Attribution of Quoted Speech in Portuguese Text

Eckhard Bick
University of Southern Denmark
eckhard.bick@gmail.com

Abstract

This paper describes and evaluates a rule-based system implementing a novel method for quote attribution in Portuguese text, working on top of a Constraint-Grammar parse. Both direct and indirect speech are covered, as well as certain other text-embedded quote sources. In a first step, the system performs quote segmentation and identifies speech verbs, taking into account the different styles used in literature and news text. Speakers are then identified using syntactically and semantically grounded Constraint-Grammar rules. We rely on relational links and stream variables to handle anaphorical mentions and to recover the names of implied or underspecified speakers. In an evaluation including both literature and news text, the system performed well on both the segmentation and attribution tasks, achieving F-scores of 98-99% for the former and 89-94% for the latter.

1 Introduction

In text linguistics, quote attribution is the task of identifying the person or entity behind a quoted utterance, as well as delimiting the quote itself. Automatic tools capable of robustly performing this twin task may be used in a variety of scenarios, such as the extraction of character networks from novels (e.g. Elson et al., 2010; Vala et al., 2016; Santos et al., 2022), voice assignment in text-to-speech systems, information extraction from news channels (Sarmento and Nunes, 2009) or the validation of social media claims (Janze and Risius, 2017). In this paper we will distinguish between speakers and sources, associating the former with direct quotes (1a) and the latter with indirect quotes (1b) and in-text source references (1c). Both speakers and sources may be either narrative characters (including a first-person narrator) or real-life people, organizations and institutions, depending on whether the text in question is fiction or non-fiction (e.g. news).

(1a) ["/--]The attack caused widespread fires["], the mayor said.

(1b) The mayor also said that the attack caused widespread fires.

(1c) According to the mayor, Vitali Klitschko, the attack caused widespread fires.

Quote attribution must be distinguished from a wider approach to source identification that would include, for instance, photo or article sources (2a), or the bibliographic attribution of scientific findings (2b).

(2a) Russian reservists leaving for the front. EPA/YURI K.

(2b) Yearly precipitation increased by 24% over the decade (Moulder & Huggins, 2016)

The work presented here excludes these source types and focuses on quote attribution.

Though there is a growing body of research in the field, most work has been carried out for English. Among the (few) publications about our own target language, Portuguese, are (Mamede & Chaleira, 2004) and (Sarmento and Nunes, 2009), who address characters in children's' books and news quotes, respectively. Most systems use various machine-learning (ML) techniques exploiting, besides frequency and recency of mention, features from morphosyntactic and other available linguistic annotation, not least named entity recognition (NER). For instance, Elson and McKeown (2010) treat attribution as an ML classification task, while O'Keefe et al. (2012) use ML with a sequence-labelling method. Systems also differ in task scope. Thus, Ek et al. (2018) annotate addressees, including collective addressees, in

Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar - Methods, Tools and Applications, pages 1–9

addition to speakers, which they categorize as either explicit, anaphoric or implied[1].

Our own work is different from most current research not only in terms of target language (Portuguese) and by including both literature and news data, but also because it pursues a rule-based approach, exploiting the Constraint Grammar paradigm (Bick and Didriksen 2015) to harness complex context conditions and assign relational tagging for coreference resolution and speaker mark-up. Apart from linguistic transparency and efficiency in a sparse-data situation, rule-based systems support straight forward genre adaptation, even in the complete absence of training data, by adding new rules (or exceptions to existing ones). Interestingly, O'Keefe et al. (2012) found that a simple rule-based baseline[2] outperformed their ML approach for speaker attribution in literature, and proved to be on par for for mixed news (Sidney Morning Herald). Only for the Wall Street Journal did ML work better.

The scope of our attribution annotator includes speakers/sources in both direct and indirect speech, regardless whether the information is explicit, anaphoric or implied. However, given the fact that accuracy for listener/addressee identification tends to be almost half that obtainable for speaker/source (Yeung and Lee, 2017; Ek et al. 2018), the former was not included here. Given that speaker attribution is a high-level linguistic task, we believe that the methodology of our Portuguese set-up can be generalized to other languages, or at least be used for inspiration and comparison.

2 Parsing technology

Our attribution rules are run on top of a full morphosyntactic and semantic annotation provided by the PALAVRAS parsers (Bick 2014). The system provides reliable tagging and disambiguation for lemma, POS, inflection, and semantic class including named entities, as well as dependency and frame structures. Our own rules are an extension of PALAVRAS' anaphora and coreference relations and make use of existing ID-links. While our rules make reference to many different tag types, the most important ones are, obviously, the speech- and speaker-related ones: +HUM semantic classes, speech verbs and their subjects and object clauses, as well as related semantic roles (§ATR – attribute, §ID – identity, §SP – speaker, §MES – message).

The attribution annotator does not use CG's traditional MAP, SELECT and REMOVE rules, as we do not treat attribution as a disambiguation task. Rather, tags are inserted using CG3's SUBSTITUTE rules type in a sequential fashion – supporting tags first (quote delimiters, quote heads and quoting verbs), then the primary tags (speaker/source), progressing from safe/close contexts to more heuristic long-distance contexts. In addition, we use the relatively new CG3 feature of stream variables[3] to store and retrieve text and paragraph-level information about turn-taking, previous speaker and speaker-associated noun phrases.

3 Quote and dialogue segmentation

3.1 Quote types and annotation

The density of quotes is extremely text-dependent. Thus, in their work on classical English literature, Elson and McKeown (2010) found a spread of 19-71% of text included in quotes. For our own, Portuguese data the quote density was also high, higher for news (51.7%) than for the literature sample (42.1%). Interestingly, there was a considerable difference when comparing direct with indirect quotes, with the former being frequent in literature (87.3% of all quotes), but rare in the news data (11.7% of all quotes) – a difference with a possible bearing on performance, as indirect quotes are more likely to have a close/syntactic link to a quoting verb, while direct speech may occur in isolation, with quoting information left implicit or provided in another sentence.

1 Some systems handle only the explicit category, and others who do include anaphoric pronouns and np's may do so verbatim without resolving the reference by linking to a name. Our own system handles all three types, but attempts to resolve all as names, with noun phrases as an under-specified fall-back solution.

2 search backwards from the end of the quoted sentence until the nearest speech verb, then pick the nearest named entity mention.

3 Stream variables are different from CG3's tag unification variables. While the latter are local and limited to the containing rule, stream variables are are part of the input/output stream and visible to all rules. Stream variables can be set either externally or by the CG rules themselves. In the newest edition of CG3, both names and values of stream variables can be written, matched and excerpted using regular expressions and ordinary tag variables.

In Portuguese, direct quotes may be optionally marked with either opening quotation marks (3c, 4b, common in news text) or a dash (3a-b, common in literature), both of which will be repeated if the quote continues after a "backward" quoting verb (3a). Opening quotation marks are always matched with closing quotation marks (3c), but closing dashes are only used before quoting verbs, not if the latter precedes the quote (3b). In the absence of other punctuation, a comma is added between a quote and a "backward" quoting verb. In one quoting style (3d), the closing comma is the only visible quote delimiter.

We use three quote delimiter tags: <quote-edge> (start), <quote-end> (stop) and <quote-ana> (for quote continuation after an inquit). In addition, the quoting verb is tagged <v-quote>. The quote itself is marked on its syntactic top node (3a-d), with <quote> for direct speech, or <quote-ind> for indirect speech.

(3a) -- <quote-edge> *É verdade, faz* <quote> *medo, mas é bonito* - <quote-end> *acrescentou* <v-quote> *Eulália*. - <quote-ana> *Hei* <quote> *de ir sempre ver.* [it's true, it is frightening, but it is beautiful, Eulália added. I have to watch it all the time]

(3b) *Era, pois, sincero, quando, de joelhos, exclamou* <v-quote>: – <quote-edge> *Porque te amo* <quote>. [He was, thus, sincere, when he, on his knees, called out: "Because I love you"]

(3c) " <quote-edge> *A situação na região de Odessa é muito difícil"* <quote-end>, *começou por dizer* <quote> *o Presidente ucraniano* [The situation in the region of Odessa is very difficult, the Ukrainian president said when he took the floor]

(3d) *Em Odesa, registaram-se* <quote> *ataques com "drones" durante a noite, que deixaram grande parte da região sem eletricidade,* <quote-end> disse <v-quote> o chefe do Governo local, Maxim Marchenko. [In Odessa, during the night, drone attacks were registered, which left a large part of the region without electricity, the local governor, Maxim Marchenko, said.]

Automatic quote annotation has to distinguish quote segmentation from other uses of quotation marks (e.g. titles, literatim-markers [4a] or special words [3c]) and dashes (e.g. parenthetical material), and it has to take into account "forward" quoting constructions with a colon, but potentially no other delimiter. Also, a quote may encompass more than one utterance, so possible quotation marks or hyphens may be outside the window of analysis.

(4a) *O Presidente ucraniano agradeceu a Washington pela "forte parceria" e descreveu a visita de Biden como "histórica, oportuna e corajosa"*. [The Ucrainian president thanked Washington for its "strong partnership" and described the Biden visit as "historical, timely and courageous".]

Given the underlying dependency annotation, the <quote>/<quote-ind> tags are enough to extract the quote even without the use of delimiter tags. If present, quoting verbs are either forward- (colon-style) or backward-pointing, but unlike the <quote> marker, <v-quote> may also be absent or implied (4b).

(4b) *A avaliação é do Ministério da Defesa britânico: "A contínua priorização de infra-estrutura nacional crítica (…)"* [The evaluation is the British Ministry of Defense's: "The continued targeting of critical national infrastructure (…)]

All six types of quote markers double as CG rule barriers (or barrier exceptions), especially when searching across multiple sentence windows, telling the rule "cursor" when it leaves or enters a quote, whether a sentence is part of a multipart quote, or – in dialogue – if a quote is adjacent or isolated by narrative body text.

Below are two simplified examples of <v-quote> mapping rules[4], relying on a speech-verb frame tag (<fn:speak>) and the presence of either a <quote-end> token (first rule) or a post-positioned human subject (second rule):

```
SUBSTITUTE (V) (<v-quote> V)
  TARGET VFIN + @FV + <fn:speak>
  (*-1 <quote-end> BARRIER NON-KOMMA);

SUBSTITUTE (V) (<v-quote> V)
  TARGET VFIN + @FV + <fn:speak>
  (cr @<SUBJ + N-HUM-person);
```

4 In their simplified form, the rules could be combined by using the OR convention to include both context conditions in one rule: ((*-1 ….) OR (*1 ….))

While direct quotes have the syntactic structure of main clauses and may constitute independent sentences, indirect quotes have subclause structure and a dependency link to the quoting verb. In this case, we add <quote-ind> to the top node of the subclause[5], again facilitating dependency-based quote extraction (5a-b). Note that in Portuguese, indirect quotes may be infinitives (5b), a construction common in the news domain.

(5a) *Ela disse que não o queria* <quote-ind> *fazer*. [She said that she didn't want to do it.]

(5b) *Ele disse ser* <quote-ind> *absurda a alegação.* [He said that the allegation was absurd.]

The identification of speech verbs is of great importance for other annotation tasks. Thus, it triggers the recognition of a comma as <quote-end>, and a preceding clause as <quote>. Above all, however, speech verb identification facilitates speaker identification, either directly, through a proper noun subject dependent, or indirectly through reference links for zero-subjects and pronouns. Like Elson and McKeown (2010), we use an external semantic resource to identify speech verbs, but instead of their lexical WordNet categories, we use verb classes from PALAVRAS' framenet annotation (Bick 2022), which have the advantage of being disambiguated and linked to tokens carrying semantic role tags for speaker (§SP) and message (§MES), respectively. Verbs without a speech frame proper (e.g. *wonder*, attack) may still be identified if they are reinforced by a post-positioned +HUM subject in the pattern:

…, + finite_verb + human_subject

For continuation verbs (*continue, add, insist*) the subject is usually omitted in Portuguese (corresponding to pronoun use in English), but in sentence final position, the pattern is still a reliable speech indicator:

…, + continuation_verb [adverbials] .

3.2 Dialogue and turn-taking variables

In addition to quote annotation, dialogue segmentation has obvious benefits for speaker attribution (Yeung and Lee, 2017)[6]. For instance, a vocative mention in one quote paragraph may help identify the speaker of an immediately preceding or following quote paragraph. In our system, we set a turn-taking variable when a quote-opening token is found, alternating the variable value between 1 and 2 in order to keep track of speaker turns. The variable is un-set after the paragraph that contained the quote-opener. The turn-taking variable either directly as a CG3 local variable (LVAR) or as a tag mapped on relevant tokens. It allows us to unify speakers across alternating turns, propagating information from e.g. the first, explicitly quoted, turn to later turns[7] by the same speaker in the same dialogue chain. The rule example captures a speaker variable \(...\) from an established SPEAKER tag in the same turn type (here: turn 1) either left or right in the window span (*0W). To make sure the turns belong to the same dialogue span, there is a BARRIER for top node verbs (@FV) that are not quoted (<quote>) or quoting (<v-quote>), i.e. that represent ordinary, narrative text. The captured variable ($1) is then inserted as SPEAKER on the target quote.

(R-1) SUBSTITUTE
(<quote>) (<quote> <SPEAKER:$1>v)
TARGET (<quote> <turn-1>)
(*0W (<SPEAKER:\(.*\)>r <turn-1>)
BARRIER @FV - <quote> - <v-quote>) ;

4 Attribution methods

We assign a <SPEAKER:...> tag to each direct quote, and a <SOURCE:...> tag to each indirect quote, mapped on the <quote> and <quote-ind> tags, respectively. Ideally, the value for SPEAKER and SOURCE should be a name, but as a fall-back, definite noun phrases are accepted. With the exception of anonymous or group utterances, quotes in literature should ultimately be traceable to a character name, but

5 The semantic parser will already have marked the subclause as §MES (message), but on its main verb, which may be different from the top-node finite verb.

6 Naturally, this is relevant only if the work in question contains structured, but unexplicit dialogue. In their own experiment, Yeung and Lee (2017) did not find any improvement for the New Testament.

7 This kind of speaker propagation also works backward. To achieve this, we have to run the attribution grammar twice: Because CG works sequentially from left to right, "future" (later-in-text) information will only be accessible for reference in a repeat run of the grammar. A section rerun would have the same effect, but the cg3 formalism only foresees this for disambiguation grammars, not pure substitution or relation grammars.

in the news domain, sources may be institutions (e.g. the Ministry of Defense) or officials not linked to a human name, but to a function (e.g. the local mayor or a Red Cross representative). The rule examples and variable use discussed in this section focus on the <SPEAKER:...> tag, but mostly hold for the <SOURCE:..> tags as well.

4.1 Direct attribution: Quoting verbs and source references

The safest speaker identification is through an associated quoting verb (<v-quote>), found either (a) as a dependency head, (b) immediately after a <quote-end> tag or before a <quote-ana> tag, or (c) as the closest top-level verb before a quote-opening colon. Departing from the speech verb, the prioritized order of speaker extraction will then be the following:

1. subject dependent: name
2. subject dependent: noun phrase or pronoun with a <REF:....> name tag, or r:ref relation leading to a name
3. no surface subject, but a <REF:...> or <SUBJ:...> name tag, or a r:ref name relation, on the verb

For (b) and (c), in order to reach the relevant left or right quote delimiter and quoting verb, sentence boundaries may have to be crossed in the case of multi-sentence quotes. As mentioned in section 3, this can be made safer by using the various <...quote...> markers as barriers or barrier exceptions, but ultimately there is the risk of confusing a sentence boundary with a paragraph boundary and retrieving a wrong speaker value. We therefore introduced a paragraph-numbering variable that can be checked to see if the encountered <v-quote> and is located within the same paragraph. In dialogue, where each turn fills a whole paragraph, the turn-taking variable can be exploited to the same end.

Another scenario for direct attribution is the use of reference pointers in adverbial constructions with *segundo, conforme* and *de acordo com* (all meaning 'according to'). Independent of word order and syntax, these proved to be very safe source indicators for citations[8] occurring in the same sentence. Rule R-2 looks left or right (*0) for the trigger words *segundo* or *conforme*, harvesting a referent name or lemma either directly form their argument (c=child) or from the +HUM subject in a dependent clause. Typically, PALAVRAS will have assigned these constructions a §META role.

```
(R-2) SUBSTITUTE
  (V) (VSTR:<SOURCE:$1> V)
  TARGET <fmc>
  (*0 ("segundo") OR ("conforme")
    LINK (0 §META LINK c @P<)
      OR (c §META)
      OR (1 VFIN LINK 0 <fn:speak>
        LINK cr N/PROP-HUM + @<SUBJ)
    LINK 0 (<REF:\(.*\)>r) OR ("<(.*)>"r)) ;
```

4.2 Indirect or implied attribution and speaker propagation

Similarly, if a quote has no associated quoting verb of its own, but the preceding sentence contains a direct or indirect quoting construction, its speaker may be copied ($2 variable below) as long as both sentences are in the same paragraph[9] (cf. the 'par' $1 variable in the rule below).

```
(R-3) SUBSTITUTE
  (<quote>) (<quote> <SPEAKER:$2>v)
  TARGET (<quote>)
  (0 (VAR:par=/\([0-9]+\)/r))
  (*-1 <quote-edge> OR <quote-ana>
     BARRIER <quote-end> OR <v-quote>
     LINK -1 >>> LINK *-1W ALL-ORD
     LINK *-1 <v-quote> BARRIER <quote>
     LINK cr @<SUBJ
     LINK *-1 (<SPEAKER:\(.*\)>r)
     LINK 0 (VSTR:LVAR:par=$1)) ;
```

If there is no speaker mention or speech verb subject reference found in the same paragraph, and not even an anonymous speaker can be assigned, the grammar defaults to speaker alteration using a pair of stream variables, speaker (R-5) and oldspeaker (R-4). When a new speaker is established, the speaker variable is reset and its previous value stored as "oldspeaker". Unless a turn is marked as "continuing" (verb frame), it is "oldspeaker" that

8 It should be noted, though, that such citations, unless framed in quotation marks in their entirety, may be rephrasings or gists, and need not exhibit the same literatim fidelity as direct or indirect speech with a quote verb.

9 if not, they may belong to different turns, with different speakers.

will be used for hitherto unattributed quotes (R-6).

```
(R-4) SETVARIABLE (oldspeaker) (VSTR:$1)
  TARGET (<SPEAKER:.*>r)
  (0 (VAR:speaker=/\([^\\^].*\)/r)) ;

(R-5) SETVARIABLE (speaker) (VSTR:$1)
  TARGET (<SPEAKER:\([^\\^].+\)>r) ;

(R-6) SUBSTITUTE (<quote>)
  (<quote> VSTR:<SPEAKER:$1>)
  TARGET (<quote>)
  (0 (VAR:oldspeaker=/\(.*\)/r))
  (NEGATE *1W <v-quote>
    BARRIER @FV - <quote>
    LINK 0 <fn:continue> OR <fn:add>
  (NEGATE *0 @VOK
    LINK 0 ("$1"v) OR (<REF:$1>v)) ;
```

4.3 Reference links, tags and variables

An important aspect of our attribution method is keeping track of who is who through stream variables and through the use of co-reference links and tags. The task is a formidable one: 40-50% of quote chunks do not contain a quoting verb, leaving the speaker implied or obliquely mentioned. Of those that do feature a quoting verb, the latter may lack a surface subject (28% in our literature data, 14% in news), or the surface subject may be an anaphorical pronoun or underspecified noun phrase.

For anaphora resolution, we use CG3's ADDRELATIONS operator to establish referent links between pronouns and underspecified noun phrases and a target referent, optimally a named entity (NE). The equivalent solution for subject-less verbs are elliptic-subject relations. In both cases, name targets will also be mapped, as <REF:name> tags, on the anaphorical element itself. This is useful for "promoting" the antecedent information, if link targets are themselves anaphorical (e.g. chains of pronouns or subject-elliptical verbs), in which case the ultimate name referent may be outside the rolling CG focus window (set to ±6 sentences in this grammar). In (6), for instance, the second quoting verb, *disse* [said], is subject-elliptic, but the anaphora rules will link it to the nearest top-level human subject to the left, in this case the explicit subject (*Biden*) of the first quoting verb, *afirmou* [asserted]. To this end, contextual syntax- and semantics-informed rules are much safer than e.g. just going for the closest NE or even human NE, which in this case would yield the wrong speaker, *Zelensky*.

(6) *"A Ucrânia resiste. A democracia resiste", afirmou Biden, ao lado de Zelensky. "Putin achou que a Ucrânia era fraca e que o Ocidente estava dividido", disse.* ["Ukraine resists. Democracy resists", Biden asserted, with Zelensky by his side. "Putin thought Ukraine was weak and the West divided," he said.]

In addition to anaphora links, we use stream variables to store relevant established information across analysis windows. Apart from the afore-mentioned turn-taking and paragraph variables, we store "new speaker name" and "old speaker name" (cf. section 4.2). For anaphora resolution, we set a variable for most recent top-level subject and a "social function" variable (professions and functional titles) for nouns referring to names. This type of information storing goes beyond simple fixed variables, as it can't be known beforehand, which and how many social functions a text may contain.

Thus, every time a common-noun reference has been resolved to a proper noun (7a), the latter is stored as the value ($1 in R-7) of a *newly created* variable ($2 in R-7) carrying the name of the former, appended to a prefix *nattr-* (name attribute). The prefix allows a blanket resetting of all noun-speaker variables at major breaks in the text, such as chapter or news article headlines.

```
(R7) SETVARIABLE
  (VSTR:nattr-$2) (VSTR:$1)
  TARGET ("<(.+)>"r PROP £hum)
  + (<NA:Hprof/\(.*\)>r) ;
```

The name value can then be retrieved (as $2 in R-8) for "underspecified" speakers ($1 in R-8), i.e. speaker mentions that were nouns rather than names (7b), exploiting information from an earlier paragraph (7a) in the same article or chapter.

(7a) *Bombardeamentos (…). O anúncio foi feito pelo* ***governador*** *da região,* ***Pavlo Kyrylenko****, no Facebook: "(...)".* [Bombardments … The announcement was made by the region's governor, Pavlo Kyrylenko, in Facebook: ...]

(7b) *A cidade é (…). Segundo o* ***governador****, (…) , "é impossível determinar ..."* [The town is

… According to the governor, … “it is impossible to determine ...”]

(R-8) SUBSTITUTE
(<SPEAKER:.*>r) (VSTR:<SPEAKER:$2>)
TARGET (<SPEAKER:\([a-z].*\)>r)
(0 (VSTR:VAR:nattr-$1=/\(.*\)/r)) ;

As a fall-back alternative to variable-based name retrieval, a single-noun reference can be expanded through rules harvesting and adding its post-nominal dependents. This way, *governador* (governor) can be specified as *governador de Lugansk* (the Luhansk governor), and sources such as ministries and intelligence services may be specified for resort or nationality.

5 Evaluation

With two main applications in mind, character cast extraction and information extraction, the system was evaluated on two very diverse sets of data, historical literature to cover the former, and news text to cover the latter. Specifically, we used the first 7% of José do Patrocínio’s “Os Retirantes”, published in 1879 (13164 parser tokens, ca. 380 quotes), and a collection of articles[10] from the *Público* newspaper covering the Ukraine war between 3 August 2022 and 20 February 2023 (35076 parser tokens, ca. 610 quotes). In addition to extreme differences in vocabulary, syntax, style and orthography, there was a marked difference in quotation style, with 97.6% direct (SPEAKER) quotes in “Os Retirantes” and 63.9% indirect[11] (SOURCE) quotes in the war news. The relative number of quoted sentences was higher in the literature sample, but the quotes were longer in the news text.

5.1 Quote recognition and segmentation

Quote recognition worked well or both text types and both quotation styles (table 1), with F-scores of around 99% for direct speech (<quote>), and 97%[12] for indirect speech (<quote-ind>). The good results for quote recognition are not surprising given that most quotes in the literature sample were in separate paragraphs and marked with an opening dash, while the prevailing indirect quotes in the news sample were dependency-linked to a speech verb.

mark-up	**literature**			**news**		
	R	**P**	**F**	**R**	**P**	**F**
quote	98.1	99.7	98.9	99.1	98.7	98.9
v-quote	100	98.9	99.4	100	97.8	98.9
quote-edge	100	98.1	99.0	100	98.6	99.3
quote-end	97.6	97.6	97.6	98.6	100	99.3
quote-ana	96.1	96.1	96.1	100	100	100
quote-ind	(100)	(100)	(100)	96.1	98.4	97.2

Table 1: Precision, recall and F1-score for quote recognition and segmentation

Annotation of the quoting verbs (<v-quote>) and segmentation markers (<quote-edge>, <quote-ana> and <quote-end>) for direct speech was also very robust, but a little less so for the (sometimes ambiguous) dashes used in literature for <quote-end> and <quote-ana> than for the quotation marks used in news. For indirect quotes, segmentation was implicit and hence unevaluated, with the quoting verb assumed to be the dependency head of the <quote-ind> node, and segmentation implied by the dependency structure.

5.2 Speaker/source attribution

The second part of the evaluation concerned the more difficult task of speaker and source attribution. Most existing research has focused on the former rather than the latter. For English, in a cross-author testing, Ek et al. (2018) achieved F-scores of 41.3-73.4 (mostly around 70). Elson & McKeown (2010) achieved a higher F-Score (83%), for a mixed-author quote corpus, but included the gold-annotation of the preceding quote as a feature for their classifier. He et al. (2013) report 74.8-82.5 for speaker identification in direct quotes, with a ±1-paragraph window. As expected, explicit speakers were unproblematic (F=100), while anaphoric and implied speakers were harder, with F-scores of 76.4 and 63.1, respectively.

Our own system for Portuguese achieved an F-score of 94.7% for speaker identification in

10 https://www.publico.pt/2022/02/24/infografia/russia-invade-ucrania-guia-visual-entender-guerra-661 [retrieved 23 February 2023]

11 While direct quotes are clearly marked as such, the borderline for indirect quotes is a little more fuzzy, and based on verb semantics. For instance, the object clauses in the frames *defend cognitively* and *reject* were not counted as quote, while those in *promise* frames were included.

12 There were too few instances (9) of indirect speech in “Os Retirantes” to meaningfully compute performance.

news, and 92.0% for literature, with relatively small differences between recall and precision (table 2). One obvious explanations for the difference between literature and news is that we counted (full definite) noun phrases as correct speaker references, but not pronouns, and that the former are more typical of news text than in literature, where there is a limited, but more constant, set of characters with anaphorical or implied references to names. For both text types, results are about one percentage point better if computed for correctly identified quotes only. For the news domain, accepting underspecified noun phrases as speaker also led to higher scores (F=96.1). Source identification (indirect speech and "according to"-type references) proved to be more difficult than direct quote attribution[13], with 50% higher error rates (F=91.4 for news[14]).

mark-up	**literature**			**news**		
	R	**P**	**F**	**R**	**P**	**F**
SPEAKER	91.3	92.8	**92.0**	94.9	94.5	**94.7**
on corr. quotes	92.8	93.0	92.9	95.8	95.8	95.8
w/ underspecif.	-	-	-	96.3	95.9	96.1
SOURCE	88.9	88.9	**88.9**	90.4	92.5	**91.4**
w/ undersp.	-	-	-	94.3	96.5	95.4

Table 2: Precision, recall and F1-score for speaker and source

These results, albeit measured with a "soft", inspection-based method without a pre-determined gold-standard annotation[15], compare favourably with prior research for English, where good results may depend on the exclusion of anaphora and implied mentions, e.g. (Zhang and Liu, 2022) with an F-score of 87% for explicit speakers in direct speech. The best and most comparable results for Portuguese were reported by Sarmento and Nunes (2009), who crawled direct and indirect news quotes, but pursued an extreme precision-oriented approach, achieving P=98.2% for speaker attribution by excluding all anaphorical references and accepting only explicit named-entity mentions as speaker candidates.

Error inspection revealed that about 1/3 of attribution errors in the news data could be traced back to parsing errors, mostly syntactic function / dependency errors, but also a couple of POS errors (both leading to false positives). For the literature sample, due to the prevalence of direct quotes and scarcity of syntactically linked surface speaker names, errors were mostly due to complex rule interaction problems rather than (local) base parse errors[16].

6 Conclusion

We have shown how a rule-based and context-aware (CG) system can reliably exploit existing dependency and framenet annotation for quote attribution in Portuguese, stressing the importance of long-distance referent links and the use of annotation-aware speaker and turn-taking variables.

Given that the context conditions in the attribution rules make use of higher-level, universal linguistic categories and relations rather than language-specific vocabulary or morphology, it appears likely that the rules could be ported to other languages with similar base parser support.

It is an added advantage of the approach that the same set of rules appears to work for both news and literature. For the news domain, with real-life information extraction in mind, future versions could exploit external resources to link NE mentions to unique identifiers and to resolve definite noun phrase mentions that are not clear from immediate context, e.g. for politicians and officials referenced with their function rather than their name.

13 However, the portion of errors related to 'according-to' constructions, were due to an easy-to-fix rule bug, corrected post-evaluation. Thus, the under-performance for SOURCE attribution is now likely smaller than reported.

14 The same holds for literature, but given the few instances of SOURCE in our sample, this should be reevaluated on a different novel, with more indirect speech.

15 Gold annotations are typically piece and parcel of ML methodology, because they are used for training, too. For a rule-based approach, gold data would be evaluation-only, and hence relatively more expensive. It would also counteract the fast improvements and genre adaptation typical of rule-based development, because changes in e.g. category inventory or tokenization will make the (fixed) gold data difficult to use.

16 One specific problem was that quotes were sentence-split from a following quoting verb if the quote ended in a question or exclamation mark, as the latter were treated as window delimiters by the CG. This was fixed by disambiguating between "breaking" and "non-breaking" question and exclamation marks.

References

Eckhard Bick. 2014. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. In: Tony Berber Sardinha & Thelma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpor*a, pp 279-302. London/New York:Bloomsburry Academic.

Eckhard Bick and Tino Didriksen. 2015. CG-3 - Beyond Classical Constraint Grammar. In: Beáta Megyesi: *Proceedings of NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. pp. 31-39. Linköping: LiU Electronic Press.

Eckhard Bick. 2022. PFN-PT: A Framenet Annotator for Portuguese. In Heliana Mello & Fernanda Farinelli (eds.), *The computational treatment of Brazilian Portuguese*. Domínios de Linguagem. [S. l.], v. 16, n. 4, pp. 1401-1435.

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam, Gintarė Grigonytė & Sofia Gustafson Capková. Identifying Speakers and Addressees in Dialogues Extracted from Literary Fiction. In *Proceedings of LREC 2018.* ELRA. pp- 817-824

David K. Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. ACL.

David K. Elson and Kathleen McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1),. pp. 1013-1019. https://doi.org/10.1609/aaai.v24i1.7720

Hua He, Denilson Barbosa and Grzegorz Kondrak. 2013. Identification of Speakers in Novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* pp. 1312-1320.

Christian Janze and Marten Risius. 2017. Automatic Detection of Fake News on Social Media Platforms. In *Proceedings of the 21st Pacific Asia Conference on Information Systems (PACIS).*

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A Sequence Labelling Approach to Quote Attribution. In *Proceedings of EMNLP 2012.* ACL. pp- 790-799

Nuno Mamede and Pedro Chaleira. 2004. Character identification in children stories. In *Advances in natural language processing*, pp. 82–90. Springer.

Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires and Rebeca Schumacher. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. In Vládia Pinheiro et al. (eds.) *Proceedings of PROPOR 2022* (Fortaleza, Brazil). pp. 413-419. Springer

Luis Sarmento and Sergio Nunes. 2009. Automatic extraction of quotes and topics from news feeds. In *DSIE'09 - 4th Doctoral Symposium on Informatics Engineering* (Porto, Portugal, 5-6 February 2009).

Hardik Vala, Stefan Dimitrov, Davud Jurgens, Andrew Piper, and Derek Ruths. (2016). Annotating characters in literary corpora: A scheme, the Charles tool, and an annotated novel. In Nicoletta Calzolari et al. (eds), *Proceedings of LREC 2016.* ELRA → extraction of character networks

Chak Yan Yeung and John Lee. (2017). Identifying speakers and listeners of quoted speech in literary works. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 2: Short Papers). pp. 325–329. Asian Federation of Natural Language

Yuanchi Zhang and Yang Liu. 2022. DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 6959–6966. ELRA