

Corref-PT:A Semi-Automatic Annotated Portuguese Coreference Corpus

Renata Vieira¹, Amália Mendes², Paulo Quaresma³ Evandro Fonseca¹, Sandra Collovini¹ and Sandra Antunes²

`renata.vieira@pucrs.br, amaliamendes@letras.ulisboa.pt, pq@di.uevora.pt,
evandro.fonseca@acad.pucrs.br, sandra.abreu@acad.pucrs.br,
sandra.antunes@gmail.com`

¹Pontifical Catholic University of Rio Grande do Sul

²University of Lisbon, Center of Linguistics

³University of Évora

Abstract. This paper describes the Portuguese coreference corpus Corref-PT, annotated semi-automatically using the coreference annotation tool CORP, and manually revised with the editing tool CorrefVisual. It includes a total of 182 texts, mostly news (corpus CSTNews, corpus LE-PAROLE, FAPESP magazine) but also articles from Wikipedia. The result is a corpus that includes a total of 3898 reference chains. We present the coreference annotation tool CORP, which was built on the basis of deterministic rules, and the editor CorrefVisual used for manual revision. We report on the annotation agreement and on the feedback provided by the annotators regarding the editor and the complexity of the task. Examples of technical and linguistic issues encountered during the annotation are given and the pros and cons of such approach for corpus construction are discussed. Our motivation was to use of a semi-automatic approach to increase the set of available resources for coreference resolution applications for Portuguese.

1 Introduction

Coreference resolution basically consists of finding different references to a same entity in a text, as in the example: *France resists as the only country of the European Union that doesn't allow gene patenting*. The noun phrases [the only country of the European Union that doesn't allow gene patenting] and [France] are considered coreferent. In other words, they belong to the same coreference chain. Coreference resolution may provide important input for other NLP tasks. It usually requires previous annotated data in order to be studied and also to train or evaluate proposed systems.

In this paper we present an effort to increase the number of Portuguese annotated data for coreference, by using a coreference resolution system as an intermediate step of the annotation. One of the motivations for creating an annotated corpus semi-automatically is to increase the number of annotated coreference data for Portuguese. Instead of creating such annotated corpus from scratch, we adopted a different approach,

and we proposed the edition of coreference chains produced by a coreference resolution tool. We then discuss the inherent difficulty of the task, highlighting the pros and cons of adopting this approach.

2 Previous Coreference Annotated Corpora

Coreference resolution is very important in understanding texts; thus, it is a crucial step in many high-level natural processing tasks, ranging from information extraction to text summarization or machine translation [29]. In general, the evaluation of systems devoted to this task depends on reference corpora (golden standards). There are, for example, English coreference annotated corpora that have been used in coreference resolution tracks such as SemEval, ACE and CoNLL [3,28,23,8,22,21]. SemEval (Evaluation Exercises on Semantic Evaluation) includes, among others tasks, the Coreference Resolution task [23], considering multiple languages (Catalan, Dutch, English, German, Italian and Spanish). This task involved automatically detecting full coreference chains, composed of named entities, pronouns, and full noun phrases. The datasets used in SemEval task were extracted from five corpora: 1) the AnCora corpora [24] for Catalan and Spanish; 2) the KNACK-2002 corpus [14] for Dutch; 3) the OntoNotes Release 2.0 corpus for English [22]; 4) the TurBa-D/Z corpus [13] for German; and 5) the LiveMemories corpus [25] for Italian. CoNLL-2011 Coreference Task included a closed (limited to using the distributed resources) and an open track (unrestricted use of external resources). The task was to automatically identify mentions of entities and events in texts and to link the coreferring mentions together to form coreference chains. For this, the participants could use information from other structural layers including parsing, semantic roles, word sense and named entities. It was based on OntoNotes 4.0 [21].

The OntoNotes is a large-scale corpus of general anaphoric coreference not restricted to noun phrases or to a specified set of entity types [22,21]. In addition to coreference, the corpus provides other layers of annotation: syntactic trees; propositions structures of verbs; partial verb and noun word senses; and 18 named entity types. OntoNotes is a multi-lingual resource with annotations available in three languages: English, Chinese and Arabic.

OntoNotes corpus is of crucial important for data modeling of linguistically easier cases of coreference. Complex cases are being investigated more recently, one of the main reasons for this is the lack of appropriated datasets [28]. The ARRAU dataset is a multi-domain corpus with large-scale annotations of various linguistic phenomena related to anaphora. A second release of the ARRAU is presented in [28], and the authors not only focused on increasing the number of documents, but also invested a considerable effort into improving the data quality. The data is manually labeled for tasks such as coreference resolution, bridging, mention detection, referentiality and genericity. The documents were annotated for anaphoric information, using the MMAX (Multi-Modal Annotation in XML) tool, which is specific for corpus annotation, with main focus in the annotation of coreference [20]. The annotation followed the ARRAU

guidelines, which focused on a more detailed representation of linguistic phenomena related to anaphoric and coreference. The authors present the main differences between ARRAU and two coreference corpus: ACE and OntoNotes. The difference between these corpora stands out, ARRAU considers different types of noun phrases, including markables that do not participate in coreference chains (singletons and non-referentials). Also, this corpus combines coreference with bridging, and for the third release of ARRAU, the authors plan to focus on bridging.

One of the difficulties for the creation of annotated corpora is the availability of specialists for this task. An alternative is crowd-sourcing approach, which uses a non-expert crowd to annotate text, driven by cost, speed and scalability [15]. In [3] Phrase Detectives, an interactive online game for creating annotated anaphoric coreference corpora using GWAP (game-with-a-purpose) approach is presented. The Phrase Detectives Corpus 1.0 contains 45 documents from Wikipedia articles and narrative text, with 6,452 markables.

3 Portuguese Coreference Corpora

HAREM is one of the first joint evaluation efforts for Portuguese (Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas) [26]. This contest had the purpose of studying expressions regarding proper names (named/mentioned entities). The Second HAREM took place in 2008 and it included the task of identifying the semantic relations between entities [11]. This task, although maintaining the restriction to named entities, was also a source of coreference annotation, since the authors proposed the detection of relations between named entities, including coreference, represented by the relation of *Identity* (entities with the same referent, defined to all the categories and whose instances must had the same category).

Another related Portuguese corpus is the Summ-it corpus [4,1]. It is a corpus gathering annotations of various linguistic levels, including coreference, but also morphological, syntactic and rhetorical relations. Summ-it has a total of 560 coreference chains with an average of 3 noun phrases for chain, where the largest chain has 16 members (noun phrases). Recently, a new version of the Summ-it corpus was enriched with two layers: named entities and the relations that hold between these entities [6,5]. This version is called Summ-it++¹ and is described in [1]. Garcias's corpus [12] also contains coreference annotation, but only for Person entities. It is a multilingual corpus² including Portuguese, Galician and Spanish. It also adopts the SemEval format.

4 Corref-PT

Corref-PT was annotated semi-automatically. A coreference resolution tool (CORP) [10] was applied to the texts and an editing tool [27] was

¹ http://www.inf.pucrs.br/linatural/summit_plus_plus.html

² <http://gramatica.usc.es/marcos/coling14.tar.bz>

provided to the annotators to correct its output, together with an annotation manual explaining the task, the concept of coreference and some examples, and also the editor manual.

Corref-PT is composed by texts from the CSTNews corpus [18]; from the Parole corpus (miscellaneous texts from books, magazines, journalistic, among others) [7]; Wikipedia articles, selected randomly; and also a few scientific texts from Fapesp Magazine³. Metrics about number of texts, tokens, mentions, coreferential mentions, coreference chains, chains sizes, and the distribution per text type are shown in Table 1.

Corpus	Texts	Tokens	Mentions	Coreferent Mentions	Coreference Chains	Largest Chain	Avg. Chain Size
CST-News	137	54445	14680	6797	1906	25	3.6
Le-Parole	12	21607	5773	2202	573	38	3.8
Wikipedia	30	44153	12049	4973	1308	53	3.8
Fapesp Magazine	3	3535	1012	496	111	33	4.5
Total	182	123740	33514	14468	3898	53	3.7

Table 1. Corref-PT - Corpus Metrics

The corpus was annotated as an effort made by seven teams, with a total of twenty-one Portuguese native speakers annotators, varying among students and professors in the area of computational linguists.

Annotation Tools CORP is a coreference resolution tool for Portuguese [9] which was built on the basis of deterministic rules, in the line with previous tools proposed for English [17,16].

CorrefVisual is a tool developed in order to allow the edition of coreference chains annotated with CORP. It provides a graphical interface for visualizing and replacing NPs in other coreference chains. It also allows the editing of noun phrases, creation and deletion of chains and persistency of changes.

The final annotated corpus is available in CORP’s XML and SemEval format [23] used by other well known coreference corpora, such as Ontonotes [21], Summ-it++ [1] and Garcia’s corpus [12]. Corref-PT is available for download⁴. In Table 2, we show the SemEval format. It is available in a single file, containing all texts. Each text document is contained within a “#begin document ID” line and another line containing only “#end document”. Each sentence’s information is organized vertically, with one token per line, and a blank line after the last token of each sentence. The information associated with each token is available in columns (separated by a tab character - “\t”). The annotation columns contain, respectively: Token’s ID in sentence; the word or multiword itself; lemma; each word’s Part-of-speech tagging; features (gender and number); Head, denoting if the word is a head word in the NP (if so, this field receives ’0’) and

³ <http://revistapesquisa.fapesp.br/>

⁴ [http://\[blind\]](http://[blind])

ID	Token	Lemma	POS	Feat	Head	Corref
1	Segundo	segundo	prp			
2	informações	informar	n	$\bar{F}=P$	$\bar{0}$	$\bar{-}$
3	de	de	prp		$\bar{-}$	$\bar{-}$
4	a	o	art	$\bar{F}=S$	$\bar{-}$	$\bar{-}$
5	assessoria	assessoria	n	$\bar{F}=S$	$\bar{0}$	$\bar{-}$
6	de	de	prp		$\bar{-}$	$\bar{-}$
7	o	o	art	$\bar{M}=S$	$\bar{-}$	$\bar{(2)}$
8	apresentador	apresentador	n	$\bar{M}=S$	$\bar{0}$	$\bar{2)}$
9	,		,			
10	ele	ele	pron-pers	$\bar{M}=3S=NOM$	$\bar{0}$	$\bar{(2)}$
11	não	não	adv		$\bar{-}$	$\bar{-}$
12	poderia	poder	v-fin	$\bar{COND}=3S$	$\bar{-}$	$\bar{-}$
13	comparecer	comparecer	v-inf		$\bar{-}$	$\bar{-}$
14	a	a	prp		$\bar{-}$	$\bar{-}$
15	o	o	art	$\bar{M}=S$	$\bar{-}$	$\bar{-}$
16	Deic		prop	$\bar{M}=S$	$\bar{0}$	$\bar{-}$
..	...					

Table 2. Corref-PT - Semeval format

coreference information, where each coreferent noun phrase starts with “(”, followed by the chain’s ID. Note that the “)” just occurs in the last NP token. Basically, coreferent NPs receives the same chain ID.

5 Annotation agreement and task evaluation

We measured annotation agreement for the revised chains on the basis of Kappa statistics. Kappa is usually used to measure concordance among canonical elements. For the coreference task, we need to calculate the agreement of complex elements: coreference chains. Basically, a coreference chain may have two or more noun phrases. Thus, for the correct calculation of agreement, we need to transform these chains into items that may be analysed as a category. In that way, the resulting Kappa was 0.51.

The annotators evaluated the task regarding a few issues inquired through a survey on Google Forms. Fifteen of the 21 participants sent their answers. They were asked about their confidence level in the annotation, whether the previous automatic annotation was helpful for the task and about the necessity of noun phrase edition for the task (considering that noun phrase identification was made automatically by a parser). We can see in Figure 1 that few annotators had high confidence in their annotation. Most participants were not sure about this issue.

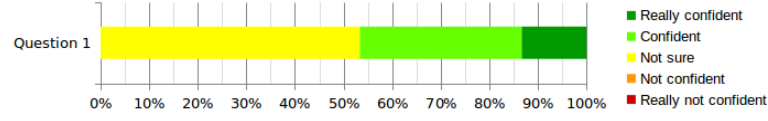


Fig. 1. Question 1 - confidence level

Regarding previous annotation, most participants were ambivalent whether this helps or not the process, but a greater number thought it was helpful.

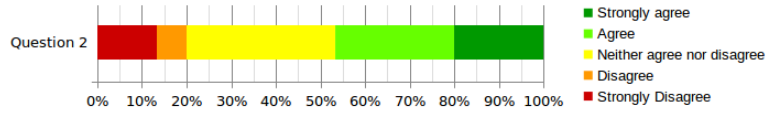


Fig. 2. Question 2 - usefulness of previous annotation

Regarding noun phrase edition, 60% of participants strongly agreed that it is indispensable for the annotation task. That is indeed a crucial pre-processing requirement for building the chains, that the references are correctly identified. The main problem here was that the task was in fact mostly fixed regarding mention detection, and it was based on the parser's NP chunks. Suggestions given by the annotators were mostly related to CorrefVisual's usability - one major problem was related to noun phrase edition. That was very difficult to handle by the annotators, since the mention detection is required for identifying coreference chains correctly, but the tool was not primarily meant for that.

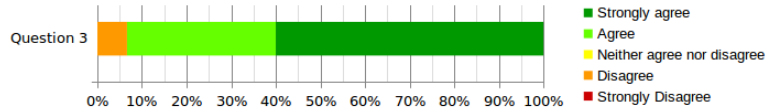


Fig. 3. Question 3 - noun phrase edition required

6 Linguistic Issues in annotating Corref-PT

Although we assumed that pre-identified coreference chains would speed the annotation and that correcting chains would be straightforward, in the process we faced some important issues that are related to the complexity of coreference and that had to be taken into account. In this section we list and exemplify some of them.

Tokenization Some of the problems that we experienced in the identification of the nominal phrases are due to the preprocessing of the texts, namely tokenization and the grouping of tokens as named entities. For instance, titles were grouped together with the first token of the following sentence as one named entity, as in *Hospital de Castelo Branco O Hospital Distrital* and also *Linha de o Corgo Está*. Some named entities that were automatically identified contained more lexical material than required. For instance, in the sequence *Extinção do Gabinete de Planeamento e de Coordenação do Combate à Droga*, we couldn't eliminate the first two tokens *Extinção do* and had to select the whole sequence.

Also, the post-verbal accusative or dative clitic is usually treated as an independent token that can be identified separately as part of a reference chain (as in (1-a)), but in some cases, the tokenization didn't separate verb and clitic, and both tokens had to be selected as part of the chain, as illustrated in (1-b).

- (1) a. o homem não devia obedecer a a natureza, mas sim vencê -la (dn88218). (man shouldn't obey nature, but instead defeat it)
- b. desafiar e vencer a natureza *contrariando-a* (dn88218) (to challenge and defeat nature by going against it)

Near Identity In many contexts, it is difficult to establish with absolute certainty that two NPs are coreferent [19]. For instance, in the initial training phase, we decided to treat as coreferent the NP *os primeiros cães domésticos* (the first domestic dogs) and the NP *cães domésticos* (domestic dogs) that occurs in the larger NP *fósseis de cães domésticos* (fossils of domestic dogs). It can be debated whether the two are coreferent: although the second NP refers to fossils which are consequently old, it might not refer exactly to the fossils of the first domestic dogs. The two NPs were treated as coreferent to avoid dividing the data into many reference chains. This brings about the question of Near-Identity, which will be treated according to the level of granularity and the general goals of the annotation. Another example from the training phase is the NP *diversidade genética* (genetic diversity) and the noun *diferenças* (differences) that were treated as coreferent because the differences were interpreted contextually as genetic differences. This reference chain was already automatically pre-identified in the CorrefVisual editor.

In another case, we annotated two near coreferent NPs as part of different reference chains. In example (2), *ser humano* (human being) and *a humanidade* (the humanity) are treated as non coreferent due to the explicit mention of their different scope in the context, although they appear to be used in the rest of the text as synonyms.

- (2) o *ser humano* e, por extensão, *a humanidade* (dn81201) (the human being and, by extension, the humanity)

Nominal phrases may be lexically distinct but very similar in terms of their reference, as in examples (3-a) and (3-b), where *estações de recolha*

(collection stations) and *estações meteorológicas* (weather stations) refer to the same entity and *à escala mundial* (world wide) and *o planeta* (the planet) refer to the same scope of the network. We considered the two NPs as part of the same reference chain.

- (3) a. a rede de estações de recolha a a escala mundial (pu92214)
(the network of collection stations world wide)
- b. a rede de estações meteorológicas de o planeta (pu92214)
(the network of weather stations of the planet)

The following case raises even more questions about what can be considered as coreferent. In example (4-a), the nouns *modelos matemáticos* (mathematical models) and *computadores* (computers) are modified by an adjectival phrase with very specific lexical material. In example (4-b), the same nouns are modified by the less informative adjective *melhores* (better). Could we consider that the better models and computers that the second example mentions are the ones capable of modeling the weather and the meteorological conditions? We consider that it is indeed the case, based on the context, and annotated as coreferent.

- (4) a. Há alguns anos , faltavam estações de observação e não havia modelos matemáticos nem computadores *capazes de modelizar o clima e as condições meteorológicas* (pu92214)
(Some years ago, there was a lack of observation stations and there were no mathematical models nor computers capable of modeling the climate and the weather conditions.)
- b. Considera que os principais problemas consistem em a falta de dados de base , de *melhores* modelos matemáticos , de melhores computadores ...? (Do you think that the main problems are the lack of base data, of better mathematical models, of better computers ...? (pu92214)

Modality The presence of epistemic modal markers (i.e, lexical markers that express values such as uncertainty, possibility) raises issues in terms of the annotation of coreference [2]. For instance, in example (5) the writer presents the view of someone else (the employer) about work organization (the NPs are in *italic*, while the modal marker is underlined). This means that the coreference between the two underlined NPs *aquela organização do trabalho* (that work organization) and *um dos primeiros passos num caminho que tende a levar longe* (one of the first steps in a road that tends to lead far ahead) stands in the perspective of the employer, but obviously not of the author that clearly disagrees with this perspective: *o sacrossanto poder do patrão na empresa* (the sacrosanct power of the employer in the company). Coreference would then be dependent on the view of each source in the text.

- (5) *Aquela organização de o trabalho* , a o conferir poderes a os trabalhadores sobre as condições de o trabalho , é vista por o sacrossanto poder de o patrão em a empresa como um de os primeiros passos em um caminho que tende a levar longe . de aí resistências (That work organization, by giving power to the

workers over the work conditions, is seen by the sacrosanct power of the employer in the company as one of the first steps in a road that tends to lead far ahead. Thus the resistances.) (Texto11.txt)

Embedded and coordinated NPs One of the first issues faced during the annotation was whether an embedded NP could be part of another reference chain. For instance, the NP illustrated in (6) would refer to the two reference chains indicated in the example. We treated embedded NPs as part of other reference chains, when applicable.

- (6) o estudo das sociedades primitivas (the study of the primitive societies) (dn81201)
 reference chain 1: o estudo das sociedades primitivas (the study of the primitive societies)
 reference chain 2: as sociedades primitivas (the primitive societies)

Length of the NPs One of the recurrent questions in our annotation was the amount of information to include in the NP of a reference chain. Restrictive relatives were included in the NP because they contribute to establish its reference. There was, however, some hesitation in the annotations of relative clauses.

Another type of context that leads to some discussion about the length of the NPs is illustrated in (3-a). Here, the issue was whether *à escala mundial* (word wide) should be included in the NP or not. The fact that a similar NP, illustrated in (3-b), occurred in the text lead to the selection of the whole sequence. In contexts such as (7), the parenthetical segment (in italic) wasn't included because it is not essential to the reference of the NP (just as a non restrictive relative would also be left out). The length of the NPs is obviously one of the sources of lack of agreement among the annotators.

- (7) a definição e concretização de uma estrutura associativa empresarial sólida e eficaz, *essencialmente de base regional* (the definition and implementation of a corporate associative structure strong and effective, essentially regionally based) (dn81625)

Implicit content The antecedent of an anaphoric element can be implicit in the context. For example, the meetings referred in (8-b) are the meetings of the Commission referred in (8-a), so this entity is implicit in the NP: *essas reuniões [da Comissão Intergovernamental de Negociação]* (those meetings [of the Intergovernmental Commission of Negotiation]). The question is whether the NP in (8-b) should be considered as part of the reference chain of the entity *Comissão Intergovernamental de Negociação* (Intergovernmental Commission of Negotiation).

- (8) a. Uma Comissão Intergovernamental de Negociação (pu92214)
 (An Intergovernmental Commission of Negotiation)
 b. Essas reuniões (Those meetings)

7 Conclusion

In this paper, we presented a coreference corpus for Portuguese with nearly 4000 chains. Considering Summ-it++, a previous available resource of the kind, with around 500 chains, we now have a coreference annotated corpus with 8 times as many chains. The resource is available both in the SemEval format and in XML. The annotated corpus can be visualized in the CorrefVisual tool⁵.

This task comprises a high level of difficulty, involving high linguistic knowledge. The use of pre-identified NPs was one of the main problems: when they are incorrect there is an overhead to the process having, as a consequence, the need to undo the initial annotation manually.

For the next steps, we have to improve questions regarding automatic mention detection, which seems to be a major pre-processing issue for this task, and similarly we have also to improve the ways for their manual editing, if we consider further semi automatic annotation tasks.

Although moderate agreement was achieved for the annotation editing, as future work, a revision of this annotation should be done to improve annotation quality.

In spite of all the described problems we believe this task allowed us to better understand the complexity and the details of coreference annotation and to contribute to the creation of a reference annotated corpus for the Portuguese language.

References

1. A. Antonitsch, A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini. Summ-it++: an enriched version of the summ-it corpus. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2047–2051, Paris, France, 2016. European Language Resources Association (ELRA).
2. G. Bouma, W. Daelemans, I. Hendrickx, V. Hoste, and A. Mineur. The corea-project, manual for the annotation of coreference in dutch texts. Technical report, 2007.
3. J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2039–2046, Paris, France, 2016. European Language Resources Association (ELRA).
4. S. Collovini, T. I. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, and R. Vieira. Summ-it: Um corpus anotado com informações discursivas

⁵ <http://www.inf.pucri.br/linatural/wordpress/index.php/recursos-e-ferramentas/correfvisual/>

- visando a sumarização automática. In *Proceedings of V Workshop em Tecnologia da Informação e da Linguagem Humana*, Rio de Janeiro, RJ, Brasil, pages 1605–1614, 2007.
5. S. Collovini de Abreu and R. Vieira. Relp: Portuguese open relation extraction. *Knowledge Organization*, 44(3):163–177, 2017.
 6. D. O. F. do Amaral and R. Vieira. NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamatica*, 6(1):41–49, 2014.
 7. M. F. B. do Nascimento, A. Mendes, and L. Pereira. Providing on-line access to portuguese language resources: Corpora and lexicons. In *Proceedings of the International Conference on Language Resources and Evaluation*, Portugal, 2004.
 8. G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The automatic content extraction (ace) program: Tasks, data, and evaluation. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation – LREC 2004*, pages 837–840, Lisboa, 2004.
 9. E. B. Fonseca, V. Sesti, A. Antonitsch, A. A. Vanin, and R. Vieira. Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. *Linguamatica*, 9(1):3–18, 2017.
 10. E. B. Fonseca, R. Vieira, and A. Vanin. Corp: Coreference resolution for portuguese. In *12th International Conference on the Computational Processing of Portuguese, Demo Session (PROPOR)*, 2016.
 11. C. Freitas, C. Mota, D. Santos, H. G. Oliveira, and P. Carvalho. Second HAREM: advancing the state of the art of named entity recognition in portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valletta, Malta*, 2010.
 12. M. Garcia and P. Gamallo. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference - LREC*, pages 3229–3233, 2014.
 13. E. W. Hinrichs, S. Kübler, and K. Naumann. A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 13–20, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
 14. V. Hoste and G. De Pauw. Knack-2002: a richly annotated corpus of dutch written text. In *Proceedings of The Fifth international conference on Language Resources and Evaluation*, pages 1432–1437, Genoa, Italy, 2006. European Language Resources Association, European Language Resources Association.
 15. J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition, 2008.
 16. H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. volume 39, pages 885–916. *Computational Linguistics - MIT Press*, 2013.

17. H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics, 2011.
18. E. G. Maziero, M. L. del Rosario Castro Jorge, and T. A. S. Pardo. Identifying multidocument relations. In *Natural Language Processing and Cognitive Science, Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2010, In conjunction with ICEIS 2010, Funchal, Madeira, Portugal, June 2010*, pages 60–69, 2010.
19. A. Mendes. Organização textual e articulação de orações. pages 1691–1755. Gramática do Português, vol. II. Lisboa: Fundação Calouste Gulbenkian, 2013.
20. C. Müller and M. Strube. Mmax: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Adaptive Text Extraction and Mining - IJCAI 2001*, Seattle, Washington, 2001.
21. S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics, 2011.
22. S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. Ontonotes: A unified relational semantic representation. In *Proceedings of the International Conference on Semantic Computing, ICSC '07*, pages 517–526, Washington, DC, USA, 2007. IEEE Computer Society.
23. M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics, 2010.
24. M. Recasens and M. A. Martí. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44(4):315–345, 2010.
25. K. J. Rodríguez, F. Delogu, Y. Versley, E. Stemle, and M. Poesio. Anaphoric annotation of wikipedia and blogs in the live memories corpus. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation - LREC*. European Language Resources Association, 2010.
26. D. Santos, N. Cardoso, N. Seco, and R. Vilela. Breve introdução ao harem. *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*, Linguatca, 2007.
27. M. d. O. Tubino and M. M. S. Silva. Visualização, manipulação e refinamento de correferência em língua portuguesa. Trabalho de conclusão de curso, Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

28. O. Uryupina, R. Artstein, A. Bristot, F. Cavicchio, K. Rodriguez, and M. Poesio. ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2058–2062, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).
29. K. van Deemter and R. Kibble. What is coreference, and what should coreference annotation be? In *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp '99, pages 90–96, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.