QUEMDISSE? Reported speech in Portuguese

Cláudia Freitas, Bianca Freitas, Diana Santos
PUC-Rio & Linguateca, PUC-Rio, Linguateca & University of Oslo
claudiafreitas@puc-rio.br, bianca.frejes@gmail.com, d.s.m.santos@ilos.uio.no

Abstract

This paper presents some work on direct and indirect speech in Portuguese using corpus-based methods: we report on a study whose aim was to identify (i) Portuguese verbs used to introduce reported speech and (ii) syntactic patterns used to convey reported speech, in order to enhance the performance of a quotation extraction system, dubbed QUEMDISSE?. In addition, (iii) we present a Portuguese corpus annotated with reported speech, using the lexicon and rules provided by (i) and (ii), and discuss the process of their annotation and what was learned.

Keywords: quotation verbs; reported speech; corpus annotation; Portuguese

1. Introduction and Motivation

A considerable amount of language activities involves reporting what others have said. In certain contexts, such as the journalistic discourse, the use of reported speech is crucial. (Bergler et al., 2004) found that there are pieces of news in which over 90% of the sentences include a quotation.

In natural language processing (NLP), automatic identification of reported speech is called Quotation Extraction (QE), which aims to identify quotations in text, and relate them to their authors. This is a task often associated, or subsidiary, to sentiment analysis, and it is the distinctive subtask for "opinion-oriented information extraction" (Pang and Lee, 2008). Its focus is to identify what is said, who said it, and the speaker's and the writer's judgments on what was said.

New contexts for the exploration of reported speech have recently been provided by research in Digital Humanities, see (Mambrini et al., 2012). Through reported speech, one gives voice to different characters both in a piece of news or in fiction. Consequently, both the presence and the absence of particular characters indicate choices by the text author. By means of quotation identification, it is possible to measure which characters are given voice and, by contrast, which ones are silenced. (Smith et al., 2014), researching female characters in popular movies in 11 countries, established, among other findings, that only 23% of these women have lines in action movies. Exploring large amounts of texts – either news or fictional works – and who they quote or silence may provide us with new findings about our society.

Given a relatively regular structure of reported speech – although quite different in different languages, and even across varieties, see (Santos, 1998) for some discussion –, rule-based approaches to QE are often extremely successful. However, purely formal marks that indicate the presence of a quotation, such as quotes in English, are not unique to this purpose, hence recognizing the specific verbs that are used in these contexts is highly relevant. Additionally, not all reported speech has the aforementioned formal marks. Indirect quotations, constituting almost half of the reported speech in English news (Pareti et al., 2013), are more difficult to identify, and not always covered by QE systems. On the other hand, 96% of the clues for reported speech in English found by (Pareti et al., 2013) are verbs, which makes us conclude that a lexicon of reported speech verbs[1] is of great value for quotation extraction.

2. Discussion

As everything linguistic, there is not an easy way to decide what is and what is not a given phenomenon. Furthermore, different languages give more or less (and always different) attention to whatever one chooses to investigate.

Language norms, in addition, are cultural norms, not absolute laws, and different written cultures (especially those influenced by translation) are especially prone to change and to experimenting in different ways.

So, the presentation of direct speech in Portuguese is a case in point, from a rigid separation between oral scenes and narrative text to Saramago's prose and free indirect speech in lusophone literature, to the influence (and victory) of a completely different graphic form of conveying it – the anglophone one, using quotes, which is used overall in Portuguese-speaking countries in newspaper text.

Interestingly, the punctuation of direct speech is one of the areas in which American English differs most from British English, see (Jones, 1996); see also Hofstader's famous complaint when reading the quintessential American Salinger in British clothing (Hofstader, 1997).

News is probably the text genre where Portuguese suffers more influence of globalization (and global English) and, therefore, where an anglophone style is more pronounced and influential. In fact, it is probably uncontroversial to say that in Brazil and in Portugal it changed completely into English style: using quotes as direct speech, or rather, direct quotation.

The focus and raison d'être of direct speech in narrative fiction (or even non-fiction) is obviously different: while a fiction text tries to reproduce or create an oral exchange[2], a news text is, on the other hand, interested in assigning responsibility of an utterance to other (identified) actors. Instead of a dialogue or a conversation among several speakers (whose turns are indicated by long dashes in fiction),

[1] There is a larger class of speech verbs, but not all of these are used to report speech, hence the full name *reported speech verbs*. In addition, a particular speech verb can sometimes be used to report speech, and other times not: for example, *falar* in *Ele falou alto* ('he spoke loudly') vs. *Ele falou que viria* ('he said he would come').

[2] Although a fictive one, see (Brumme and Espunya, 2012) for more on this subject.

we have a report for responsibility assignment. It is not the colour, the dialect or the emotion that is at stake, but what was said for "objective" coverage.

In fact, it is this accountability issue that makes the study of reporting and (direct and indirect) written speech relevant for a wider audience than literary experts or language learners. In a time of extremely short-lived fame, "who said that" is mostly relevant if it is modified by, or served as "has just said so", and one needs automatic reporting boots.[3]

On the other hand, it should be stressed that our interest is also linguistic, in the sense that we wish to define and study a lexical field, that of language-talk-speech, that contains the words related to this (central) property of mankind. Not only from a lexicographic perspective, but also as a semantic and contrastive topic, given the repeated statements that Portuguese differs widely from English in this respect (Caldas-Coulthard, 1996), and, interestingly, also from Arabic, as noted in the translator's preface in (Jarouche, 2013).

3. Previous Work

In their study of human tagging in English, (Bruce and Wiebe, 1999) already showed that attribution was hard, and in (Wiebe et al., 2003) the problem of automated opinion mining is further discussed.

Drury and colleagues compiled a large quotation corpus in the financial domain (Drury et al., 2011; Drury and Almeida, 2012) and used it to identify trends in that domain.

Based on another English corpus with 18,000 citations, (Pareti et al., 2013) described several machine learning experiments to identify indirect and mixed quotes. The authors did not use a specific verb lexicon in their work, they developed a classifier to detect verbs that introduce quotations.

In order to build a quote extraction system, (Sagot et al., 2010) focused on creating a lexicon of reported speech verbs in French, dealing primarily with direct quotes introduced by appositional clauses and headed by a quotation verb. Their work influenced heavily our initial attempts in this area.

As far as Portuguese is concerned, there are a number of works in this area, too. (Sarmento and Nunes, 2009) proposed the VERBATIM system, using a lexicon of 35 reported speech verbs and 19 lexico-syntactic patterns, while (Fernandes et al., 2011) used machine learning techniques to identify quotations and correctly assign their authors to them for the GloboQuotes corpus, created specifically for this purpose.

The most relevant system to date is probably the EMM News explorer (Pouliquen et al., 2007) which extracts and displays quotes in several EU languages and is in use for the general public today at `http://emm.newsexplorer.eu/NewsExplorer/`.

Outside the narrow context of QE, there is obviously a lot done in reported speech. For example, in the context of parsing, (Bick, 2000) used a specific (lexical) class for speech verbs in PALAVRAS; and we are aware of at least another corpus-based work on translations of this kind of verbs, (Loffredo et al., 2004).

4. Contribution of our Work

In this paper we report on

- the delimitation and clarification of the phenomena of interest: speech and reporting speech;
- the creation of a specific reporting speech verb lexicon, with three major classes (note that the classes are not mutually exclusive);
- the creation of patterns to identify and annotate quotations in corpora (as the basis for subsequent extraction and assignment)
- the quantitative results of fully annotating (and revising) a small corpus of mainly newspaper text, as well as annotating larger materials.

5. What we are talking about

In our work we distinguish between

- Verbs whose meaning refers to some form of saying, what we call VERBOS DE DIZER. They can refer to intensity, sound, difficulty – like *gritar* ("shout"), *rugir* ("browl") and *gaguejar* ("stammer"), respectively[4] – or any specific task or speech act you can do with words, such as *prometer* ("promise"), *pedir* ("ask"), *proibir* ("forbid") or *rezar* ("pray").
- Verbs which are acceptable as conveyors of direct speech (and therefore appear with it), but whose meaning is either more general, judgemental, or refers to things that one can do while talking. Examples are, respectively, *brincar* ("joke"), *entusiasmar-se* ("became enthusiastic") and *rir* ("laughed"). They might not be considered speech verbs, but they appear frequently as direct speech indicators. Or, one might say, instead of the obvious speech verb.
- Verbs which are acceptable as conveyors of indirect speech (and therefore appear with it). As in the previous group, they might not be considered speech verbs, but they appear frequently as indirect speech indicators. Examples are *defender* (to defend) and *lembrar* (to remind).

We call (and tag) each case separately, as DIZER, DIZER-RELATODIRETO (direct reporting) and DIZER-RELATOINDIRETO (indirect reporting).

While the only relevant cases for quotation extraction are those that involve some kind of RELATO, we thought it also convenient from a linguistic point of view to identify the full range of speech verbs, especially when we noticed that, contrary to our initial expectations, most of the speech verbs (DIZER) do **not** occur in reported speech.

[3] Note the discussion about Twitter's future (on being real time or not) in the beginning of 2016.

[4] This is incidentally similar to what (Snell-Hornby, 1983) calls "verb-descriptivity".

Before we discuss how we collected our lexicon and the actual patterns, we present authentic examples of the kinds of contexts we found in corpora, encompassing: direct quotation, indirect quotation, and a third category, mixed quotation (tagged as DIZER-RELATOMISTO), in which the reporter makes minor (or not so minor) interventions in the text but still gives direct access to at least parts of the original utterance.
The first three examples show direct quotation, the next four indirect, and the last one shows an example of mixed quotation.[5]

1. *"Talvez tenha sido mal-interpretado", disse.* (' "Maybe he has been misunderstood", she said.')
2. *Até que uma amiga minha passou por ele e disse: "Oi, Fábio".* ('Until a friend of mine passed by and said "Hi Fabio",')
3. *"Se não tivessem sido feitas", disse, "Portugal era hoje um país ao nível do Leste Europeu".* ("'If they hadn't been done," he said, "Portugal would be at Eastern European level."')
4. *Cauteloso, ele disse que não receberá empresários e empreiteiras.* ('Cautiously, he said that he won't entertain/receive entreprenours or building companies')
5. *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.')
6. *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.')
7. *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.')
8. *O PP, por meio de Rocco Buttiglione, respondeu, aconselhando a leitura de "Mein Kampf" (Minha Luta), de Adolf Hitler, para entender porque o líder da Liga, senador Umberto Bossi, "age com um Führer (guia, como os alemães chamavam Hitler)".* (PP, through Rocco Buttiglione, answered by suggesting that they read Adolf Hitler's Mein Kampf to understand why the Liga's leader, senator Umberto Bossi, "acts as a Führer (guide, as the Germans called Hitler)".')

Examples 4 and 5 ilustrate finite and infinite subordinate clauses, while 6 shows reflexive use with adjectival clauses and 7 is an example of explicit adverbial subordinated clauses (like "according to"-phrases in English, which are however a nominal and not a verbal construct).

[5]On purpose, all examples concern the verb *dizer* ('say'), in order to achieve a more systematic presentation. Any other verb of the corresponding classes could be used.

5.1. Refinement of the reported speech definition

Contrary to what QE works may suggest, the identification of reported speech is not unequivocal. The unclear cases typically relate to

- The use of conditional as hedge: *Eu diria que...* ('I would say that...'): *Hoje* **diria** *que há um movimento que se gera a partir do Me e da movimentação de base que existe no Técnico.* ('Today I **would say** there was a movement that came to life through the student movement and the base agitation that existed in IST.')
- The unactualization of reported speech, as in *se disserem* ('if they say...'), *que diga*, *dirá?* ('will he say?'). In all these cases the actual saying is not presupposed: *Só se tem efeito formativo real e se consegue colocar know how na prática profissional das pessoas se* **disser** *que tem aqui uma actividade...* ('You only have a real pedagogical effect (...) if you **say** that you have here an activity...')
- The use of the modal *poder* ('can') as hedge: *Pode-se dizer* ('it can be said') and then saying it: **Poderíamos dizer** *que foi mais tempo Ministro que qualquer outro político no pós-25 de Abril.* ('We could **say** he was a minister longer than any other politician after the 25th April.)
- The frequent omission of the **sayer** but not of what was said: **Dizia-se** *que os estudantes tinham enlouquecido e só faziam coisas aberrantes.* ('It was **said** that the students had become mad and only did wild things.')
- The presence of a report verb in the inflected in the 1st person, present tense: Are we reporting when we say "I say that and that" or is this the actual saying, not reporting? *Hoje* **digo** *que a culpa foi minha.* ('Today I **say** that it was my fault.')
- The presence of negation: *Há quem o veja como candidato presidencial e ele* **nunca disse** *que não.* ('Some people see him as a presidential candidate and he **never said** that he wouldn't be one.')
- Nominalization of speech, like in *ele* ***falou*** *da sua promoção* ('he **spoke** about his promotion'), that could have been uttered to report a *Fui promovido!* ('I was promoted!'). It is somehow reporting, but more condensed. That the boundaries can be blurred is obvious in the following example, that illustrates a kind of mixed quotation which is hard to identify because there are no indirect speech markers at all. *Em entrevista telefónica na TVI 24, o atual comentador* ***falou*** *de um homem com "uma inteligência vastíssima", com grande "empenhamento na ação cívica" e um "incansável combatente da ignorância".* (On a phone interview in TVI 24, the present commentator **talked** about a main with "a very wide inteligence", with a strong "commitment for civical action" and an "untiring fighter against ignorance".)

In all these cases, we considered the presence of a reported speech.

5.2. Further challenging examples

Although we cannot give here an overview of annotation difficulties, we would like to discuss some problems:
First, the definition of `relatoMISTO` is not as clear as previously expressed, as the two following examples show:

- *Numa conferência sobre o 25 de Abril, há um ano, disse que em Portugal fez-se muito pela investigação: "Não conheço país nenhum que tenha conseguido o feito de Portugal".* (In a conference about the 25th April, one year ago, he said that Portugal had done a lot for research: "I do not know any country where as much as been achieved".)

- *A última vez que estivemos juntos disse que queria vir à Madeira visitar o M-ITI e perceber as coisas "estranhas" que andamos a fazer.* (Last time we were together he told us he wanted to come to Madeira and see the "strange" things we are doing.)

The first could be classified as an example of an indirect speech followed by a direct quotation, while the second requires that we interpret the quotes as the original words and not a ill-chosen word from the writer.
Second, what is the exact scope of what is stated or hinted is often not clear.

- *O DN sabe que Gago sofria de cancro, mas terá sido vítima de morte súbita, segundo disse à Lusa a sua secretária nos últimos 30 anos, Maria José Miguel.* (DN knows that Gago had cancer, but he was victim of sudden death, according to his secretary for the last 30 years, MJM.)

In the previous sentence, what did the secretary actually say? Was it overlapping with what the newspaper already knew, or not? And in the next, we classify differently the first speech verb, *lembrar* as MISTO and the second, *dizer*, inside the relative clause, as simply indirect.

- *Em declarações telefónicas à Lusa, a partir de Díli, Nuno Crato* **lembrou** *o antigo ministro da Ciência, que* **disse** *conhecer desde finais da década de 60, como um "homem que dedicou a sua vida ao desenvolvimento, à divulgação e à promoção da ciência em Portugal"* (On the phone from Dili, NC recalled the previous Science minister, whom he claimed to know since the end of the 60s...)

An interesting case is when it is presented as direct speech, but we know from the context that it must be a translation:

- *Muitos cientistas estrangeiros perguntam-nos: "Vocês têm uma agência só para a cultura científica?"* (Many foreigners ask us: "do you have an agency dedicated solely to scientific culture?")

6. The lexicon of reported speech verbs

In this section, we report on the construction of the verb lexicon. More detailed information is to be found in (Freitas, 2015).

6.1. Gathering reported speech verbs to be used as seeds

In order to detect the target verbs, we started with COMPARA (Frankenberg-Garcia and Santos, 2003), a bidirectional parallel corpus of English and Portuguese. COMPARA contains original and translated (fiction) texts in these two languages that have been linked together sentence by sentence. We used COMPARA to look for Portuguese translations of the English form *said*. We chose the past tense because quotations are commonly reported in this tense. Given that the identification of a given verb as a reported speech verb is not always straightforward (for example, *imaginar* ('to imagine') or *interromper* ('to interrupt') can be used in many other senses), the use of translations of *said* as a starting point aimed at ensuring that the Portuguese verb did occur in a reported speech context:

Source text *'Don't,' I* ***said****, in a muffled voice.*

Target text *– Não?! –* ***interrompi*** *(lit. to interrupt) numa voz abafada.*

A set of one thousand translations were analyzed manually, providing us with 58 different Portuguese verbs.

6.2. Expansion of the original verb lexicon

In order to enlarge the list of verbs, we then chose six verbs out of the original 58 to work as seeds in a larger Portuguese monolingual newspaper corpus (125 million words), CHAVE (Santos and Rocha, 2005). The idea was to identify patterns typically used to introduce these verbs, and then use the patterns to discover more verbs. The verbs chosen were *dizer* ('say'[6]), *perguntar* ('ask'), *responder* ('answer'), *admitir* ('admit'), *contar* ('tell') and *continuar* ('proceed'). The first three verbs were selected because they are often referred to as typical reported speech verbs, while the other three were typical cases of describing other speech properties, as shown in the examples below, lightly adapted from COMPARA:

- *A Jean* ***contou*** *à Betty e a Betty* ***contou****-me.* ('Jean **told** Betty, and Betty **told** me')

- *Ele* ***contou*** *até três e tocou a campainha.* ('Ele **counted** to three and pressed the bell')

- *– As escolas são de boa qualidade –* ***continuou*** *ele.* (' "The schools are pretty good," he **said**')

- *Ela* ***continuou*** *calmamente o seu caminho.* ('She **proceeded** calmly on her way')

- *– Foi uma longa viagem –* ***admitiu*** *Bernardo.* ("'It was a long journey," Bernardo admitted')

- *A concorrência é muito grande e não são* ***admitidos*** *erros* ('The competition is severe and mistakes are not **admitted**')

[6]The translations into English are only rough aids for a reader who does not understand Portuguese. They obviously cannot illuminate the differences in meaning and style between the two languages.

From the analysis of the concordance lines of the six aforementioned verbs, we extracted the previously mentioned eight patterns, classified as direct, indirect and mixed quotation.

We the repeated the queries, but now using Floresta (Freitas et al., 2008), a monolingual corpus smaller than CHAVE (6.7 million words) and leaving the verb position blank in order to gather new verbs. We revised the new verb list, looking at each verb with a frequency greater than one. As a result, we produced a list with patterns and 285 reported speech verbs associated to it.

The association between a verb and its pattern is a crucial clue to get the correct sentences, since there are verbs that are used as reported speech verbs only in specific contexts. The verb *considerar* ('to consider'), for example, is a reported speech verb only in direct quotations, as shown in the first sentence below. In the second one, although the verb appears in an indirect quotation pattern, it does not introduce a quotation, but refers to the opinion of the subject.

- *"Não precisa chamar o garçom, sai mais barato e é bem mais prático",* ***considera*** *o advogado Beto Tonsig, 28.* ("No need to call the waiter, it is cheaper and way more practical", considers/says Beto Tonsig, a 28 year old lawyer.')
- *Dirceu* ***considerou*** *que a mudança prejudicava seu projeto porque o parlamentar continuava podendo renunciar.* ('Dirceu meant/judged that the change harmed his project because the member of parliament could stil renounce.')

Further work with the material suggested that a tripartite grouping was enough, as referred above, but the pattern-separated verbs can still be found in (Freitas, 2015).

7. Corpus results

We annotated a particular corpus available from AC/DC (Santos, 2014a), created in honour of the late Mariano Gago, and whose contents have a large number of texts produced in the occasion of his death, in April 2015. This corpus, presented in (Santos, 2015), was created precisely to try out and develop several language engineering tasks that were made possible by his support and funding of Linguateca, one of them being automatic opinion and quotation attribution.

But the idea and the results described in the present paper (namely, improved annotation rules and lexica) have also been applied to all corpora available through Gramateca (Santos, 2014b), although their annotation has not been revised (yet).

So, the Mariano Gago corpus was annotated in the usual way for semantic material: after a lexicon phase, we used the `corte-e-costura tool` (Santos and Mota, 2010) iteratively in order to refine the rules and produce a fully humanly revised material.

In this corpus (of around 300,000 words), we have identified 1,760 occurrences of speech verbs (157 different lemmas) that do not report what was said, plus 831 cases (73 different lemmas) where such reporting can be identified: 248 direct speech instances (40 different verbs), 270 indirect speech cases (corresponding to 44 different verbs) and 313 cases of mixed (37 different verbs).

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

Table 1: Speech in the JMG corpus

It is interesting to note that indirect speech is more frequent than direct speech in this corpus. Even more interesting is the high proportion of mixed quotation, a phenomenon so far ignored by most scholars, as (Freitas, 2015) stresses. Another relatively surprising finding was the amount of saying verbs in the first person: 65 in 836 reporting cases (7.8%), and 294 in 1836 non-reporting ones (33.9%).

7.1. Other quantitative results

How often does a sentence include reported speech? And how often does a text include reported speech? In the JMG corpus, out of the 661 texts we found reported speech in 268.

The genre (or kind of text) distribution is presented in Table 2 (note that "news" refers to news texts after his death, while "other texts", which are also news, were published before Mariano Gago's death).

Genre	Speech	Rep. speech	Total
news	690	446	142544
other texts	561	318	76908
interviews	221	31	30598
speeches	98	10	11888
hommage site	190	33	32536

Table 2: Speech in the JMG corpus per kind of text

7.2. Problems

Was the particular corpus well chosen to address and investigate reporting? Yes and no. On the one hand, it did include a lot of reports of what others had said. But on the other hand, there was a lot of reuse (and one might even say repeating) of the same utterances. So while the reporting numbers can be adequate to give an idea of what an ordinary reader is submitted to (with a fair share of repetition), it is obviously too repetitive in terms of source.

Also, since Mariano Gago was Portuguese, most of the news were written in Portugal and very few in Brazil, so the corpus was not at all representative of Portuguese as an international language. It was thought provoking, however, to realize that one of the few direct speech quotations without quotes came from a Brazilian source. This is something that has obviously to be better investigated.

Other corpora may give a more varied indication of how to report in Portuguese, so we present here the (unrevised) data from CHAVE (Santos and Rocha, 2005), a comparable

newspaper text corpus from Brazil and Portugal from 1994 and 1995.

Case	Number	Types	Number	Types
Speech verbs	68,865	562	1,434,177	1813
Reporting	16214	218	564,805	726

Table 3: Speech in Floresta and CHAVE

We present also data from Floresta Sintá(c)tica since it was used in the lexicon building phase, described above.
In CHAVE, we found (not revised) direct speech in 10.6% of the news, and indirect speech in as much as 31.6% of the news.

8. Concluding remarks

Our results show that the identification of reported speech in Portuguese is hard but manageable. The existence of speech-annotated corpora allows innovative applications, as well as increases the semantic and syntactic knowledge about Portuguese in a domain where this language is particularly rich.
The fact that all these data are publicly available through Linguateca's site[7] is also relevant, since then other people interested in quotation extraction in Portuguese can study the issues and contribute to improve the lexicons and/or the annotation.
Our intention is to develop, based on this work, the QUEMDISSE? system, a QE tool to identify opinion sources in Portuguese, which is also aware of mood and intention in reported speech.

9. Bibliographical References

Bergler, S., Doandes, M., Gerard, C., and Witte, R. (2004). Attributions. In *Exploring Attitude and Affect in Text: Theories and Applications, Technical Report SS-04-07*, pages 16–19.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Aarhus, Denmark.

Bruce, R. F. and Wiebe, J. (1999). Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5:187–205.

Jenny Brumme et al., editors. (2012). *The Translation of Fictive Dialogue*. Rodopi.

Caldas-Coulthard, C. R. (1996). A tradução e os problemas da representação da fala. In Malcolm Coulthard et al., editors, *Theoretical Issues and Practical Cases in Portuguese-English Translation*, pages 145–156. The Edwin Meilen Press.

Drury, B. and Almeida, J. (2012). The Minho Quotation Resource. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012*, pages 2280–2285. European Language Resources Association.

Drury, B., Dias, G., and Torgo, L. (2011). A Contextual Classification Strategy for Polarity Classification of Direct Quotations from Financial News. In *International Conference On Recent Advances in Natural Language Processing (RANLP 2011), Hissar, Bulgaria*, pages 434–440.

Fernandes, W. P. D., Motta, E., and Milidiú, R. L. (2011). Quotation Extraction for Portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 204–208.

Frankenberg-Garcia, A. and Santos, D. (2003). Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, et al., editors, *Corpora in Translation Education*, pages 71–87, Manchester. St. Jerome Publishing.

Freitas, C., Rocha, P., and Bick, E. (2008). Floresta Sintá(c)tica: Bigger, Thicker and Easier. In António Teixeira, et al., editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, volume Vol. 5190, pages 216–219. Springer Verlag.

Freitas, B. (2015). Discurso relatado: relatório parcial sobre a obtenção dos verbos do dizer. Technical report, PUC Rio, May.

Hofstader, D. R. (1997). *Le Ton beau de Marot: In praise of the Music of Language*. Basic Books.

Jarouche, M. M. (2013). *As mil e uma noites*. Globo Livros.

Jones, B. (1996). *What's The Point? A (Computational) Theory of Punctuation*. Ph.D. thesis, University of Edinburgh.

Loffredo, L., Grossman, D., Bitar, G., and Gonçalves, J. (2004). Verbos de Elocução - As Diferencas entre o Inglês e o Português. *CROP - Revista da Área de Língua e Literatura Inglesa e Norte-Americana*, 10:167–184.

Mambrini, F., Passarotti, M., and Sporleder, C. (2012). Annotation of Corpora for Research in the Humanities. Proceedings of the ACRH Workshop, Heidelberg, 5 Jan. 2012. *Journal for Language Technology and Computational Linguistics*, 26(2).

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of the 2103 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 989–999.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances In Natural Language Processing*, page 25–32.

Sagot, B., Danlos, L., and Stern, R. (2010). A lexicon of French quotation verbs for automatic quotation extraction. In Nicoletta Calzolari, et al., editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1437–1444. European Language Resources Association, 17-23 May.

Santos, D. and Mota, C. (2010). Experiments in human-

[7]Lexicon and rules are available from `http://www.linguateca.pt/Gramateca/VerbosDizer.html`, and the corpora are available from the AC/DC interface.

computer cooperation for the semantic annotation of Portuguese corpora. In Nicoletta Calzolari, et al., editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1437–1444. European Language Resources Association, 17-23 May.

Santos, D. and Rocha, P. (2005). The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Carol Peters, et al., editors, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491, pages 821–832, Berlin/Heidelberg. Springer.

Santos, D. (1998). Punctuation and multilinguality: Reflections from a language engineering perspective. In Jo Terje Ydstie et al., editors, *Working Papers in Applied Linguistics*, pages 138–160, Oslo.

Santos, D. (2014a). Corpora at Linguateca: Vision and Roads Taken. In Tony Berber Sardinha et al., editors, *Working with Portuguese Corpora*, pages 219–236. Bloomsbury.

Santos, D. (2014b). Gramateca: corpus-based grammar of Portuguese. In Jorge Baptista, et al., editors, *International Conference on Computational Processing of Portuguese (PROPOR'2014)*, pages 214–219. Springer, Outubro.

Santos, D. (2015). Um novo corpo e seus desafios. In *STIL 2015*, 5-6 november.

Sarmento, L. and Nunes, S. (2009). Automatic extraction of quotes and topics from news feeds. In *Proceedings of the 4th Doctoral Symposium on Informatics Engineering (DSIE09)*.

Smith, S. L., Mac Choueiti, M., and Pieper, K. (2014). Gender bias without borders: An investigation of female characters in popular films across 11 countries.

Snell-Hornby, M. (1983). *Verb-descriptivity in German and English: A contrastive study in semantic fields*. Carl Winter Universitätsverlag.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D., and Maybury, M. (2003). Recognizing and Organizing Opinions Expressed in the World Press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*, pages 24–26.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.