Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 163-170 recibido 07-05-12 revisado 30-05-12 aceptado 05-06-12

# A First Approach to the Automatic Detection of Zero Subjects and Impersonal Constructions in Portuguese

## *Primera aproximación para la detección automática de pronombres cero y construcciones impersonales en portugués*

**Luz Rello,**[1,2] **Gabriela Ferraro**[2]
Web Research Group[1] & TALN, Centre for Autonomous Systems and Neuro-Robotics[2]
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra
luzrello@acm.org, gabriela.ferraro@upf.edu

**Iria Gayo**
Grupo de Gramática del Español
Departamento de Lengua Española
Universidad de Santiago de Compostela
iria.delrio@usc.es

**Resumen:** Este trabajo constituye un primer intento de abordar la detección automática de sujetos elididos y de construcciones impersonales en portugués de Brasil, una tarea que no nos consta que se haya llevado a cabo previamente en esta lengua. Para ello, creamos un corpus que contiene más de 5.600 casos anotados con las clases que deben identificarse: sujetos explícitos, sujetos o pronombres omitidos y construcciones impersonales. Estos casos se clasificaron mediante aprendizaje automático basado en rasgos lingüísticamente motivados. Los resultados obtenidos son modestos, aunque prometedores, y proporcionan una orientación para futuros trabajos en este ámbito.
**Palabras clave:** elipsis de sujeto, construcción impersonal, pronombre cero, sujeto nulo, aprendizaje automático.

**Abstract:** In this paper we present a first approximation to the automatic detection of zero subjects and impersonal constructions in Brazilian Portuguese. To the best of our knowledge, this is the first attempt of approaching such task using machine learning in Portuguese. We compiled a corpus containing more than 5,600 instances annotated with the classes to be identified: explicit subjects, zero subjects or pronouns and impersonal constructions. We applied machine learning using linguistically motivated features to classify the instances. The results are modest but promising and provide guidance for future work.
**Keywords:** subject ellipsis, impersonal construction, zero pronoun, null subject, machine learning

## *1 Introduction*

Portuguese is a pro-drop language (Chomsky, 1981) meaning that subject ellipsis is a highly recurring phenomenon. For instance, in our Brazilian Portuguese corpus, 21% of the subjects are elided.

Numerous natural language processing (NLP) tasks require the identification of subject ellipsis. For instance, the identification of zero pronouns is necessary for zero anaphora resolution (Mitkov, 2002; Chaves and Rino, 2008) and for co-reference resolution (Ng and Cardie, 2002). Also, it has been found to be helpful in a number of NLP applications such as machine translation (Peral and Ferrández, 2000) or text categorization (Yeh and Chen, 2003). However, to the best of our knowledge, the recognition of zero pronouns and non-referential impersonal constructions has not yet been addressed in Portuguese. The goal of this paper is to present a first approach of a method to accomplish this task.

The remainder of the paper is organized as follows. Section 2 provides a literature review while Section 3 describes the classes of Brazilian Portuguese subjects. Section 4 presents the creation and the annotation of the corpus and Section 5 discusses the features used and the preliminary results of the

ISSN 1135-5948 

machine learning (ML) method. Finally, in Section 6, conclusions are drawn and plans for future work are discussed.

## 2 Related Work

The difficulty in detecting missing subjects and non-referential pronouns has been acknowledged since the first studies on computational treatment of anaphora (Bergsma, Lin, and Goebel, 2008; Mitkov, 2010).

Identification of zero pronouns and non-referential pronouns is a crucial step in coreference and anaphora resolution systems because the identification of zero anaphors first requires that they be distinguished from non-referential impersonal constructions (Mitkov, 2010). To approach these tasks we find both ruled-based and ML-based approaches. While machine learning methods are known to perform better than rule-based techniques for identifying non-referential expressions (Boyd, Gegg-Harrison, and Byron, 2005), the most favorable approach for detecting zero subjects is under debate (Ferrández and Peral, 2000; Rello, Baeza-Yates, and Mitkov, 2012).

Among the various computational methods for anaphora resolution in Portuguese, to the best of our knowledge, there is only a rule-based system for pronoun resolution which considers specifically zero subjects (Bick, 2010). This method reaches a f-measure of 70.6% (just 74 annotated zero pronouns) for the resolution of zero anaphora, which is the most difficult to approach according to the author (Bick, 2010). However, we found no evaluation for the task of the identification of zero subjects. The rest of the approaches for anaphora resolution in Portuguese do not consider specifically zero subjects and non-referential impersonal constructions. In (Chaves and Rino, 2008) Mitkov's algorithm is used for the resolution of third person pronouns in texts written in Brazilian Portuguese while in (Rocha, 1999), the classification used does not include zero subjects or empty categories.

Apart from Portuguese, there are other pro-drop languages on which related work about zero pronoun identification has been carried out, such as Japanese (Yoshimoto, 1988) – rule-based approach–, Spanish (Ferrández and Peral, 2000; Rello and Illisei, 2009) –rule-based approach–, Chinese (Zhao and Ng, 2007) –ML based– and Romanian –ML based– (Mihaila, Ilisei, and Inkpen, 2011). For the identification of explicit non-referential constructions we found ML based studies in English (Evans, 2001; Bergsma and Yarowsky, 2011), Spanish (Rello, Baeza-Yates, and Mitkov, 2012) and French (Danlos, 2005).

Our approach is mainly inspired by the ML-based methods and combines features which were useful in (Evans, 2001; Zhao and Ng, 2007; Rello, Baeza-Yates, and Mitkov, 2012).

## 3 Classification

Subject ellipsis is the omission of the subject in a sentence. We consider not only missing referential subject (zero subject) as manifestation of ellipsis, but also non-referential impersonal constructions.

Literature related to linguistic theory (Cunha and Cintra, 1984; Mateus et al., 2003; Rello and Gayo, 2011) has served as a basis for establishing the linguistically motivated classes and the annotation criteria of this work. The features into which Portuguese subjects were distinguished are: [± elliptic] and [± referential] subjects. These two features result in a ternary classification:

- Explicit subjects: [− elliptic, + referential].
- Zero subjects: [+ elliptic, + referential].
- Impersonals: [− elliptic, − referential].

In the examples explicit subjects are presented in italics, zero pronouns are marked by the symbol Ø and impersonal constructions are not explicitly indicated. In the English translations the subjects which are elided in Portuguese are marked with parenthesis and italics.

### 3.1 Explicit Subjects

Explicit Subjects are realized usually by a nominal group: noun, pronoun, noun phrase (a), free relatives, semi-free relatives or substantival adjectives (Cunha and Cintra, 1984). The syntactic positions of subjects can be pre-verbal or post-verbal. The occurrence of post-verbal subjects is not restricted by any conditions in Portuguese (Mateus et al., 2003). Projections of non-nominal categories such as clauses containing an infinitive or a conjugated verb, interrogative indirect clauses, or indirect exclamative clauses,

can function as subjects (Cunha and Cintra, 1984).

(a) *Este Decreto* dispõe sobre o exercício das funções de regulação, supervisão e avaliação de instituições de educação superior [...].[1]
*This Ordinance* disposes about the exercise of the functions of regulation, supervision and evaluation of institutions of superior education [...].

### 3.1.1 Zero Pronouns

An elliptic subject (b) is the result of a nominal ellipsis, where a realized lexical element –elliptic subject– which is needed for the interpretation of the meaning and the structure of the sentence, is omitted since it can be retrieved from its context.

(b) Ø Dispõe sobre as sanções aplicáveis aos agentes públicos nos casos de enriquecimento ilícito no exercício de mandato, cargo emprego [...].
*(It)* disposes about the applicable sanctions to the public agents in the cases of illicit enrichment in the mandate exercise, position used [...].

In Portuguese, the noun head can be omitted (Clara, 2008) when the subject of which it is a part fulfills some structural requirements and a definite article occurs, such as *os* in (c).

(c) Em o período do estudo, foram analisadas 7.956 ligações, sendo os usuários de drogas os Ø que mais procuraram atendimento, com 2.600 ligações.
During the study period, 7,956 phone calls were analysed, being the drug users *(those that)* looked for more attention, with 2,600 phone calls.

### 3.1.2 Impersonals

Impersonal constructions are both non-referential and elliptic (Cunha and Cintra, 1984; Mateus et al., 2003). The appearance of clauses containing zero pronouns is similar to impersonal constructions. This category is composed of impersonal constructions which are formed by impersonal verbs (d) and reflex impersonal clauses, impersonal clauses with *"se"* (e).

(d) Ainda não há consenso em relação a melhor sistemática a ser empregada para apresentação de um instrumento com equivalência transcultural.
*(There)* is still no consensus in relation to the best systematic to be used for presentation of an instrument with transcultural equivalence.

(e) Optou-se por uma abordagem qualitativa.
*(It)* was chosen a qualitative approach.

## *4 Corpus*

The training data used in the learning process was obtained from a corpus created specifically for this work, the Explicit Subjects, Zero-Pronouns and Impersonal Constructions corpus (ESZIC) (Rello and Gayo, 2012).[2]

The corpus is composed of 17 documents, originally written in Brazilian Portuguese and belonging to two genres: legal and health.

The legal texts are extracted from the: Civil Code of United States of Brazil (until third book, title II), Brazilian Penal Code (until title VIII, chapter VII), Brazilian Constitution of 1988 (until title III, chapter V), Law of Administrative Dishonesty (whole text), Antitrust Law (until chapter III, article VII), Law no. 9,637 (whole text), Law no. 12,232 (whole text); and Decree no. 5,773 (until chapter II, section II).

The health texts are psychiatric papers taken from the digital journal of psychiatry: *Revista de Psiquiatria do Rio Grande do Sul*. All the papers were written from 2003 to 2009.[3]

The texts were parsed by *Palavras*[4] (Bick, 2000), a parser based on Constraint Grammar methodological paradigm (Karlsson, 1990; Karlsson, Voutilainen, and Anttila, 1995). *Palavras* returns morphological information (part of the speech (POS) and lemma), syntactic information (structure of constituents and their dependency values) and semantic information (semantic prototypes).

[1] All the examples provided are taken from our corpus.

[2] Publicly available at: `http://www.luzrello.com/Projects.html`.

[3] The full-text articles from *Revista de Psiquiatria do Rio Grande do Sul* are available online at: `http://www.scielo.br/scielo.php?script=sci_serial&pid=0101-8108&lng=en&nrm=iso`

[4] `http://beta.visl.sdu.dk/visl/pt/info/`

### 4.1 Corpus Annotation

The annotator was presented the sentences in which a verb or a group of verbs appear and prompted to classify the verb into one of classes: verb with an explicit subject, verb with a zero subject or verb with no subject (impersonal construction).

Our corpus contains 5,665 finite verbs, 77% have an explicit subject, 21% a zero pronoun and 2% are impersonal constructions (see Table 1). This fact is consistent with linguistic literature since some studies claim that Brazilian Portuguese is a partial pro-drop language, mainly due to the progressive decrease of zero pronouns usage (Kato and Negrão, 2000; Gayo and Rello, 2011).

| *N* **of instances** | **Legal** | **Health** | **All** |
|---|---|---|---|
| Explicit subjects | 1,891 | 2,462 | 4,353 |
| Zero subjects | 462 | 740 | 1,202 |
| Impersonals | 55 | 55 | 110 |
| **Total** | **2,408** | **3,257** | **5,665** |

Table 1: Number of instances per class.

### 4.2 Inter-annotator Agreement

To test the reliability, validity and stability of the annotations we have computed the inter-annotator agreement.

Among the possible metrics to measure inter-annotator reliability we chose Fleiss' Kappa statistical measure (Fleiss, 1971). This measure is a generalization of Scott's Pi statistic (Scott, 1955) and an extension of the Cohen's Kappa coefficient of agreement for nominal scales to measure agreement in ordinal scale data (Cohen, 1960). Whereas Cohen's Kappa works for only two raters, Fleiss' Kappa works for any number of raters giving categorical ratings, to a fixed number of items. Fleiss' Kappa determines the chance of agreement among arbitrary coders and do not treat all kinds of disagreements in the same manner.

Given the high cost of conducting inter-annotator studies, we choose a representative sample from our corpus to limit the scope of the analysis. We extracted 10% of the instances of each of the texts of the corpus covering the two genres. Two volunteer graduate students, native speakers of Portuguese, participated in the study.

There is no universally accepted interpretation of Kappa values. However, it is common practice among researchers in computational linguistics to consider 0.8 as a minimum value of acceptance (Artstein and Poesio, 2008).

Considering these factors, our results indicate that the annotation is reliable (see Table 2). There is a small number of categories but the Fleiss Kappa value is high. Therefore, our corpus can provide a reliable resource to study subject ellipsis in Portuguese.

| **Genre** | 2 Annotators | 3 Annotators |
|---|---|---|
| Legal | 0.8261 | 0.8255 |
| Health | 0.9589 | 0.8570 |

Table 2: Fleiss' Kappa coefficient for the inter-annotator agreement.

## *5 Detecting Subject Ellipsis*

In this section we present the linguistically motivated features and the results of our ML-based method.

### 5.1 Features

We extracted nine features taken from previous studies (Evans, 2001; Rello, Baeza-Yates, and Mitkov, 2012) in order to classify instances according to the three classes defined in Section 3. The values of the features were derived from information provided by *Palavras* parser. These are:

- **Subject** The presence or absence of a subject in the clause, as identified by the parser.
- **Lemma** The lemma of the finite verb, that is, its infinitive form.
- **Number** The grammatical number of the verb (singular or plural).
- **Person** The grammatical person of the verb (first, second, or third).
- **Total Noun Phrases** The number of noun phrases in the clause that precede the verb.
- **Previous Noun Phrases** The total number of noun phrases in the clause.
- **Se** A binary feature encoding the presence or absence of the Portuguese particle *"se"* in the clause.
- **Previous POS** The POS of the four tokens preceding the instance.
- **Following POS** The POS of the four tokens following the instance.

| Class | P | R | F |
|---|---|---|---|
| Explicit Subj. | 88.4% | 90.4% | 89.4% |
| Zero Subj. | 60.2% | 55.9% | 57.9% |
| Impersonals | 81.7% | 69.1% | 74.9% |
| Weighted Avg. | 82.6% | 83.0% | 82.8% |

Table 3: LAD Tree performance (83.04% accuracy for ten-fold cross validation).

## 5.2 Preliminary Results

The training data is composed of 5,665 vectors. Each vector corresponds to one finite verb extracted from the corpus and is composed by the values of the features derived from the corpus. There is a training set but no explicit test set, therefore we use cross-validation.

To determine the most accurate algorithm for our classification task, we compared the learning algorithms implemented in Weka (Witten and Frank, 2005). Firstly, the classification was executed using the default values of the algorithms using ten-fold cross-validation. Secondly, the highest performing classifiers were compared modifying the number of iterations.

The decision tree learning classifier LAD Tree (Breiman, 1984) using 10 iterations was the best performing one with an overall accuracy[5] of 83.04%. Table 3 shows the results for each class using ten-fold cross validation.

We used as baseline the output of parser *Palavras*. It is not possible to make a fair comparison because both classes, zero subjects and impersonals are not distinguished by *Palavras*, so we included them in the same class (verbs with no subject identified by the parser) to compare both systems accuracies. Our method outperforms *Palavras* for identifying explicit subjects and impersonals. Although *Palavras* presents a higher accuracy for the identification of verbs with no explicit subjects, or method distinguishes between referential and non-referential elided subjects.

A confusion matrix show us that most of the wrongly identified instances are zero pronouns (36.10%) classified as explicit subjects, given that this class is the most heterogeneous one.

[5]Accuracy is understood as the percentage of the correctly classified instances.

| Algorithm | Explicit Subject | Zero Subject | Impersonals |
|---|---|---|---|
| *Palavras* | 71.4% | 77.3% | |
| Our method | 94.43% | 55.87% | 69.09% |

Table 4: Summary of accuracy comparison with the baseline.

| Class | Explicit Subject | Zero Subject | Impersonal |
|---|---|---|---|
| Explicit Subj. | 3957 | 409 | 10 |
| Zero Subj. | 493 | 633 | 7 |
| Impersonals | 24 | 10 | 76 |

Table 5: Confusion matrix.

## *6 Conclusions and Future Work*

This is only a first approach to the automatic classification of zero pronouns and impersonal constructions in Brazilian Portuguese. However, to the best of our knowledge, it is the first time to tackle this task for this language using a ML-based approach. The results are modest but promising and provide insights for future work.

A first goal for future work is to improve our method by extending the set of features and performing feature analysis to select the best performing ones. For the creation of new features we will conduct an errors analysis of the wrongly classified instances taking into account their linguistic context in the corpus. Another future research is the extrinsic evaluation of our system by integrating our system in NLP tasks.

## *References*

Artstein, R. and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bergsma, S., D. Lin, and R. Goebel. 2008. Distributional identification of non-referential pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT-08)*, pages 10–18.

Bergsma, S. and D. Yarowsky. 2011. Nada: A robust system for non-referential pronoun detection. *Anaphora Processing and Applications*, pages 12–23.

Bick, Eckhard. 2000. *The Parsing System Palavras - Automatic Grammatical Anal-*

*ysis of Portuguese in a Constraint Grammar Famework*. Aarhus University Press, Aarhus.

Bick, Eckhard. 2010. A dependency-based approach to anaphora annotation. In *Extended Activities Proceedings, 9th International Conference on Computational Processing of the Portuguese Language (PROPOR)*.

Boyd, A., W. Gegg-Harrison, and D. Byron. 2005. Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing. 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 40–47.

Breiman, L. 1984. *Classification and regression trees*. Chapman & Hall/CRC.

Chaves, A. and L. Rino. 2008. The mitkov algorithm for anaphora resolution in portuguese. *Computational Processing of the Portuguese Language*, pages 51–60.

Chomsky, N. 1981. *Lectures on government and binding*. Mouton de Gruyter, Berlin, New York.

Clara, Daniela. 2008. *A aquisição da elipse nominal em português europeu - produção e compreensão*. Ph.D. thesis, Universidade Nova de Lisboa, Lisboa.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cunha, Celso and Lindley Cintra. 1984. *Nova Gramática do Português Contemporâneo*. Sá da Costa, Lisboa.

Danlos, L. 2005. Automatic recognition of French expletive pronoun occurrences. In Robert Dale, Kam-Fai Wong, Jiang Su, and Oi Yee Kwong, editors, *Natural language processing. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 73–78, Berlin, Heidelberg, New York. Springer. Lecture Notes in Computer Science, Vol. 3651.

Evans, R. 2001. Applying machine learning: toward an automatic classification of *it*. *Literary and Linguistic Computing*, 16(1):45–57.

Ferrández, A. and J. Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 166–172.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gayo, I. and L. Rello. 2011. El fenómeno pro-drop en portugués brasileño y español peninsular. In *III Congreso Internacional de Lingüística de Corpus (CILC 2011)*.

Karlsson, F. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 168–173. Association for Computational Linguistics.

Karlsson, F., A. Voutilainen, and A. Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Walter de Gruyter.

Kato, M. A. and E.V. Negrão. 2000. *Brazilian Portuguese and the Null Subject Parameter*. Vervuert-Iberoamericana, Frankfurt.

Mateus, Maria H., Ana M. Brito, Inês Duarte, Isabel H. Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário, and Alina Villalva. 2003. *Gramática da Língua Portuguesa*. Editorial Caminho, Lisboa, 5 edition.

Mihaila, C., I. Ilisei, and D. Inkpen. 2011. Zero pronominal anaphora resolution for the romanian language. *Research Journal on Computer Science and Computer Engineering with Applications POLIBITS*, 42.

Mitkov, R. 2002. *Anaphora resolution*. Longman, London.

Mitkov, R. 2010. Discourse processing. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The handbook of computational linguistics and natural language processing*. Wiley Blackwell, Oxford, pages 599–629.

Ng, V. and C. Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International*

*Conference on Computational Linguistics (COLING-02)*, pages 1–7.

Peral, J. and A. Ferrández. 2000. Generation of Spanish zero-pronouns into English. In D. N. Christodoulakis, editor, *Natural Language Processing - NLP 2000. Proceedings of the 2nd International Conference on Natural Language Processing (NLP-2000)*. Springer, Berlin, Heidelberg, New York, pages 252–260. Lecture Notes in Computer Science, Vol. 1835.

Rello, L., R. Baeza-Yates, and R. Mitkov. 2012. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in spanish. In *Proceedings of the European chapter of the Association for Computational Linguistics (EACL 2012)*. Association for Computational Linguistics.

Rello, L. and I. Gayo. 2011. Classification criteria for the automatic identification of subject ellipsis and impersonal constructions in Portuguese. In *Proceedings of the 25th Sociedad Española de Lingüística (SEL 2011)*.

Rello, L. and I. Gayo. 2012. A comparable Portuguese-Spanish corpus with ellipsis annotations. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

Rello, L. and I. Illisei. 2009. A rule-based approach to the identification of Spanish zero pronouns. In *Student Research Workshop. International Conference on Recent Advances in Natural Language Processing (RANLP-09)*, pages 209–214.

Rocha, M. 1999. Coreference resolution in dialogues in english and portuguese. In *Proceedings of the Workshop on Coreference and its Applications*, pages 53–60. Association for Computational Linguistics.

Scott, W.A. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*.

Witten, I. H. and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, London, 2 edition.

Yeh, C. and Y. Chen. 2003. Using zero anaphora resolution to improve text categorization. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC-03)*, pages 423–430.

Yoshimoto, K. 1988. Identifying zero pronouns in Japanese dialogue. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 779–784.

Zhao, S. and H.T. Ng. 2007. Identification and resolution of Chinese zero pronouns: a machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CNLL-07)*, pages 541–550.