doi: http://dx.doi.org/10.5007/1984-8412.2017v14n3p2266

REPORTING VERBS IN PORTUGUESE: A CORPUS-BASED DESCRIPTIVE STUDY MOTIVATED BY COMPUTATIONAL LINGUISTICS

VERBOS DE ELOCUÇÃO EM PORTUGUÊS: UM ESTUDO DESCRITIVO COM BASE EM GRANDES CORPORA E MOTIVADO PELA LINGUÍSTICA COMPUTACIONAL

LOS VERBOS DECLARATIVOS EN PORTUGUÉS: UN ESTUDIO DESCRIPTIVO BASADO EN GRANDES CORPUS Y MOTIVADO POR LA LINGÜÍSTICA COMPUTACIONAL

Bianca Freitas Saburi Costa[*]
Cláudia Freitas[**]
Pontifícia Universidade Católica do Rio de Janeiro

ABSTRACT: In this article, we present the results as well as the procedures of a wide descriptive, corpus-based study on reporting verbs in Portuguese. The motivation for this research comes from a task carried out in Computational Linguistics – quotation extraction. In order to obtain reporting verbs used in Portuguese, our starting point was a list of translations of the verb "said" in literary contexts. Then, by using large monolingual corpora, we searched for lexical-grammatical patterns that characterize this verb class, in an attempt to broaden the list. Through this methodology, we identified 293 reporting verbs and distributed them amongst eight general patterns in which these verbs are typically found. The comparison between our research and other studies in Portuguese, English and French that investigate reporting verbs attests the success of the methodology that was developed.
KEYWORDS: Reporting verbs; *Verba dicendi*; Portuguese language description; Monolingual corpora.

RESUMO: Neste artigo, apresentamos os resultados e as etapas de um amplo estudo descritivo, com base em *corpora*, sobre os verbos de elocução em português. A motivação para o trabalho vem de uma tarefa da Linguística Computacional – a identificação de citação. Partimos, inicialmente, de traduções do verbo "said" em contextos literários para a depreensão de verbos de elocução. Em seguida, utilizando grandes *corpora* monolíngues, buscamos padrões léxico-gramaticais característicos dessa classe de verbos, a fim de ampliar a lista. Com a metodologia, foram identificados 293 verbos de elocução, distribuídos em oito padrões gerais típicos desse grupo de verbos. Os resultados são comparados com outros trabalhos em português, inglês e francês, que tematizam os verbos de elocução e atestam o sucesso da metodologia utilizada.
PALAVRAS-CHAVE: Verbos de elocução. Verbos *dicendi*. Descrição do português. *Corpora* monolíngues.

[*] Mestre em Estudos da Linguagem pela Pontifícia Universidade Católica do Rio de Janeiro. E-mail: bianca.saburi@gmail.com.

[**] Doutora em Estudos da Linguagem pela Pontifícia Universidade Católica do Rio de Janeiro. Professora assistente do Departamento de Letras da PUC-Rio. E-mail: claudiafreitas@puc-rio.br.

RESUMEN: En este artículo, presentamos los resultados y las etapas de un amplio estudio descriptivo, basado en corpus, sobre los verbos declarativos en portugués. La motivación para realizar el trabajo proviene de una de las tareas de la Lingüística Computacional: identificar las citas. Inicialmente, partimos de las traducciones del verbo "said" en contextos literarios para identificar los verbos declarativos. A continuación, mediante la utilización de grandes corpus monolingües, buscamos patrones léxico-gramaticales característicos de este tipo de verbos con el objetivo de ampliar la lista. Siguiendo esta metodología, se identificaron 293 verbos declarativos, distribuidos en ocho patrones generales típicos de este tipo de verbos. Los resultados se comparan con los de otros trabajos realizados en portugués, inglés y francés, que abordan el tema de los verbos declarativos y demuestran el éxito de la metodología utilizada.
PALABRAS CLAVE: Verbos declarativos. *Verba dicendi*. Descripción del portugués. Corpus monolingüe.

1 INTRODUCTION AND MOTIVATION

Human languages are characterized by, among other things, their polyphonic and dialogic trait – a great deal of our linguistic activities involve, to some extent, reporting what other people said. This reporting, which we call reported speech, is of great interest to Computational Linguistics/NLP (Natural Language Processing), more specifically to the quotation extraction task. The main goal of such task is to identify quotations within a text and to match it to its authors. The task focuses on identifying what is said and who says it. There is a preference for direct speech, since its formal marks make automatic detection easier.

Since the reported speech structure is relatively regular in Portuguese language, approaches that make use of rules tend to be successful. On the other hand, formal marks that indicate the presence of a quotation, such as quotation marks and dashes, are not exclusive of this type of speech, thus making it relevant to have a list of verbs that point to the existence of reported speech. The occurrence below, for instance, could be falsely detected as reported speech if punctuation were the only formal clue in the development of the task:

(a) O ataque do ombudsman ao vírus jornalístico apareceu na coluna «Chega de Ébola e Internet», de 28 de maio.
The ombdusman's attack on the journalistic virus came up in the 28th May column «Chega de Ébola e Internet».

Furthermore, not every reported speech presents the aforementioned formal clues, as these are characteristic in direct speech:

(b) «Já fui mais nervoso», disse.
«I have been more nervous», he said.
(c) Maria de Fátima disse que planeja sair do Rio.
Maria de Fátima said that she plans to move out of Rio.

Precisely for being so difficult to identify, indirect speech quotations are not always covered by automatic detection systems, even though such quotations correspond to nearly half of the reported speech occurrences, according to Pareti et al. (2013).

Moreover, even if the quotation extraction task focuses on the author and the content, instead of the verb, we know that the way people introduce reported speech – that is, the verb chosen to report – may also indicate stances. Therefore, the existence of a lexicon of verbs which may signalize the presence of reported speech, associated with the lexical-syntactic patterns in which they are used, could be a useful resource for NLP in the Portuguese language.

This article aims to describe the creation process of a reporting verbs lexicon (henceforth called RV) based on large corpora. Such verbs are also known as *verba dicendi* or "reported speech introducing verbs" (MOURA NEVES, 2000, p. 47).

A wide descriptive study of reporting verbs may also concern other fields, such as Portuguese teaching – when direct and indirect speech is being taught – and Pragmatics, especially regarding Austin (1962). Although there are a number of works on the topic, with either a more stylistic or pragmatic perspective, the academy could use more descriptive studies, particularly based on large

corpora. An exception would be a Brazilian grammar called *Gramática de usos do português*, by Maria Helena de Moura Neves (2000). Her work, which is corpus-based, covers reporting verbs and lists them according to their presence in direct and indirect speeches. Our aims in this paper are different – and the results of each work, thus, complement each other. Besides, the corpora we used compose the AC/DC project (COSTA et al., 2009)[1] and are available to the public, which allows other people to confront or continue what is being presented here. All the examples found in this article were taken from the project's corpora.

Several studies that describe verbs, particularly those concerning the English language, approach what they call "communication verbs", a group that encloses many more verbs than RV. Communication verbs correspond to any verb related to communication or speech, regardless of whether there is anything to be reported. As our focus is on reporting, such terminology and scope did not seem to be as interesting as the one we chose.

The remaining of the article is structured as follows: in section 2, we present some studies which have already approached RV, both in Portuguese and in other languages; in section 3, we describe in detail the collecting process for the reporting verbs; and, in section 4, we put forward some final comments.

2 FROM LANGUAGE STUDIES TO COMPUTATIONAL LINGUISTICS: A LITERARY OVERVIEW ON REPORTING VERBS

In Brazil, there is a considerable amount of studies that deal with reporting verbs. However, most of them take either a stylistic, a pragmatic or a normative approach, which does not mean they did not contribute to the study of this verb class in Portuguese. When we consider languages other than Portuguese, Beth Levin's work, *English verb classes and alternations* (1993), stands out.

Levin (1993) bases her work on the principle that the syntactic behavior of a verb is mostly determined by its meaning, especially when it comes to expressing and interpreting the verbs' arguments. Hence, the syntactic behavior of the verbs could be used to investigate aspects of the meaning behind this word class (LEVIN, 1993, p. 1). Levin presents 49 verb classes which are "syntactically relevant" and "semantically coherent" (LEVIN, p. 22). For each class, the author provides a list of verbs as an example, but emphasizes that those lists are not meant to be exhaustive.

Among the 49 proposed verb classes, we highlight a class called "verbs of communication". This group, according to Levin (1993), contains the verbs related to communication and transfer of ideas. These verbs are then included in nine semantic categories, which are related to speech, but not necessarily reported speech. At first, the "say verbs" subcategory would better suit our purposes, but, even so, the criteria that delimit each subcategory are not clear, and other subcategories seem to cover some of the verbs that we regard as RV. Some of the groups that could be considered as RV are the "complain verbs", such as *complain* and *object*, and the "advise verbs", such as *advise* and *warn*. In Freitas (2016), we present and discuss in detail the nine categories that Levin (1993) proposes. Levin states clearly that her analysis of the verbs of communication is quite brief and corresponds basically to the verbs whose properties do not involve complement clauses. The author chose to be brief due to the fact that the systematic study of complement clauses of verbs is not within the scope of her book (LEVIN, 1993, p. 202). Despite the limited scope in syntactic terms, Levin managed to gather 163 verbs of communication. Furthermore, the work developed by the author is an important reference for description as well as NLP studies, since her classification is the foundation of one of the greatest lexical resources used in automatic language processing, that is, VerbNet (KIPPER et al., 2006) and, by extension, its Brazilian version, VerbNet-BR (SCARTON et al., 2012).

Unlike Levin (1993), this study did not exclude complement clauses, which are extremely productive in a language, especially for reporting speech. On the other hand, we cover a group of verbs stricter than Levin's "verbs of communication".

[1] The characteristics of each corpus explored in our study will be presented throughout this article, more specifically in sections 3 and 4.

Regarding Portuguese language, we single out two references, Othon M. Garcia (2010) and Maria Helena Moura Neves (2000). The former consists of a manual which follows a normative tone through and through, whereas the latter corresponds to a descriptive grammar book.

Othon M. Garcia defines what he calls "*dicendi* verbs" as those "whose main function is to indicate the interlocutor who holds the turn to speak" (GARCIA, 2010, p. 149, our translation). In what he considers a "chaotic list", Garcia (2010) lists 50 verbs. In addition to that, he comments on the use of verbs which, originally, are not related to speech: "Even the most imaginative people come to employ verbs that have no connection whatsoever with the idea of speech; and this, from the point of view of syntax, could be considered inadmissible, since dicendi verbs should be, at least theoretically, transitive or admit transitivity." (GARCIA, 2010, p. 149) [2].

Then, Garcia (2010, p. 150) presents the *sentiendi* verbs, "which are not exactly 'say', but rather 'feel' verbs and that, by analogy, could be called *sentiendi* […]". These are verbs "that express a mental state, a psychological reaction of a character, emotions", among other feelings. The author explains that *sentiendi* verbs are "vicarious" for the *dicendi* verbs and characterize the psychic manifestations of the characters. Garcia states, however, that some intransitive verbs should precede the quote, when in direct speech:

(a) Mas João de Deus, vendo que Vasco não lhe dá atenção, **explode**:
Você pensa, seu Vasco, que estou disposto a aturar suas malcriações?
But João de Deus, noticing that Vasco pays no attention to him, ***explodes****:*
Do you think, Mr. Vasco, that I am willing to put up with your bad manners?
(VERÍSSIMO apud GARCIA, 2010, p. 155)

(b) Você pensa, seu Vasco, que estou disposto a aturar suas malcriações? – *explodiu* João de Deus.
Do you think, Mr. Vasco, that I am willing to put up with your bad manners? – exploded João de Deus.

Example (a), provided by Garcia (2010), would be the only possibility of use for the verb "explodir" (*explode*), originally intransitive, in direct speech. According to the author, example (b) would be unacceptable. For a verb like "explodir" to be admitted after the quote, it would be necessary for a *dicendi* verb to accompany it, and, even so, the order of the verbs would have to be altered, as the example below illustrates:

(c) – O coitadinho tem andado aborrecido! – disse ela **lamentando-se**.
– The poor thing has been feeling upset! – she said, ***lamenting****.*

Despite the limitations that Othon M. Garcia attributes to *sentiendi* verbs, we turned to CHAVE corpus (SANTOS; ROCHA, 2005) – a monolingual corpus which is part of the AC/DC project (COSTA et al., 2009) and is used in our research – in order to verify if those restrictions were applicable. As it turns out, we could find the frequent use of the verb "lamentar" (*lament*) after the quotation, as it can be attested in the following examples:

(d) «Pensei que conseguiria vencer e fiquei animado», **lamentou** Wilander.
«I thought I could win and I got carried away», Wilander ***lamented****.*

(e) «Qualquer dia eles me tiram», **lamenta**.
«They will pull me out any day now», he ***laments****.*

(f) «Até hoje não obtivemos resposta», *lamenta* Pádua.
«We had no answer until now», Pádua ***laments****.*

Corpus research enables us to observe the fact that not always norms in grammar books and writing manuals apply to real data. In a single corpus, 392 occurrences of the verb "lamentar" after the quotation were found. It is thus evident that real data make the

[2] "Chegam mesmo, os mais imaginativos, a empregar verbos que nenhuma relação têm com a ideia de elocução, o que, do ponto de vista da sintaxe, poderia ser considerado como inadmissível, pois os dicendi deveriam ser, teoricamente pelo menos, transitivos ou admitir transitividade." (GARCIA, 2010, p. 149).

limitations that Garcia had stipulated for the *sentiendi* verbs relative, if not invalid. The data also emphasize the potential use of corpus for this type of study.

Gramática de usos do português, by Maria Helena Moura Neves (2000), has more affinities with our study than any other. This grammar book was developed from a database of 70 million occurrences, which are stored in the Lexicographic Studies Center of São Paulo State University (Unesp) and are composed of "novels, technical texts, oratorios, news articles and plays". The selection of texts, according to the author, "ensures the diversity of genres and allows the approach of different speech situations, including interaction, and it is worth mentioning that spoken language is represented in the simulation of speech found in theatrical plays" (MOURA NEVES, 2000, p. 14, our translation).

However, unlike the study we are reporting here, the material consulted by Moura Neves (2000) is not available to the public, which makes it difficult to reproduce the research conducted for the grammar book. Within the group of "verbs that require complement clauses", Moura Neves presents the reporting verbs.

Moura Neves uses the terminology "say verbs" for the "actual reporting verbs", that is, "action verbs whose direct complement is the content of what is said" (MOURA NEVES, 2000, p. 48). Aside from the say verbs, the author proposes a category for "verbs that introduce speech, but not necessarily indicate speech acts". This group also integrates the RV. In this group of "say verbs", also called *dicendi* in the grammar book, the author includes the verbs "falar" and "dizer" (*speak* and *say*, respectively), which would be "neutral" verbs, together with other verbs "whose meaning brings, in addition to the basics, information on the way speech is performed". The verbs "gritar" (*shout*), "berrar" (*yell*) and "sussurrar" (*whisper*) would serve as examples. Verbs that add "notions about the speech chronology", such as "retrucar" (*reply*) and "repetir" (*repeat*), could also be considered say verbs (MOURA NEVES, 2000, p. 48).

Moura Neves highlights that, among the say verbs, there are also those in which the way we speak is lexicalized in the verb itself (MOURA NEVES, 2000, p. 48). The author cites as examples the verbs "queixar-se" (*complain*), "comentar" (*comment*) and "responder" (*answer*).

Once the RV are enclosed in the group of verbs that admit complement clauses, Moura Neves provides some tables in which she analyzes (1) the types of complement accepted by the RV as well as (2) the types of complement in which reported speech is presented. We counted 103 verbs compiled by the author as RV.

In the context of NLP, we call attention to a study by Sagot et al. (2010), in which they present the process for creating a quotation verbs lexicon. Their main goal was to develop an automatic quotation extraction tool.

From a preexistent list of 110 verbs, which were compiled in a previous study, Sagot et al. (2010) analyzed a corpus of 5,000 newspaper articles, looking for structures that would indicate the presence of direct, indirect and mixed quotations. After filtering the results manually, the authors narrowed it down to 836 quotation configurations that were associated with quotation verbs. Once strategies for obtaining verbs in such structures were developed, the authors reached 232 verbs.

The path we followed is, in a way, similar to the one treaded by Sagot et al. (2010): based in corpora, we carried out a semiautomatic survey of the RP in Portuguese. We assessed the structures in which such verbs were usually found and, from there, we obtained hundreds of verbs used in those configurations. The following section covers this process.

3 TRANSLATION AS A STARTING POINT: "SAID"

To begin the compilation of RV, the first step was to conduct a survey on the verbs that we use to introduce speech in Portuguese. In order to do that, we chose a material in which reported speech is abundant: literary texts. Therefore, we obtained our initial data

by consulting COMPARA (FRANKENBERG-GARCIA; SANTOS, 2002), a parallel, bilingual, English-Portuguese corpus, composed of literary texts and its corresponding translations into English or Portuguese[3].

COMPARA boasts morphosyntactic annotation and is available online. We chose COMPARA because of two different reasons. On the one hand, we know that the identification of a given verb as a reporting verb is not always obvious – "imaginar" (*imagine*) or "interromper" (*interrupt*), for instance. Thus, our choice for the initial exploration of COMPARA, using the verb form in English *said*, aimed at ensuring that the selected verbs were appropriate: when the original text uses *said*, there is no doubt that the correspondent translation to Portuguese refers to reported speech:

SOURCE TEXT: «Don't,» I **said**, in a muffled voice.
TARGET TEXT: – *Não!* – **interrompi** *numa voz abafada.*
– *No!* – *I* ***interrupted****, in a muffled voice.*

On the other hand, COMPARA is composed of edited literary works, aligned with its corresponding translations. The fact that those translations were edited by professionals strengthens our trust on the quality of the translations[4] – and on the validity of the verbs chosen. In COMPARA, we searched for translations for the verb form *said* (in English). We used the search expression [word="said"][5], limiting the results to English to Portuguese translations only. The verb form *said*, instead of *say*, was chosen due to the fact that the verb mostly introduces speech in its past form; such predominance is confirmed by Biber et al. (1999, p. 374). The search for translations of *said* provided us with an extensive material as a starting point. It was, then, unnecessary to resort to other forms of the verb *say* or to other equivalent verbs.

When searching in COMPARA, the results are limited to a sample of 1,000 random occurrences, due to copyright issues. Although the corpus had located a total of 3,560 occurrences of *said*, we decided to concentrate our study on one random sample, analyzing the 1,000 provided results. Figure 1 shows COMPARA's interface when searching for "said".

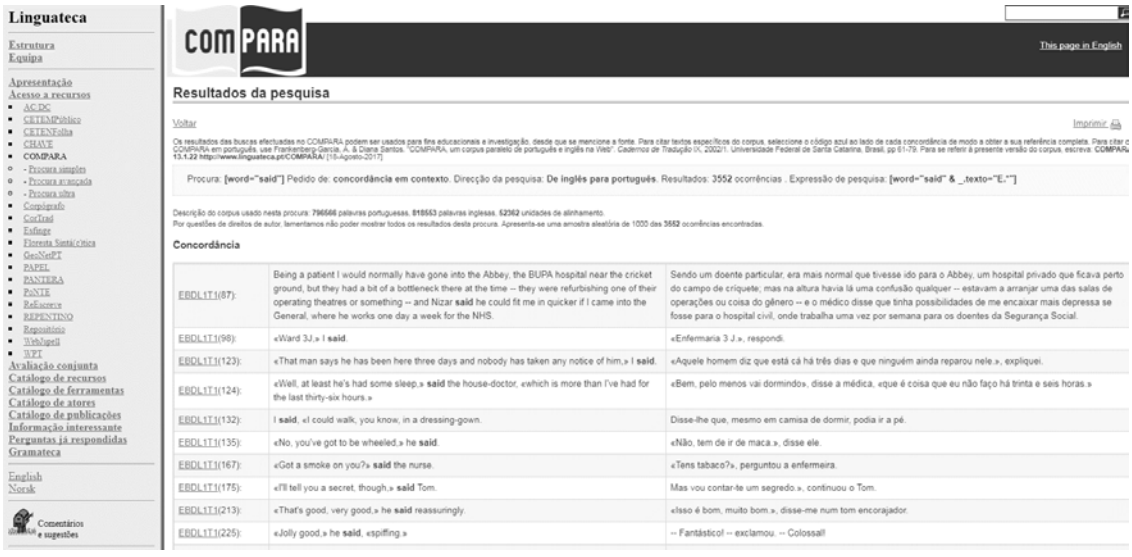


Figure 1: Search interface of COMPARA.

[3] It is worth mentioning that COMPARA consists of originals as well as translations in Portuguese from Portugal, Angola, Mozambique and Brazil. However, given the scope of our study, we did not consider it relevant to exclude any of the language's varieties.

[4] It should be taken into account the fact that COMPARA consists of a compilation of literary works so eminent that were worthy of being translated into other languages. All translations were developed by publishing houses (unlike some "free" translations from the internet, for example), which means that the translations we found in the corpus were validated by linguistic professionals. This gives us more confidence to accept the verbs that were chosen as translations for said as reporting verbs. Henry James, Lewis Carroll and Oscar Wilde are some of the authors whose translation into Portuguese constitute COMPARA.

[5] The expression indicates that we only wanted the word "said" in this exact form, discarding any possible variations.

As we analyzed the data, we discarded 42 occurrences, because they did not correspond to reported speech, as it may be noticed in the following example:

> SOURCE TEXT: «What they do, what Polly *said*, oh please let's Ange!»
> TARGET TEXT: – O que eles fazem, o que *disse* a Polly, por favor, Ange!

Table 1 lists the other 958 results of *said*, ordered by the number of occurrences of each verb. In 53 cases, the verb was omitted in the translation.

VERB	No. OF OCCUR.	VERB	No. OF OCCUR.	VERB	No. OF OCCUR.
dizer	561	comunicar	3	anuir	1
responder	88	concluir	3	anunciar	1
perguntar	60	confessar	3	balbuciar	1
omissão	53	confirmar	3	berrar	1
comentar	20	falar	3	brindar	1
explicar	17	informar	3	censurar	1
acrescentar	9	observar	3	concordar	1
afirmar	9	retrucar	3	desabafar	1
contrapor	9	ripostar	3	escrever	1
exclamar	9	agradecer	2	espantar-se	1
retorquir	9	argumentar	2	indagar	1
continuar	8	contar	2	justificar-se	1
declarar	7	cumprimentar	2	lembrar	1
insistir	6	esclarecer	2	prometer	1
interromper	6	propor	2	saudar	1
prosseguir	6	queixar-se	2	sondar	1
repetir	5	querer saber	2	sublinhar	1
pedir	4	admirar-se	1	sugerir	1
replicar	4	admitir	1	tornar	1
assegurar	3	advertir	1	**TOTAL**	**958**

Table 1: Verbs used as translations for *said* in COMPARA, ordered by frequency.
Source: Freitas (2016)

Although the number of occurrences of the verb "dizer" (*say*) is significantly higher than the other verbs' occurrences, reaching a total of 58.56% of translations, around 40% of the cases translators chose not to use the direct equivalent of *say*, which would be *dizer*. Instead, they used 58 different verbs, ranging from other frequent reporting verbs, such as "perguntar" and "responder" (*ask* and *answer*), to verbs not so commonly used to report speech, such as "sondar" (*investigate*), "agradecer" (*thank*), "brincar" (*tease*), etc.

Verbs such as these are rarely cited as reporting verbs, and therefore are not included in lists such as the ones elaborated by Garcia (2010) and Moura Neves (2000).

The main aim of the next step was to broaden the list of 58 verbs, by means of a semiautomatic strategy. In order to achieve that aim, we explored, in separate occasions, three large monolingual corpora from the AC/DC project (COSTA et al., 2009): CHAVE (SANTOS; ROCHA, 2005), composed of journalistic texts (98 million words); OBras, of literary texts in the public domain (1.2 million words), and Floresta (FREITAS et al., 2008), mostly journalistic (6 million words). The material is publicly available at the AC/DC project webpage[6].

4 OBTAINING USAGE PATTERNS

From the 58 initial verbs, we chose 6 to work as "seeds", in order to help us identify lexical-grammatical patterns typically used to introduce reported speech. Behind the identification of usage patterns of those verbs, there is the idea that the patterns themselves could make it easier to pinpoint other RV in the corpus. We elected the following verbs: "dizer" (*say*), "perguntar" (*ask*), "responder" (*answer*), "admitir" (*admit*), "contar" (*tell*) e "continuar" (*continue*).

The first three verbs were selected for being frequently referred to as RVs; indeed, they were the most frequent ones in the translations (see Table 1). Besides, these verbs are normally associated with *dicendi* verbs in grammar books and writing manuals (such as Othon M. Garcia's work) and, for this reason, could be considered prototypical. On the other hand, the verbs "admitir", "contar" e "continuar" were picked due to the fact that they are often associated with uses which are not always related to speech[7], such as:

> *O evento* **contou** *com a presença do arcebispo de Fortaleza, d. Aloísio Loscheider.*
> *The event* ***boasted*** *on the presence of the archbishop of Fortaleza, d. Aloísio Loscheider*[8].
> *Às 13h, a reportagem da Folha* **contou** *83 pessoas na manifestação.*
> *At 1 p.m., the Folha reporting crew* ***counted*** *83 people in the demonstration.*

Despite being the largest revised parallel corpus of Portuguese and English in the world, COMPARA only compiles 723,807 words and is limited to fictional texts, mostly novels. Therefore, in order to observe the seed verbs in context and consequently obtain their lexical-syntactical patterns, we turned to a more extensive corpus: CHAVE. Additionally, the fact that the texts in both corpora are from different genres also contributed to our choice, since the wide range of texts could provide us with a broader notion of usage of this verb class. The high frequency of RVs in CHAVE, on the one hand, confirms that we have chosen the appropriate material, but, on the other hand, emphasizes the unviability to analyze all the occurrences individually (over 350,000). This scenario forced us to develop strategies to deal with the large amount of information.

With that in mind, first we assessed the three verbs that typically indicate speech: "dizer", "perguntar" and "responder". We went over hundreds of occurrences and examined in which cases these verbs were introducing speech. From the analysis of the context in which these verbs were used, we managed to obtain eight frequent usage patterns. Then, we checked if the remaining verbs – "admitir", "contar" and "continuar" – were also used in these patterns.

The most common way of describing reported speech is to do so in accordance with the type of quotation involved: direct, indirect or mixed quotations. The type of quotation is of higher importance due to its role in highlighting what is being reported. In our case, we were interested in the patterns (rather than specifically in the type of quotations) on account of its relevance for searching in

[6] The AC/DC Project stands for "Acesso a Corpora/Disponibilização de Corpora" ("acess and availability of corpora") , and it is part of Linguateca.

[7] The verbs "admitir", "contar" and "continuar" may admit different meanings in Portuguese, which do not correspond to the verbs admit, tell and continue in English. In the case of a translation from Portuguese to English, the translator would have to opt for another verb in order to keep the intended meaning.

[8] In Portuguese, the verb "contar" may indicate a positive addition or contribution to a situation, such as the presence of a famous person in an event. Since the verb count does not have the same use in English, we opted for a substitute that makes the sentence understandable in English.

corpora: the patterns – or formalizations – correspond to different materializations of the speech structures. One of the consequences of emphasizing the types of quotation is the underlying idea that there is a direct connection between a given pattern and a given type of quotation. For instance, the pattern RV + "que" (*that*) inevitably would be associated with indirect speech. The observation of the contexts, however, showed a single pattern may be found in different types of quotation. Table 2 displays the distribution of patterns by type of quotation, and it is interesting to notice how the same pattern may be related to several ways of quoting.

PATTERN	DIRECT	INDIRECT	MIXED	EXAMPLES
1	x			**D:** «O aumento não vale para as diárias, só mensalistas », **afirmou** Paulo Octávio.
2	x			Até que uma amiga minha passou por ele e **disse**: «Oi, Fábio».
3	x	x	x	**D:** «Na França », **diz** ela, « a história é uma atividade muito prestigiada, e portanto muito masculina». **I:** Antes de tentar um acordo, ele passou em outra concessionária (não informou o nome), que, **disse**, teria lhe oferecido arcar com parte do II. **M:** A contribuição de Robin Williams a «Aladdin », **salienta** ele, « nenhum diretor, escritor ou animador poderia suprir».
4		x	x	**I:** Cauteloso, ele **disse** que não receberá empresários e empreiteiras. **M:** Ambas acabam por **reconhecer** que « roubos, sempre os houve», mas **acrescentam** que «agora é muito pior».
5	x	x	x	**D:** Quando os jornalistas perguntaram se se tratava da reeleição dele ou de Fidel, FHC respondeu, rindo: «Para ele, não é reeleição, ele quer a eternidade». **I:** O banco respondeu, informando-os que o depósito em causa tinha sido liquidado em 30 de Agosto de 1985, e que havia pago um cheque ao portador nesse montante, assinado pelo João. **M:** Vital Moreira **respondeu**, afirmando manter «tudo o que disse» e que «a comunicação social escreveu, sem que tenha havido desmentidos».
6		x	x	**I:** Ontem, Farouk Kaddoumi, dirigente do departamento político da OLP, se **disse** favorável a realização de uma nova conferência multilateral de paz, conforme proposta da Rússia. **M:** Ontem, na reapresentação na Vila Belmiro, ele se **disse** abatido, «principalmente por ter sido contra o São Paulo, de quem eu queria ganhar».
7	x	x	x	**D:** Quando os jornalistas perguntaram se se tratava da reeleição dele ou de Fidel, FHC respondeu, rindo: «Para ele, não é reeleição, ele quer a eternidade». **I:** Charles respondeu dizendo que não podia se envolver nessa questão. **M:** A acusação partiu de Miltiadis Evert, o líder da Nova Democracia, e o porta-voz do governo socialista, Telmahos Hyritis, **respondeu** acusando-o de «críticas irresponsáveis».
8	x	x	x	**D:** Ou ainda, como **escreveu** o também poeta Joseph Brodsky: «A verdadeira biografia de um poeta é quase idêntica à dos pássaros. **I:** O presidente Itamar Franco deve vetar o projeto, segundo **informou** o ministro da Fazenda, Fernando Henrique Cardoso. **M:** Se há «graves impropriedades» no relatório, como **afirma** o Planalto, elas devem decerto ser identificadas e contestadas.

Table 2: Distribution of patterns by type of quotation.
Source: Freitas (2016)

The eight patterns will now be presented in detail, taking into account each of the six analyzed verbs. To make it more reader-friendly, we decided to insert our analyses throughout the description of the patterns, instead of at the end.

Description of patterns

(1) Complete quotation + RV:

(a) *«Ninguém da minha família participa de sequestro», **disse** Silva.*
*«No one in my family takes part in kidnapping», Silva **said**.*

With direct quotations, it is simple to identify the arguments of the quotation in this pattern. The typical punctuation of this pattern would be one of the facilitating factors to pinpoint the arguments. However, it is worth mentioning that punctuation is not always present:

(b) *Os carros estavam saindo de traseira, **explicou** Keith Wiggins, o chefe da equipe.*
*The cars were leaving in reverse, Keith Wiggins, the head of the team, **explained**.*

Example (b) constitutes a case in which speech is reported directly, but without any punctuation typically associated with dialogues (quotation marks or dashes, in Portuguese). In regard to punctuation in Portuguese, quotation marks usually appear in journalistic texts, whereas in other genres, such as literary works, quotation marks are often replaced by dashes, according to publishing houses' guidelines. In the case of quotations with dashes, we usually omit the comma, as recommended by writing manuals. Garcia (2010), for example, states that "dashes render any other punctuation marks dispensable, except for question marks, exclamation points and ellipses". The following example, which is one of the 1,000 samples retrieved from COMPARA, illustrates this guideline:

(a) *– E como podemos fazer uma nova série sem Debbie? – eu **disse**.*
*– And how can we have a new series without Debbie? – I **said**.*

This pattern was the first one to be identified in the analysis of occurrences of "dizer", but it is also used with all the other verbs, which indicates its productivity. From this pattern, we assumed we would have higher chances of finding other RVs. This hypothesis will be confirmed later on.

(2) RV + complete quotation:

(a) *Zico, em conversa exclusiva com a Folha, **disse**: «Estou muito contente com essa homenagem».*
*Zico, in an exclusive interview with Folha, **said**: «I am really happy with this tribute».*
(b) *Uma hora após Gaviria ter evitado responder de forma categórica à pergunta, a Folha **perguntou** a Fujimori: «Houve transparência nas eleições peruanas?»*
*An hour after Gaviria categorically avoided answering the question, Folha **asked** Fujimori: «Was there transparency in the Peruvian elections?»*

This pattern belongs to direct speech, as well as Pattern (1) does, and it is extremely usual in Portuguese. The punctuation commonly used in this pattern (semicolons, dashes and quotation marks) may ease the identification of the arguments. In this pattern, the quotation marks can also be replaced by dashes, as illustrated by example (c):

(a) *Ao ouvir isto, Sofia virou-se para sua mãe, e empunhando a escova de lavar pratos, **disse**:*
– Saiba que ele é perfeitamente bom da cabeça.
*As she heard this, Sofia turned to her mother and, holding up the dish brush, **said**:*
– For your information, he is perfectly sane.

Similarly, Pattern (2) was initially found in the occurrences of "dizer" and, then, was widely found with the other verbs.

(3) RV inserted in the quotation

(a) *«Na Califórnia – **diz** ela – é cada vez maior o número de pessoas que come pouca carne ou que se tornam vegetarianas».*
*«In California – she **said** – the number of people who eat little meat or become vegetarian is increasingly higher».*
(b) *Como, **perguntará** o leitor, o execrando Tratado de Maastricht prevê uma coisa dessas?*

*How, the reader **will ask**, can the abominable Maastricht Treaty stipulate such a thing?*

*(c) «Vossa Santidade», **respondi**, «não só eu estou contente, todos nós estamos muito contentes».*

*«Your Holiness», I **answered**, «I am not the only one who is glad, we are all glad».*

*(d) «Sem dúvida alguma, ele é o melhor boxeador do mundo!», **admite**, «mas enquanto ele continuar dizendo que é o melhor lutador, estou pronto a desafiá-lo».*

*«Without a doubt, he is the best boxer in the world!», he **admits**, «but for as long as he keeps saying he is the best fighter, I am ready to challenge him».*

*(e) O helicóptero do líder do ANC foi apedrejado por manifestantes do Inkhata e a própria residência do rei foi atacada, **contaram** os seus conselheiros, obrigando o soberano a procurar refúgio numa quinta nos arredores.*

*The helicopter of the ANC leader was stoned by Inkhata protesters and even the king's residence was attacked, his advisers **told**[9], forcing the monarch to seek refuge in a farm nearby.*

*(f) Em 93, **continua**, a produção mundial foi de 18,7 milhões de toneladas e o consumo mundial, de 19,4 milhões.*

*In 93, he **continues**, the worldwide production corresponded to 18.7 million tons, and the worldwide consumption, to 19.4 million.*

Pattern (3) admits the three types of quotation and can be found both in journalistic and literary texts. In this pattern, it is not obligatory to make all arguments explicit, more specifically the agent[10]. Furthermore, it is quite common for the agent of the action to be in a previous sentence.

(4) RV + quotation introduced by a subordinate conjunction

(a) *Sobre a lateral esquerda, Parreira **disse** que quem deve jogar amanhã é Leonardo.*

*As for the left-back position, Parreira **said** that Leonardo is the one who should play tomorrow.*

(b) *Me **perguntaram** se esse era o momento oportuno.*

*They **asked** me if this was the proper moment.*

(c) *Genro, em tom duro também, **respondeu** que aquela não era uma visão de um dirigente de expressão nacional como Dirceu.*

*Genro, also in a harsh tone, **answered** that it was not a vision of a leader of nation-wide significance like Dirceu.*

(d) *O teólogo católico **admite**, no entanto, que o tema das mulheres é um problema real, ao qual a Igreja Católica também tem que dar resposta.*

*The catholic theologian **admits**, however, that the subject of women is a real issue and must also be addressed by the Catholic Church.*

(e) *FHC **contou** aos parlamentares que coube ao próprio Arida a indicação de Loyola.*

*FHC **told** the congressmen that it was up to Arida himself to appoint Loyola.*

(f) *Mas ante um persistente e interrogativo olhar do administrador, como que a intimá-lo a falar, ele **continuou** que não lhe tinha emprestado as pedras por uma questão de confiança, porque neste aspecto até poderia dizer que não confiava nem nos seus próprios dentes porque de quando em vez lhe mordiam a língua...*

*But under a persistent and questioning gaze from the administrator, as if urging him to speak, he **continued** that he had not lent the rocks for a matter of trust, because in this aspect he could even say that he didn't trust his own teeth because once in a while they would bite his tongue…*

Pattern (4) is limited in terms of syntactic dislocation. The author of the quote occupies the conventional subject position, preceding the subordinate clause. In occurrence (a), for instance, we have a quotation in its indirect form ("quem deve jogar amanhã é Leonardo"), the action ("disse") and the author of the quotation ("Parreira"). We must highlight, though, that the presence of the author is not mandatory; in that case, it is possible to find the referent in a previous sentence.

There is a variation of Pattern (4), adapted for direct speech (by including a colon):

[9] The verb "contar" in Portuguese does not require the same complements as its equivalent in English does. In Portuguese, the interlocutor of the message may frequently be omitted (sb tells sth).

[10] In Portuguese, a verb does not need to be accompanied by the agent of the action.

(g) ***Acrescentou*** *ainda que: «A prazo, o financiamento do Ensino Superior deverá crescer, a par do aumento da sua frequência e dos resultados alcançados.»*

*He also **added** that: «In the long term, higher education financing should increase, considering its higher frequency and the obtained results.»*

Pattern (4) is very recurrent in Portuguese, on account of the wide range of verbs that can be used in such structure. However, it is precisely this variety that makes it difficult to identify RVs, since this structure is not necessarily associated with reported speech:

(h) *Pessoalmente, **considero** que Mariano Gago foi, de longe, o melhor ministro da era «democrática».*

*Personally I **consider** that Mariano Gago was, by far, the best minister in the «democratic» era.*

In occurrence (h), we would have, at first, an indication of mixed quotation. However, the verb "considerar", as the verbs "admitir", "contar" and "continuar", holds meanings unrelated to reported speech. In example (h), the fact that the verb is in the present and in its first person singular form is already an impediment for reported speech, since we do not consider it possible to report something that has not been said yet. Precisely because of the verb tense and the first-person form, we believe that, in (h), the verb expresses the meaning of "to decide, after reflection, about (a certain thing); to regard" (HOUAISS, 2016, our translation). Besides, acting as a formal clue for identifying dialogues is not the only function of quotation marks. Many times, we use them merely as a resource for emphasis or highlight, and we believe the case above matches this use.

The noun clause as direct object that composes Pattern (4) is also possible in its incomplete form, with "yes" and "no" answers.

(i) *Nem o governo, nos momentos de maior sinceridade, seria capaz de **responder** que sim.*

*Not even the government, in its most sincere moments, would be able to **answer** yes.*

The verb "perguntar" (*ask*) is used, in this pattern, differently from the other verbs, but similar to other question verbs. As it corresponds to an indirect question, we often find the conjunction "if" instead of "that", as it can be perceived in occurrence (b), not to mention other conjunctions and wh-words:

(a) ***Perguntei*** *por que eles queriam que justo a Executive Outcomes fizesse isso.*

*I **asked** why they wanted specifically Executive Outcomes to do this.*

(b) *Tivemos o desabafo do presidente da República quando um jornalista **perguntou** o que ele faria se dependesse do salário mínimo.*

*We heard the President of the Republic vent when a journalist **asked** what he would do if he depended on minimum wage.*

(5) VE + citação em oração reduzida de infinitivo[11]

(a) *Um dia antes de disputar a eleição, Sanguinetti **disse** achar que seu país está chegando ao Mercosul em desvantagem com os demais.*

*One day before running for the elections, Sanguinetti **said** to believe his country is arriving at Mercosur at a disadvantage when compared to the others.*

(b) *Sem muita convicção, um deles **respondeu** saber de tal urgência.*

*Without much conviction, one of them **answered** to know about such urgency.*

(c) *O ministro russo de Cultura, Evguêni Sidorov, **admitiu** ter tocado no tesouro.*

*The Russian Minister of Culture, Evguêni Sidorov, **admitted** to having touched the treasure.*

[11] Some of the examples were translated literally in order to help the reader understand the structure of the sentences in Portuguese, even though they do not always sound correct in English. Examples (e) and (f) were not translated because the subtle difference between them could not be reproduced in English.

(d) ***Contou*** *ter feito a campanha de lançamento na Argentina da marca Cica.*
He ***told*** *to have done the launching campaign for Cica company in Argentina.*

From a search in CHAVE, we obtained 870 occurrences of RVs with infinitive complements in Portuguese. We found 73 verbs in the complement position.

As for the arguments, Pattern (5) presents the same structure and the same process of identification for the quotation, the author and the reporting verb that we pointed out in Pattern (4). The inflexibility of dislocation of the arguments is also constant in this pattern, which facilitates the identification process.

We only found 21 occurrences of the verb "continuar" followed by infinitive[12]. In all of them, "continuar" nevertheless has the meaning of *persist*. It is highly unusual for the complement of the verb "continuar", when in infinitive form, to appear without the preposition "a", as the occurrence (e) shows:

(e) *Tirar um curso superior em Portugal* ***continua*** *depender das condições económicas.*
(f) *O grupo dos times pequenos* ***continua*** *a existir.*

(6) RV + object pronoun + adjective or participle[13]

(a) O governo do México se ***disse*** *disposto a negociar uma trégua.*

This pattern is relatively frequent for the verb "dizer". However, it is complex to determine if the verb, in this pattern, is being used as a RV, given that it intends to attribute a characteristic to the author of the alleged quotation. In example (a), "o governo do México" (author) considered that it was "disposto a negociar uma trégua" (*willing to negotiate a truce*), that is, the government of Mexico attributed the characteristic of *willing* to itself.

(b) Ele se ***disse*** *traído por Viljoen.*

In example (b), which differs from example (a) due to its complement being a participle instead of an adjective[14], we observe the same limitation as to the intention of the speaker. Is "traído por Viljoen" what was said? Do we have a quotation indeed?
The fact that very few verbs are placed in this structure – only five verbs, as we show in Table 3 – already minimizes the chances of Pattern (6) being a good indication of a RV. Among the six analyzed verbs, only "dizer" was seen in occurrences that fit the pattern and, in a later search, we only obtained five verbs in Pattern (6):

LEMMA	OCCURRENCES
dizer	26
declarar	6
revelar	3
reconhecer	2
defender	1

Table 3: Verbs found in Pattern (6).
Fonte: Freitas (2016)

[12] For this search, we used the expression [pos="V.*" & lema="continuar"] @[pos="V" & temcagr="INF"].
[13] The examples in this subsection will not be translated, as their structure is too distant from a correspondence in English.
[14] The classification of participle forms in Portuguese is quite controversial, as Trugo (2016) discusses in her work.

(7) Quotation + RV + gerund phrase

We did not find the verb "dizer" introducing speech in this pattern. The occurrences we retrieved confirm the theory that the gerund phrase simply adds new information to the sentence, instead of just reiterating the reporting verb that precedes it:

(a) *«É, na época eu era aluno de pós-graduação», **diz** brincando.*
*«Yes, at the time I was a grad student», he **says**, joking.*

Similarly, the verb "perguntar" was not found in this pattern. The occurrences of "perguntar" in such structure refer to other usages of gerund phrases, such as indicating simultaneous or consecutive actions:

(b) *"Este aqui?", **perguntou**, apontando para um homem moreno, cheio de brilhantina no cabelo e bigodinho à Errol Flynn.*
*"This one?", he **asked**, pointing to a dark-haired man, with his hair full of grease and with an Errol Flynn moustache.*
(c) *"Quem recebe é corrupto", **perguntou**, provocando risos na plateia.*
*"The person who receives it is corrupt", he **asked**, making the audience laugh.*

The verbs "responder" and "admitir" are also questionable cases. For example, would the gerund phrase be a characterization of "responder"? Or would it be additional to the action of answering? The occurrences that we found for both verbs did not seem to validate Pattern (7) as a strong candidate for usage pattern of RVs.

(a) *A hierarquia católica costuma **responder** dizendo que a igreja não é uma democracia.*
*The catholic hierarchy usually **answers**, saying that the church is not a democracy.*
(b) *Mais modesto é, sem dúvida, o balanço em relação à evolução conseguida nas taxas de juro: «tem vindo a verificar-se uma evolução no sentido favorável, embora permaneçam no mercado de crédito taxas de juro reais elevadas», **admite**, sublinhando no texto entregue aos jornalistas que «a descida das taxas de juro não tem sido uniforme nos vários segmentos (dívida pública, crédito a grandes empresas, «PME Prestígio» e outras PME e particulares)».*
*The balance sheet in relation to developments in interest rates is undoubtedly more modest: "there has been a favorable trend, although real high interest rates remain on the credit market," he **admits**, stressing in the text handed to journalists «the decline in interest rates has not been uniform across sectors (public debt, credit to large companies, «SME Prestígio» and other SMEs and individuals)».*

In the few occurrences of gerund phrases after the verb "contar", the verb in the phrase does not help to introduce speech; instead, it attributes some type of characteristic or specificity to the act of telling, as it may be perceived in the example below:

(a) *Ele relata as histórias que lhe **contaram**, usando palavras que mostram que ele sabia aquilo a que se referia ao escrever.*
*He narrates the stories that people **told** him, using words which show that he knew what he was referring to when he was writing.*

We could not find any occurrence of "continuar" as a RV in accordance with Pattern (7). In all the occurrences in which it is followed by a gerund form, "continuar" has the meaning of *proceed*:

(a) *Disse que reconhecia nele a virtude de **continuar** dizendo o que fez.*
*He said that he recognized in him the virtue to **continue** saying what he did.*

(8) RV in a "conformational" adverbial clause[15] + quotation

(a) *E como **diz** o locutor Fiori Gigliotti em suas transmissões de futebol, «o tempo passa, torcida brasileira».*
*And as the radio announcer Fiori Gigliotti **says** in his soccer broadcasts, «time goes by, Brazilian supporters».*

[15] In Portuguese grammar, this category expresses an idea of accordance with something, which could be a rule or, in the case of the examples shown in this paper, what other people said.

(b) *Então, como **perguntou** um professor presente no Curso de Verão, «não será necessário mudar o essencial da relação pedagógica, da filosofia do ensino e da organização da escola?»*

*Then, as a teacher attending the Summer Course **asked**, «won't it be necessary to change what is essential in pedagogical relationship, teaching philosophy and school organization?"*

(c) *«Boa pergunta!», como **responderia** ele.*

*«Good question!», as he **would answer**.*

(d) *Coca-Cola, eles provaram pela primeira vez no navio Merida, como **contou** à Folha o frentista Hugo Berto Isquierdo, 22.*

*Coca-cola, they tried it for the first time in the ship called Merida, as the gas station attendant Hugo Berto Isquierdo, 22, **told** Folha.*

(e) *«O reforço das acções necessários» para a prossecução desses objectivos de luta contra a evasão e a fraude é, como **continuou**, «uma das prioridades da DGCI» e para o qual tanto este departamento como a Inspecção-Geral de Finanças eram «organismos vitais».*

*«The reinforcement of the necessary actions» for pursuing these objectives to combat tax evasion and fraud is, as he **continued**, «one of DGCI's priorities» and for which both this department and the Inspectorate-General for Finance were «vital bodies».*

This pattern is characteristic in a very peculiar case of RVs when it comes to arguments. In the examples above, it may be noticed that there is not mere speech reporting, but the appropriation of something that was previously said. In this pattern, we have access to the "original" author of the quote, but not necessarily do we know who is the speaker at that moment, since it is the speaker who is reporting speech. In contrast with the seven preceding patterns, Pattern (8) counts on a slightly more rigid syntactic structure. More specifically, in this pattern, it is not possible to omit the subject of the RV. The explanation possibly lies on the peculiarity of the speaker appropriating a prior speech; since the "original owner" of the speech is not the subject of the clause that introduces it, revealing this subject becomes mandatory.

Here are some variations of this pattern, which is observed in all types of speech:

*(a) A América, como **dizia** Miles Davis, é um país maravilhoso onde você pode reunir mil vozes para gravar um uníssono.*
*America, as Miles Davis **used to say**, is a wonderful country where you can gather a thousand voices to record in unison.*

*(b) A resposta, meu amigo, é sussurrada pelos ventos, como **diria** a canção.*
*The answer, my friend, is whispered to the winds, as the song **would say**.*

*(c) E como **disse** o pagodeiro na MTV: tudo na vida é passageiro, menos o cobrador e o motorista.*
*And as the pagode musician **said** on MTV: everything in life passes, except the bus at night.*

*(d) Até chegar a uma «produtora estruturada», como **diz** Sandra, foram «meses de angústia».*
*Until we could find a «well-structured producer», as Sandra **says**, it has been «months of distress».*

The verbs "perguntar", "responder" and "continuar", according to our searches, are not frequently used in Pattern (8); each one of them had three, two and one occurrences, respectively. Although the numbers are not very high, the possibility of usage of the six verbs in the pattern makes it a viable candidate for helping us select reporting verbs.

Analyzing the productivity of the patterns

Our analysis from the previous subsection is schematized in Table 4, that displays which patters were admitted by each verb:

PATTERN / VERB	1	2	3	4	5	6	7	8
DIZER	yes	yes	yes	yes	yes	yes	**no**	yes
PERGUNTAR	yes	yes	yes	yes	**no**	**no**	**no**	yes
RESPONDER	yes	yes	yes	yes	yes	**no**	yes	yes
ADMITIR	yes	yes	yes	yes	yes	**no**	yes	yes

CONTAR	yes	yes	yes	yes	yes	**no**	**no**	yes
CONTINUAR	yes	yes	yes	yes	**no**	**no**	**no**	yes

Table 4: Occurrences of the six analyzed verbs in each proposed pattern.

Fonte: Freitas (2016)

Based on Table 4, it is possible to draw two conclusions. The first one is that all the analyzed verbs, as they are applicable to most established patterns, could be classified as RV. Even though the verb "continuar" was not found in three out of the eight patterns, this would not exclude it from this group of verbs, particularly because, in fact, we might have to acknowledge the existence of only five patterns – which leads us to the second conclusion.

As it was previously said, the grammatical patterns used with RV would be proposed from the analysis of six "seed" verbs and, simultaneously, these six verbs would put the patterns to test. Considering that only the verb "dizer" (*say*) can be applied to Pattern (6) and only the verbs "responder" (*answer*) and "admitir" (*admit*) would satisfactorily fit in Pattern (7), the universality of these patterns is questionable and they do not help us to identify other RVs as much. That is why, from now on, these patterns will be discarded.

Pattern (5) was accepted by four out of the six verbs, which, for us, is a suggestion that the pattern might be suitable for introducing reported speech. However, its low precision would inevitably lead to the retrieval of several verbs as false candidates for RVs. For this reason, we decided to eliminate this pattern. Indeed, we had five other patters that better suited our purpose: to build a reporting verbs lexicon. Taking the five main patterns into account, we move on to the next stage, in which we extend the list of RVs.

Extending the list of RVs: building a lexicon

Based on the patterns, we created search expressions specifically for AC/DC interface, in order to extend the verbs lexicon. The expressions are listed in Freitas (2016). We soon noticed that the patterns alone would not ensure the presence of a RV, since they would obtain structures such as the examples below, thus indicating the need to carry out an individual analysis:

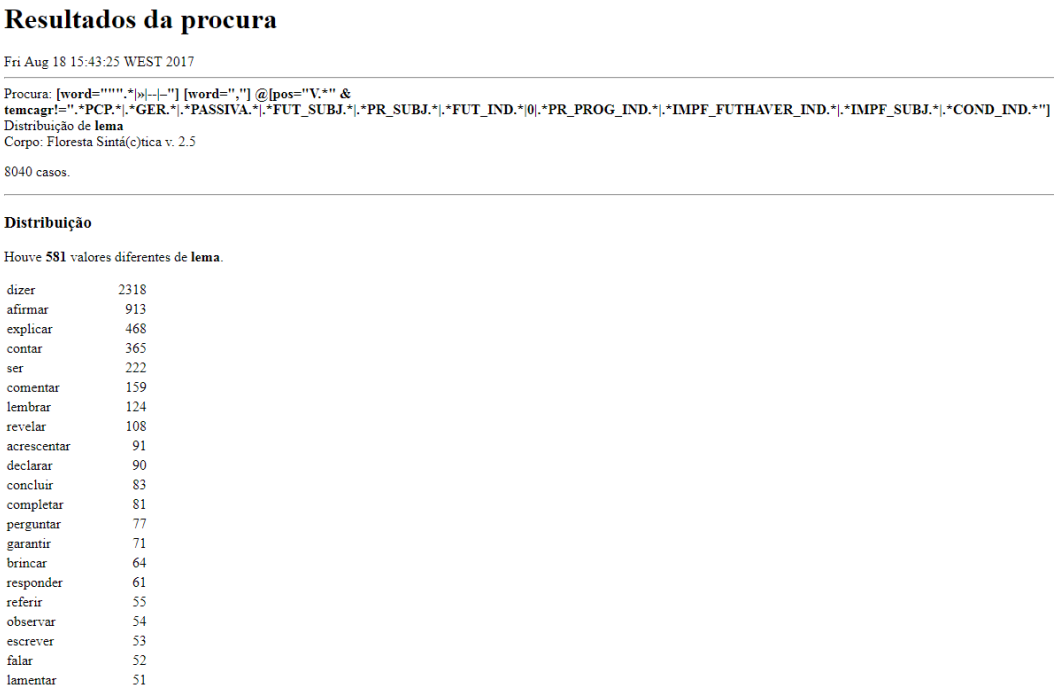
(a) *Verificada a licitude da «pré-campanha », **regressemos** ao caso do leitor.*

*Once we verified the legality of the «pre-campaign», we **return** to the case of the reader.*

(b) *Ao longo de quase dois anos, os habitantes de Reveles ainda **acreditaram** que os trabalhos de estabilização dos solos pudessem vir a evitar o pior.*

*Throughout almost two years, the Reveles inhabitants still **believed** that the soil stabilization work could prevent the worst.*

Then, we carried out the individual analysis, but this time with a corpus smaller than CHAVE, yet still considerably large: we used a corpus called Floresta (FREITAS et al., 2008), composed of 6,046,541 words. For every search expression used, we specified that the corpus provided us with the distribution of the verbs by lemma, as Figure 2 illustrates:



Resultados da procura

Fri Aug 18 15:43:25 WEST 2017

Procura: **[word="""|»|--|–"] [word="|,"] @[pos="V.*" & temcagr!=".*PCP.*|.*GER.*|.*PASSIVA.*|.*FUT_SUBJ.*|.*PR_SUBJ.*|.*FUT_IND.*|0|.*PR_PROG_IND.*|.*IMPF_FUTHAVER_IND.*|.*IMPF_SUBJ.*|.*COND_IND.*"]**
Distribuição de **lema**
Corpo: Floresta Sintá(c)tica v. 2.5

8040 casos.

Distribuição

Houve **581** valores diferentes de **lema**.

dizer	2318
afirmar	913
explicar	468
contar	365
ser	222
comentar	159
lembrar	124
revelar	108
acrescentar	91
declarar	90
concluir	83
completar	81
perguntar	77
garantir	71
brincar	64
responder	61
referir	55
observar	54
escrever	53
falar	52
lamentar	51

Figure 2: Floresta interface in a search for distribution by lemma.

In order to make the analysis viable, we did not consider the verbs that appeared in only one occurrence. Besides, we did not include in our case-by-case reading the 58 verbs already validated in the first stage.

This stage resulted in a list of 293 verbs. Table 5 shows the distribution of the verbs according to the search patterns. The data include repetitions, since the same verb may be used in different patterns. The complete list of the verbs we found is available at Gramateca (SANTOS, 2104) webpage, listed both in frequency and alphabetical order.

PATTERN	REPORTING VERBS
Pattern 1	219
Pattern 2	107
Pattern 3	137
Pattern 4	156
Pattern 8	52
TOTAL	293 (no repetitions)

Table 5: Number of verbs found with the search expressions in each pattern.
Source: Freitas (2016)

Based on Table 5, it is viable to draw some conclusions. First, it is noticeable that Pattern (1) is the most productive of all, once the percentage of verbs used in this pattern, when compared to the others, is the highest: 74.74%. Pattern (8), on the other hand, was the least productive, corresponding to only 17.74% of the compiled verbs.

Comparing results

In this part of the article, we present the results retrieved from the identification of patterns, comparing them with the results of other works on reported speech. Table 6 reveals the number of RVs obtained or listed by works that were cited in this research:

PUBLICATION	LANGUAGE	ANALYZED VERB CLASS	NUMBER OF VERBS
Sagot et al (2010)	French	reporting verbs	232
Levin (1993)	English	verbs of communication	163
Garcia (2010)	Portuguese	reporting verbs	77
Moura Neves (2000)	Portuguese	reporting verbs	103
Este trabalho	**Portuguese**	**reporting verbs**	**293**

Table 6: Comparative table of the reporting verbs presented in some studies.
Fonte: Freitas (2016)

Table 6 stresses the success of our approach in this research. However, the numbers alone could mislead us into thinking that all the authors cover the same phenomena and have the same interests in mind, which is not the case. The third column of the table shows which verb class was studied by each of the authors, according to our classification of the verbs.

Sagot et al. (2010) are concerned about automatic quotation extraction, and their interest in verbs is limited to solving the intended task; Biber et al. (1999) as well as Levin (1993) meant to carry out descriptive studies which have English as a target and, to do so, considered the broader group of verbs of communication; Garcia (2010) and the *Manual de Redação e Estilo do Estado* (MARTINS, 2016) are writing manuals; and Moura Neves (2000) corresponds to a grammar of Portuguese usage.

Moura Neves, as we stated previously, compiled 103 RVs, as opposed to the 293 verbs of this research. Among the 103 verbs of the *Gramática de usos do português* (MOURA NEVES, 2000), a total of 25 were not in our list. The 25 verbs are: **acalmar**, **agastar-se**, **aguilhoar**, antecipar, boquejar, **bronquear**, **bulir**, **caçoar**, cochichar, **conchavar**, **consolar**, **debicar**, **debochar**, **desiludir**, diagnosticar, **escarnecer**, **ferroar**, **inclinar-se**, **interceptar**, **maldizer**, participar, **remediar**, **suspirar**, sussurrar, **zombar**. All the verbs in bold belong to the same category in Moura Neves's work, *verbs that instrumentalize or circumstantiate was is said* – that is, verbs that may introduce speech, but are not necessarily speech acts. Given that number, and considering that, in spite of the great number of verbs we found, we know that our list is not exhaustive, we decided to test Moura Neves' list in the three monolingual corpora that were used in this study.

First, we wanted to understand why those 25 verbs had not appeared in our list. We consulted all the AC/DC material, and then identified three different reasons for the absence of the verbs: (i) they only featured in one occurrence (and were therefore discarded in our initial selection); (ii) they were not found among the occurrences; or (iii) the occurrences did not correspond to RVs (e.g.: "Com o vídeo "Circular", **participou** este ano do prestigiado "Curta Cinema", Festival Internacional de Curtas do Rio."/ *With the video "Circular", this year he **participated** in the prestigious "Curta Cinema", Rio International Short Film Festival*).

After that, we needed to confirm if those verbs could function as RVs – and, if so, evaluate whether they would be included in our list. From the 25 verbs, a total of 15 were used as RVs in all corpora[16].

The fact that we could not find ten of the verbs proposed by Moura Neves (2000) does not necessarily mean that these verbs must not be used as RVs. We must remember that, despite the dimension of the consulted corpora, no corpus will ever be exhaustive, nor will it account for all phenomena in a language. As the corpus Moura Neves used is not available to the public, we have no means of verifying if the occurrences that were analyzed in the making of the grammar book correspond, in our point of view, to RVs. It is worth noting, nevertheless, that the absence of these ten verbs in corpora that, together, amount to over a billion words, might indicate that the use of such verbs as RV is unusual.

[16] We would like to emphasize that a search in all corpora corresponds to looking into over a billion words, distributed across different text genres.

In addition to that, we also drew a parallel between Moura Neves's work and this study concerning the type of speech. One of the angles adopted by Moura Neves consisted of determining if the RVs were accepted in direct speech and/or in indirect speech. We reproduced the author's classification, regarding (a) the verbs in which there was disagreement as to the type of speech associated with the verb and (b) the verbs only listed by our study, thus collaborating with the author's work. Due to space constraints, we will not present the table here, but it is available in Freitas (2016).

4 FINAL CONSIDERATIONS

In this article, we presented the methodology for developing a reporting verbs lexicon in Portuguese. This paper consists of a part of a larger work, which also involves the elaboration of DISSE, a reporting verbs glossary whose main goal is to help professional translators as well as beginners (FREITAS, 2016). In order to do so, we conducted an extensive, descriptive study of this verb group. Our work was based on large Portuguese corpora and indirectly involved Computational Linguistics, as it was motivated by the quotation extraction task. As a matter of fact, the verbs and patterns that we described in this paper have already been integrated in the development of an automatic quotation detection system for Portuguese.

As another result from the lexicon and the patterns, the annotation of AC/DC corpora is also in progress, which will allow users to search by the semantic field *say* – the annotation process is described in Freitas (2016). All the materials are public and are available for online consultations.

Based on corpora, we established eight general patterns related to RVs, from which it was possible to gather these verbs and then build a large lexicon of RV in Portuguese. This lexicon is composed of 308 verbs (the 293 verbs that were found in our research as well as the 15 verbs that were listed by Moura Neves, but were confirmed by our study), and all verbs were manually validated through occurrences.

In quantitative terms, the fact that we compiled 293 verbs, that is, more verbs than any other study that we came across, whether in Portuguese, whether in other languages, highlights the successful path we have chosen – consulting large corpora. The fact that we obtained patterns involved in speech, or, in other words, the observation that the same type of quotation (direct, indirect or mixed) can be expressed in different patterns is also, in our opinion, a valuable contribution, and it brings to light the potential of corpus-based descriptive studies on the Portuguese language.

Finally, we hope that we have demonstrated, throughout this work, how productive the dialogue between language description, translation and computational linguistics can be. Descriptions, of any kind, cannot be an abstract activity, detached from intentions and purposes. This led our research to establish associations with Translation and Computational Linguistics, which are applied studies.

REFERENCES

AUSTIN, J. L. *How to do things with words*. 2. ed. Cambridge: Harvard University Press, 1962.

BIBER, D. et al. *Longman grammar of spoken and written English*. Harlow: Pearson Education ESL, 1999.

COSTA, L.; SANTOS, D.; ROCHA, P. A. Estudando o português tal como é usado: o serviço AC/DC. In: PARDO, T., NUNES, M.G.V. *The 7th Brazilian symposium in information and human language technology (STIL 2009),* São Carlos,[2009]. Disponível em: <http://nilc.icmc.usp.br/til/stil2009_English/Proceedings/stil/Costa-57572_1.pdf>. Acesso em: 18 ago 2017.

FRANKENBERG-GARCIA, A; SANTOS, D. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*, Florianópolis, v.1, n. 9, p. 61-79, 2002.

FREITAS, B. *O dizer em português*: diálogos entre tradução, descrição e linguística computacional. 2016. 116f. Dissertação (Mestrado em Estudos da Linguagem) – Programa de Pós-Graduação em Estudos da Linguagem, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

FREITAS, C.; ROCHA, P.; BICK, E. Floresta sintá(c)tica: bigger, thicker and easier. In: TEIXEIRA, A. et al. (Ed.). *Computational processing of the portuguese language*: 8th international Conference, PROPOR 2008. Gemany: Springer Verlag, 2008. p. 216-219.

GARCIA, O. M. *Comunicação em prosa moderna:* aprenda a escrever, aprendendo a pensar. 27. ed. Rio de Janeiro: FGV, 2010.

HOUAISS, A.; VILLAR, M. S. *Grande dicionário Houaiss*. Verbete "Considerar", [2009]. Disponível em: <https://houaiss.uol.com.br/>. Acesso em: 25 ago. 2016.

KIPPER, K. et al. Extending VerbNet with novel verb classes. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2006), 5., *Proceedings*...Genova, Itália, 2006.

LEVIN, B. *English verb classes and alternations*. Chicago: The University of Chicago Press, 1993.

MARTINS, E. *Manual de redação e estilo do Estado*. Verbete "Verbos mais que errados". Disponível em: <http://www.estadao.com.br/manualredacao/>. Acesso em: 10 jan. 2016.

MOURA NEVES, M. H de. *Gramática de usos do português*. São Paulo: Editora Unesp, 2000.

PARETI, S.; O'KEEFE, T.; KONSTAS, I.; CURRAN, J. R. e KOPRINSKA, I. Automatically detecting and attributing indirect quotations. In: YAROWSKY, D. et al. (Ed.). *Proceedings of the 2103 conference on empirical methods in natural language processing (EMNLP 2013)*, Washington, USA, 2013. p. 989-999. Disponível em: < http://www.aclweb.org/anthology/D/D13/D13-1101.pdf>. Acesso em: 18 ago. 2017.

SAGOT, B.; DANLOS, L.; STERN, R. A lexicon of French quotation verbs for automatic quotation extraction. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*, La Valette, Malta, 2010. p. 294-299. Disponível em: < http://www.lrec-conf.org/proceedings/lrec2010/pdf/387_Paper.pdf>. Acesso em: 18 ago. 2017.

SANTOS, D. Gramateca: corpus-based grammar of Portuguese. In: BAPTISTA, J. et al. (Ed). *Computational processing of Portuguese*:11th international conference *(PROPOR 2014).* Germany: Springer, 2014. p. 214-219.

SANTOS, D.; ROCHA, P. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In: PETERS, C. et al. (Ed.). *Multilingual information access for text, speech and images,* Workshop of the Cross-Language Evaluation Forum, 5, Bath, 2004, Revised Selected Papers. Berlin/Heidelberg: Springer, Lecture Notes in Computer Science, 2005. p. 821-832.

SCARTON, C.; ALUISIO, S. Towards a cross-linguistic VerbNet-style lexicon to Brazilian Portuguese. In: LAMBERT, P. et al. (Ed.). *Proceedings of the LREC 2012 workshop on creating cross-language resources for disconnected languages and styles (CREDISLAS 2012)*, Istambul, 2012. p. 11-18. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/index.html>. Acesso em: 15 ago. 2017.

TRUGGO, L. F. *Classes de palavras - da Grécia Antiga ao Google*: um estudo motivado pela conversão de tagsets. 2016. 113f. Dissertação (Mestrado em Letras/Estudos da Linguagem) – Programa de Pós-Graduação em Estudos da Linguagem, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeir

Recebido em 24/10/2016. Aceito em 17/12/2016.