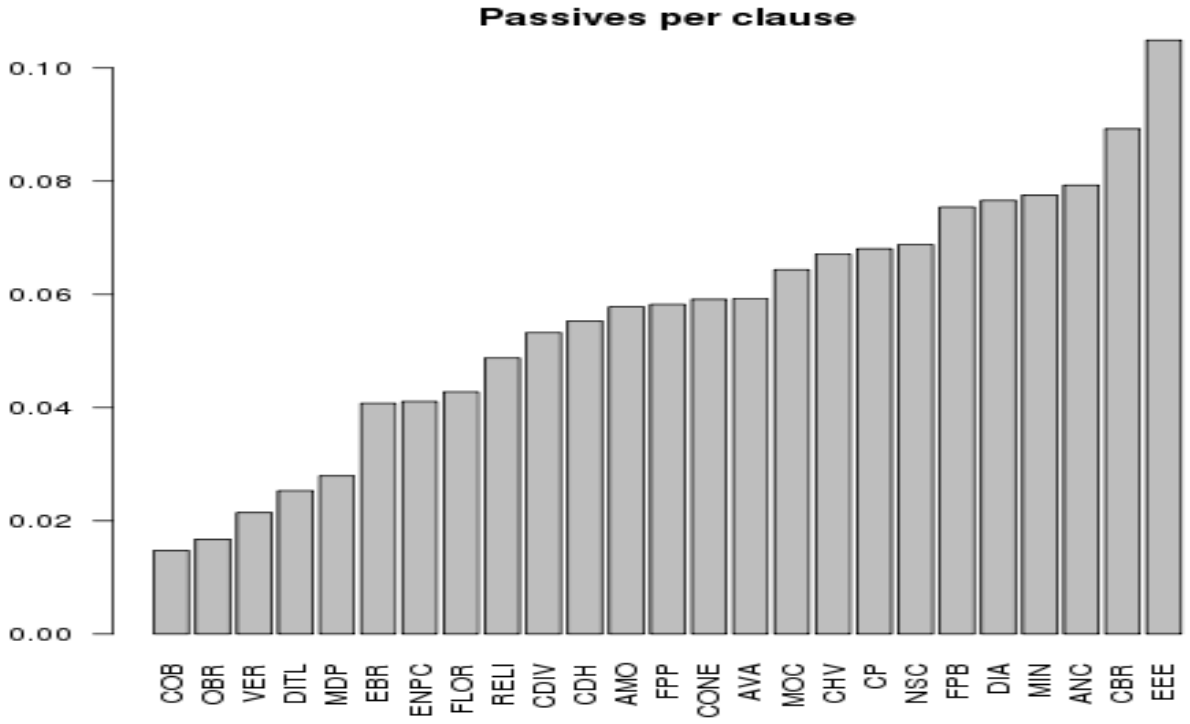
Use of passive

Passives are usually considered an elaborate way of expression, fact directed by dispromoting agents/subjects.
But, how to measure passive frequency?
Passives per number of words is not a linguistically sound measure, because not all words can be passives. Not even passives per number of verbs is sound, because not all verbos (in a verbal group) can be passive.
So, ideally, we should use the number of verbal clauses.
Then, an important information is what is considered / counted as passive.
We have used the most encompassing definition (except that we did not include *se* passives), by accepting *estar* and *ficar* as well as *ser*, although this is probably one case where different genres have different kinds of passives.

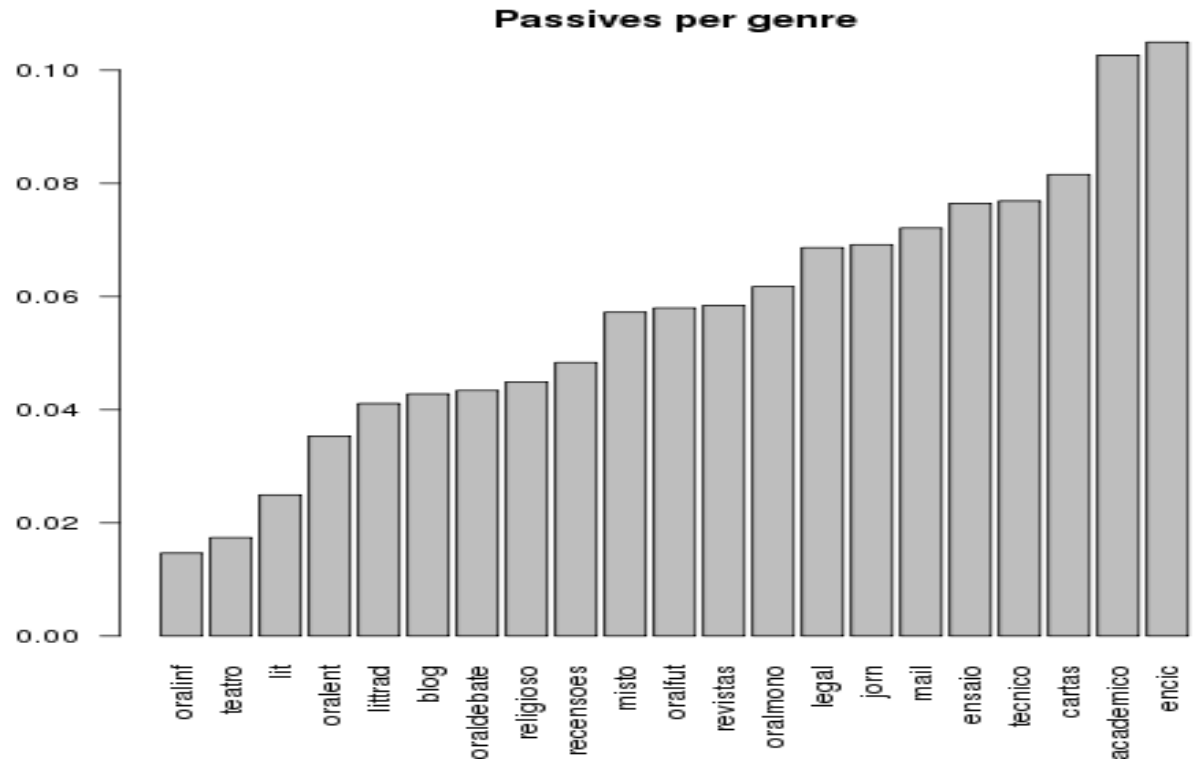
Use of passive: passives per word



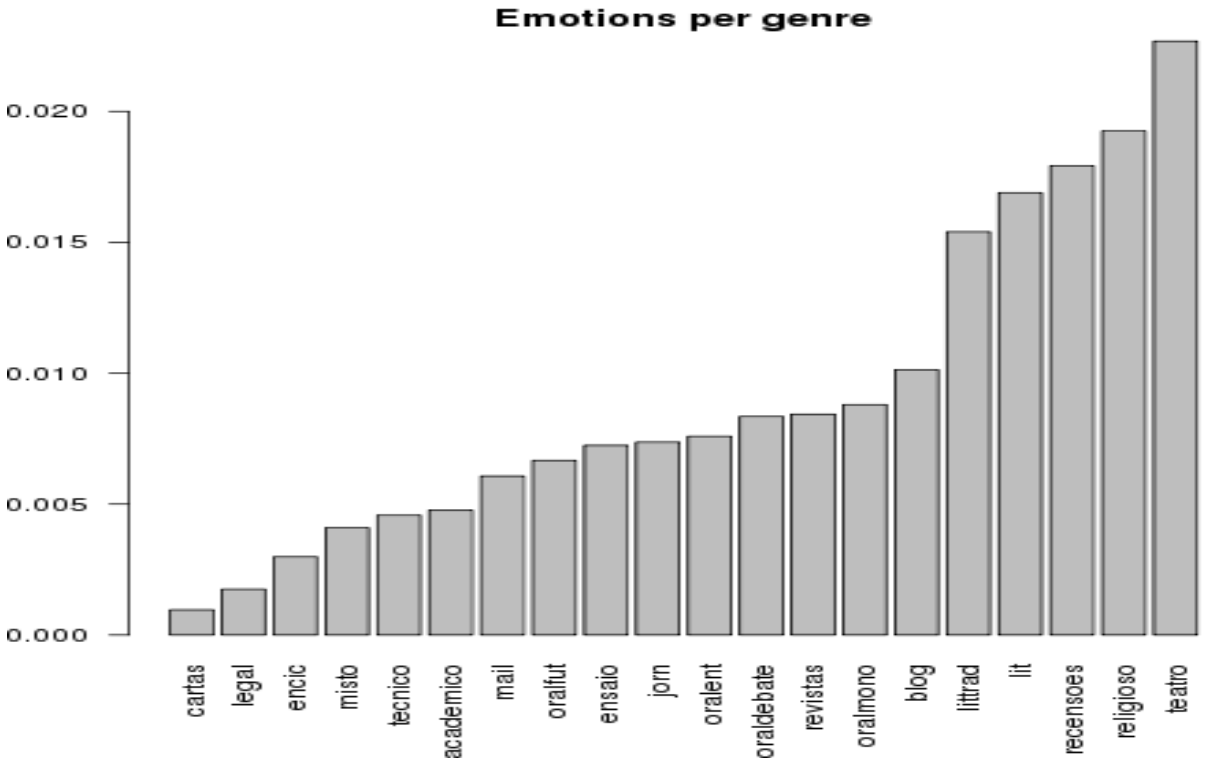
Use of passive: passives per clause



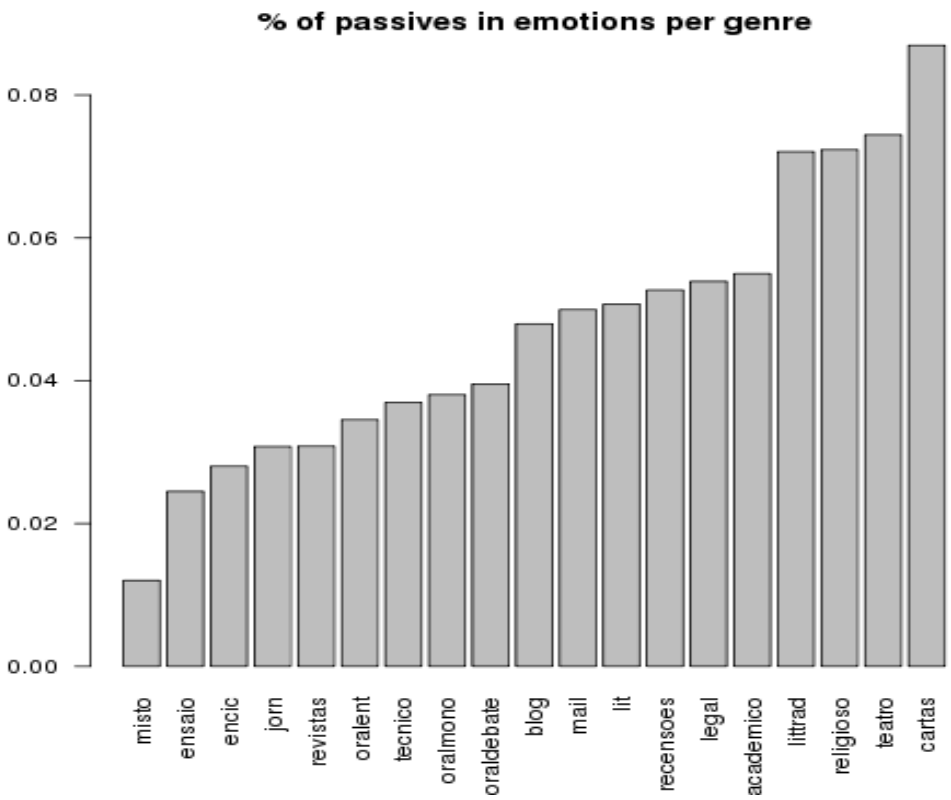
Use of passive: oral vs. written



Use of emotions



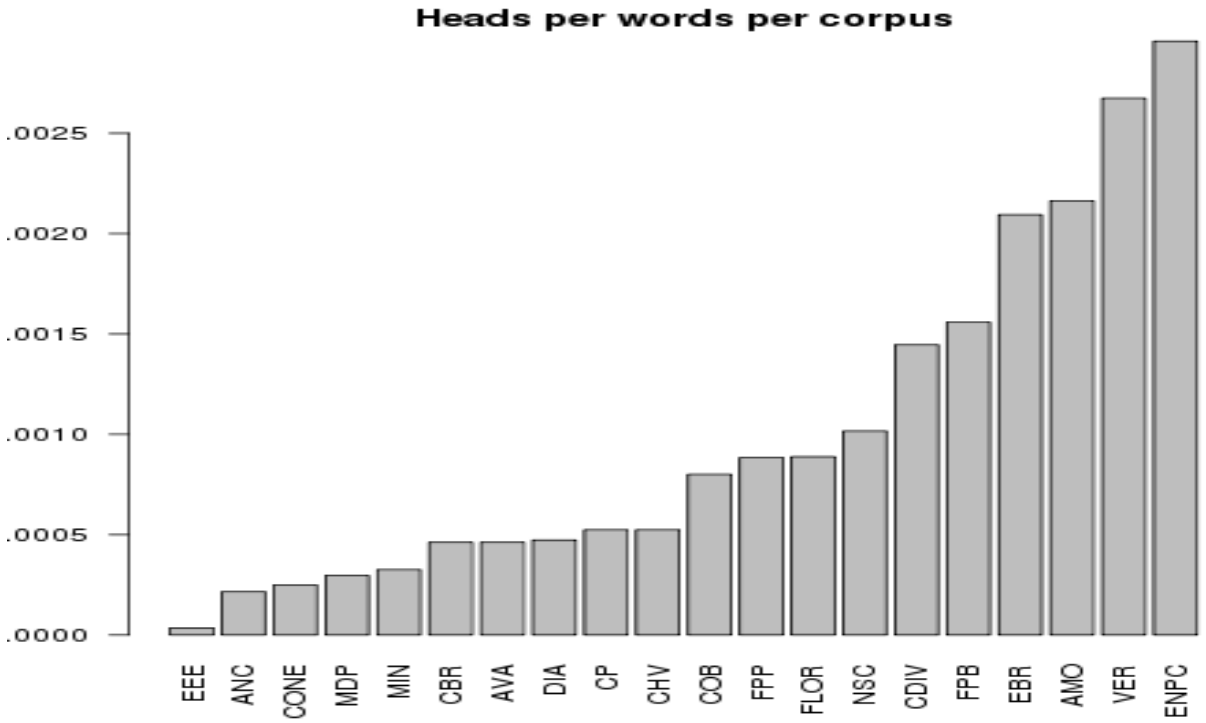
Use of emotions: beware!



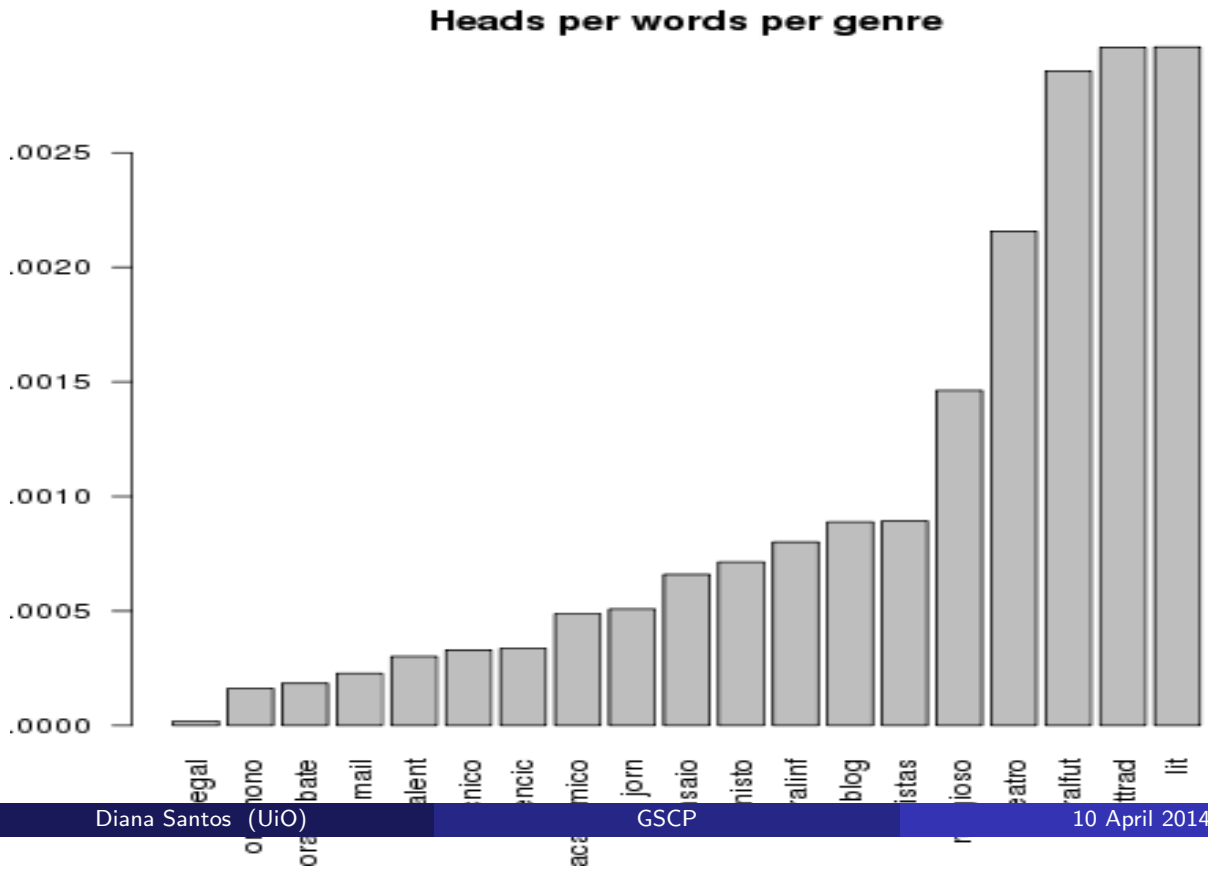
Words like *realizar*, *reconhecer*, *apreciar*, *confiar* have a legal or other meaning... And only two instances in letters...

Comparing body language

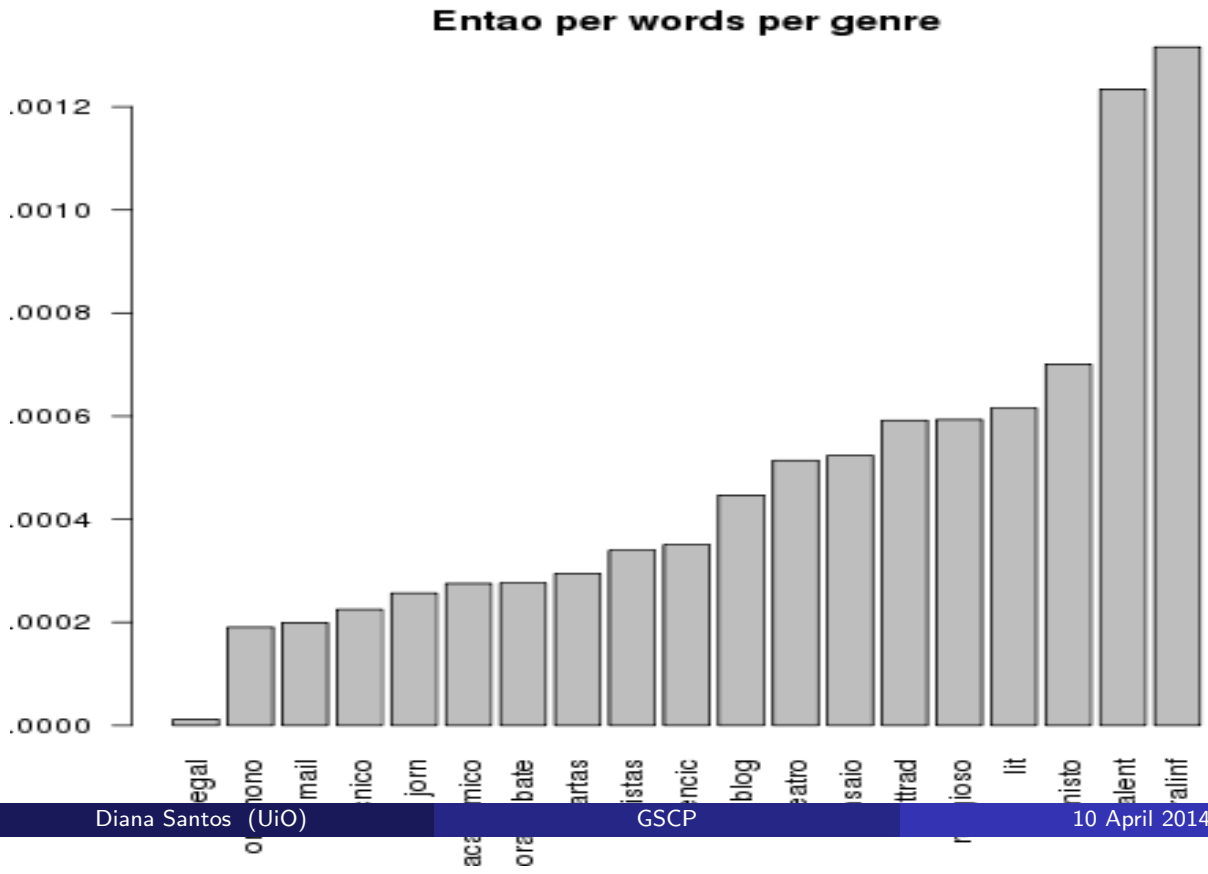
Reference to head (any part of the head)



Comparing body language: head



Comparing lexical items: *então*



Concluding remarks, about this work and about Gramateca

- 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!