QUEMDISSE?: Reported speech in Portuguese

Cláudia Freitas, Bianca Freitas & Diana Santos

ILOS

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

24 May 2016

Why study reported speech?

- News: over 90% of the sentences include a quotation (Bergler et al., 2004)
- (in)direct speech provides clues to point of view detection, cf.
  - *The Major* **advised** *the IOC of delays in public works*
  - *The Major* **admitted** *to the IOC that delays may occur public works*
- Who is being silenced? Only 23% of women have lines in action movies (Smith et al., 2014)

Local motivation?

- Reported speech verbs are highly relevant for translators
- Brazilian publishers advise translators to vary occurrences of "say" in the Portuguese translated text
- Initial question for MA thesis of Bianca Freitas: Which verbs can be used, in Portuguese, to translate "say" in reported speech contexts?

Differences between English and Portuguese

- Graphical structure of direct speech is different. Portuguese favours dialogue, with turns clearly marked, while English prefers monologue, inside a larger paragraph.
- Especially after the direct speech, almost any emotional or atitudinal verb is possible in Portuguese, provided it identifies the speaker, something which is not the case in English.
- Difference between thought and speech is marked differently in Portuguese.
- Marked difference in conventions in different genres in Portuguese: newspapers used English conventions from the start; literature continued the tradition or evolved into more radical forms like Saramago's invented punctuation.

This shows that translation betwen the two languages can be more involved than one might imagine.

Gramateca

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

AC/DC

The oldest Linguateca project: Acesso a Corpos / Disponibilização de Corpos, `http://www.linguateca.pt/ACDC/`

- Since 1999 – presented at LREC 2000 in Athens
- It provides the Portuguese-language community with access to many corpora, all parsed by PALAVRAS (Bick, 2000) and with semantic annotation (Santos & Mota, 2010; Santos, 2014)
- It has a set of search services associated (Simões & Santos, 2014)
- It is built on top of OpenCWB, `http://cwb.sourceforge.net/`
- 28 corpora, ca. 1.5 billion words

The Mariano Gago corpus

One of the most recent corpora in AC/DC is the Mariano Gago corpus (Santos, 2015): `http://www.linguateca.pt/CorpoJMG/`

- In honour of Mariano Gago
- the language minister
- who supported Linguateca
- and died in April 17, 2015

Category	Size
news	143,000
speeches	12,000
interviews	31,000
other	75,000
hommage site	43,000

Concrete results

Here, we will present

1. a lexicon for speech verbs in Portuguese
2. a set of patterns for finding quotations
3. a fully revised corpus (MG) as far as quotation is concerned, available also through the AC/DC interface, `http://www.linguateca.pt/acesso/corpus.php?corpus=JMG`
4. quantitative data on the process and the result

The lexicon

We divided the speech verbs according to their possibility of being used (or not) in direct or indirect speech:

- Verbos de dizer (speech verbs), that refer to any kind of saying or doing with words
- Verbs that are used with direct speech – interestingly, most of them are not speech verbs, but convey feelings or bodily actions that one can do while speaking
- Verbs that can be used with indirect speech, and which are a (much smaller) subset of the previous set, and may be vague between a cognitive verb or an utterance verb, such as *defender*, *ele defendeu que...*

These classes are not mutually exclusive. See `http://www.linguateca.pt/Gramateca/VerbosDizer.html`

The annotation

In context, we wanted to be able to identify

- direct quotation (tag `dizer-relatoDIRETO`)
- indirect quotation (tag `dizer-relatoINDIRETO`)
- mixed quotation (tag `dizer-relatoMISTO`)

To understand better how this semantic field works, we also identified speech verbs without quotation. (tag `dizer`)

Examples of indirect quotation

In Portuguese, indirect quotations can have the following syntactic forms

- THAT-clause: *Cauteloso, ele disse que não receberá empresários e empreiteiras.* ('Cautiously, he said that he won't entertain entrepreneurs or building companies')
- INF-clause: *Em entrevista de dez minutos à TV russa, ele disse estar controlando o país.* ('In an interview of ten minutes for the Russian TV, he said he was having full control over the country.')
- PP-clause: *O governo de Israel se disse surpreso com críticas do enviado do Vaticano a Israel, Andrea Di Montezemolo.* ('The Israeli government confessed its surprise over the criticisms of the Vatican attaché in Israel, AdM.')
- according to: *Afinal, como disse Boris Casoy, Hebe paga impostos e é assídua no trabalho*. ('After all, as Boris Casoy said, Hebe pays her taxes and comes regularly to work.')

Quantitative data for the JMG corpus

In ca. 280,000 words, ca. 12,000 sentences:

	Number	Types
Verbos de dizer	1,760	157
Direct reporting	248	40
Indirect reporting	270	44
Mixed reporting	313	37

- Surprising large amount of saying verbs in the first person: 65 in 836 reporting cases (7.8%), and 294 in 1836 non-reporting ones (33.9%)
- Many non-standard cases: conditional, subjunctive, modals, indefinite subjects, self-quotation...

Annotation process

- Annotate with the lexicon and the basic patterns
- Revise the rules and the lexicon
- Correct parsing errors or corpus errors
- Create specific rules for particular cases
- Repeat the whole process

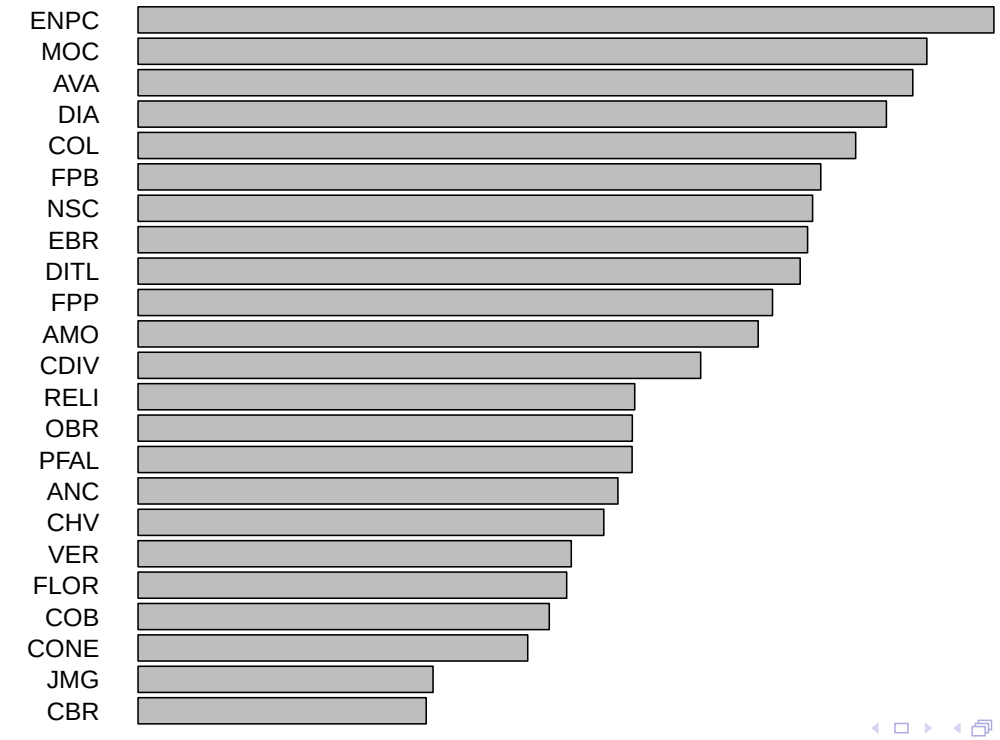
Comparing the annotation before and after revision

In order to give an estimate of the work involved:

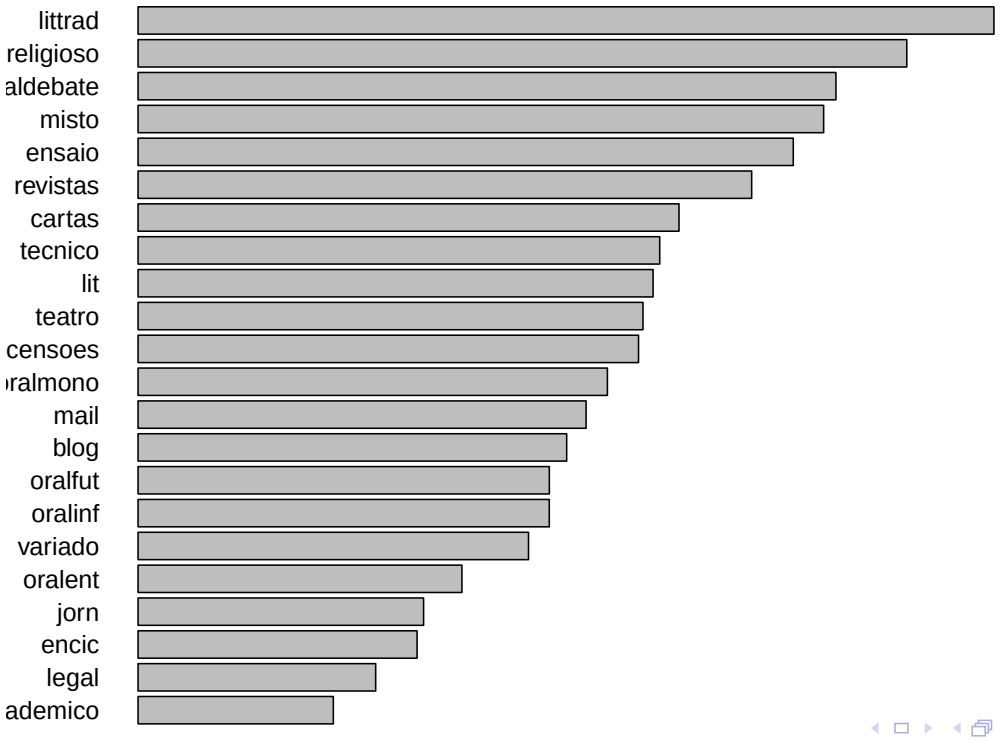
	Number before	Types before	Number	Types
Verbos de dizer	2,510	220	1,760	157
Direct reporting	208	11	248	40
Indirect reporting	362	31	270	44
Mixed reporting	371	20	313	37

- 20 correction rules
- 48 negative rules
- 189 correction rules

Proportion of saying verbs per corpus



Proportion of saying verbs per "genre"



Quotation and emotion: preliminaries

Often speech is reported together with the mood or feelings of the speaker, so it is interesting to investigate this.

- some words can mean both an atitude and a speech act: *prometer* (promise), *ameaçar* (threaten)
- some words describe speech with an emotional load: *queixar-se* (complain)
- some emotional words replace in Portuguese the speech verb altogether: *alegrar-se* (get joyful)
- some bodily actions conveying emotion replace in Portuguese the speech verb altogether: *rir* (laugh)

Work in progress.

Concluding remarks

- We offered some resources that may be interesting for the community
  - to investigate further quotation in Portuguese
  - for training and testing quotation extraction systems
  - for studying translation in this respect
- We are still planning to develop a system called QUEMDISSE on top of AC/DC to extract quotations from parsed corpora and that can be used by general users
- We also hope that this paper can foster interest in direct speech in other languages, not as an adaptation of English