A Portuguese-Spanish Corpus Annotated for Subject Realization and Referentiality

Luz Rello[1], **Iria Gayo**[2]

[1] NLP and Web Research Groups, Universitat Pompeu Fabra, Barcelona, Spain
[2] Grupo de Gramática del Español, Universidad de Santiago de Compostela, Spain
luzrello@acm.org, iria.delrio@usc.es

Abstract

This paper presents a comparable corpus of Portuguese and Spanish consisting of legal and health texts. We describe the annotation of zero subject, impersonal constructions and explicit subjects in the corpus. We annotated 12,492 examples using a scheme that distinguishes between different linguistic levels (phonology, syntax, semantics, etc.) and present a taxonomy of instances on which annotators disagree. The high level of inter-annotator agreement (83%–95%) and the performance of learning algorithms trained on the corpus show that our corpus is a reliable and useful resource.

Keywords: ellipsis, referentiality, zero subject, impersonal construction, corpus annotation, reliability study.

1. Introduction

Subject ellipsis is the omission of the subject in a clause. In pro-drop languages as Portuguese and Spanish, subject ellipsis is a recurring phenomenon. For instance, around 29% of the verbs in Spanish written language do not have an explicit subject. Also, the study of the omission of some element from the sentence or the discourse has been a challenge not only in natural language processing (NLP), but also in linguistics itself (Brucart, 1999).

Numerous NLP tasks require the identification of subject ellipsis. However, this task becomes decisive when processing pro-drop languages since subject ellipsis is a highly recurring phenomenon in these languages (Chomsky, 1981). For instance, ellipsis identification is necessary for zero anaphora resolution (Mitkov, 2002) and for coreference resolution (Ng and Cardie, 2002). In these cases, not only referential omitted subjects need to be identified but also the identification of non-referential impersonal constructions.

In this paper, we present a useful resource to improve subject ellipsis recognition (Rello et al., 2012) as well as to carry out linguistic descriptions of ellipsis (Rello and Illisei, 2009a; Gayo and Rello, 2011).

For this purpose, we created a free available comparable corpus (ESZIC)[1] for investigating subject ellipsis in Portuguese (ESZIC_pt) and Spanish (ESZIC_es). This resource is named after its annotated content "Explicit Subjects, Zero-subjects and Impersonal Constructions".

Next, we explain the singularities of the ESZIC corpus in relationship with other existing corpora. We describe the linguistic criteria which served as a basis for the design of the annotation scheme and the definition of the annotated categories in Section 3. We explain the content of the ESZIC and present a reliability study as well as the outcomes of the use of this resource in Section 4. Conclusions are drawn in Section 5.

2. Existing Corpora

There are a few reported corpora with annotated ellipsis in Portuguese and Spanish.

The Blue Book and *Lexesp* used in (Ferrández and Peral, 2000) contain together 1,599 classified verbs including 734 zero pronouns. *The Blue Book* corpus is a handbook of the International Telecommunications Union and the *Lexesp* corpus contains Spanish texts from different genres and authors, mainly taken from newspapers. AnCora-ES corpus (Recasens and Martí, 2010) includes the annotation of elliptical pronouns and is based on journalistic texts. We found 10,791 examples tagged as elliptic pronouns in subject position in AnCora-ES. Differently to *The Blue Book* and *Lexesp*, AnCora-ES contains annotated impersonal constructions (264). Finally, the Z-corpus (Rello and Illisei, 2009b) comprises legal, instructional and encyclopaedic texts, with 1,202 annotated zero subjects but impersonal constructions were not considered.

Portuguese corpora are scarcer. The ZAC corpus is the first corpus for Brazilian Portuguese annotated with subject zero anaphors (Pereira, 2009). The ZAC corpus contains annotation for the different types of subject ellipsis (1,489 annotated instances) and impersonal subjects (100 instances). Again, the genres contained in the ZAC corpus differ from ESZIC_pt. The ZAC corpus consists on a set of full and partial texts retrieved from the Web and digitalized from books, encompassing several genres, namely journalistic (news, special report and chronicle) and literary fiction text (short story and romance). A Portuguese corpus made of texts extracted from the Web (news and Wikipedia texts) with annotated zero subjects was used for testing an anaphora resolution system (Bick, 2010).

The main difference between our corpus and the corpora above is that ESZIC contains texts of different genres: legal (laws) and health (psychiatric papers).

[1] `http://luzrello.com/Projects.html`.

3. Linguistic Criteria

Literature related to ellipsis in linguistic theory together with the requirements of the NLP practices have served as a basis for establishing the categories and the annotation criteria in ESZIC. The linguistic motivation for each of the categories is shown in this section against the annotation tags to which they belong.

The terminology and the linguistic explanations relevant for this work consider both zero subjects and non-referential expressions to be different types of ellipsis. Our annotation tags typology is based on the four kinds of subject ellipsis distinguished in (Brucart, 1999) together with the linguistic issues described in (Matos, 2003).

3.1. Annotation Tags

Figure 1 shows the linguistic and formal criteria used to identify the chosen categories that served as the basis for the corpus annotation. From each annotation category, in addition to the two criteria that are crucial for this study ([± elliptic] and [± referential] subjects) a combination of syntactic, semantic and discourse knowledge was also encoded during the annotation. This knowledge includes information about whether the subject and its head are phonetically realized, whether the subject is nominal or non-nominal, whether it is an active or a passive subject or whether the subject refers to an active participant in the action, state or process denoted by the verb.

These thirteen annotation tags aim to cover all the possible elements which occur in the argumental subject position in the clause.During an annotation testing phase, we evaluated the adequacy and clarity of the annotation guidelines and established a typology of the rising borderline cases that was included in the annotation guidelines.

3.2. Annotated Categories

The features into which the subjects were distinguished are: [± elliptic] subjects and [± referential] subjects. From these two labels result four possibilities, but only three occur in Spanish and Portuguese.

- Explicit subjects: non-elliptic and referential;[2]

  (a) (Sp.) *La Constitución Española* fue refrendada por el pueblo español el 6 de diciembre de 1978.
  The Spanish Constitution was countersigned by the Spanish population on the 6th of December of 1978.

- Zero subjects: elliptic and referential;[3] and

  (b) (Pt.) Ø São formas de conhecimento que se manifestam como elementos cognitivos.
  They are forms of knowledge expressed as cognitive elements.

- Impersonal constructions: elliptic and non-referential.[4]

  (c) (Pt.) Procederá-se em a forma deste e dos arts.
  (It) will proceed as the form of this article.

As seen, a subject can be referential (zero subject) or non-referential. The distinction lies in the fact that, while the former can be lexically retrieved, the latter cannot (impersonal construction).

3.3. Borderline Cases

Additional guidelines were established for the annotation of borderline instances whose classification is a frequent source of disagreement between annotators. We classified the borderline cases in the following types:

(i) definition of particular syntactic categories which can function as subjects,

(ii) definition of cataphora cases, and the

(iii) intricate differentiation of impersonal sentences with *"se"* and reflex passive.

4. The ESZIC Corpus

The ESZIC corpus is composed of 34 documents, originally written in peninsular Spanish and Brazilian Portuguese, and belonging to two genres: legal and health. In Table 2 we detail the number of tokens, sentences and clauses per language. Clauses contain only one finite verb while sentences might contain more.

Spanish texts were analysed using Connexor's Machinese Syntax[5] which uses Functional Dependency Grammar, while Portuguese texts were parsed by Palavras,[6] a parser based on the Constraint Grammar methodological paradigm.

Four volunteer graduate students, native speakers of Spanish and Portuguese with no previous experience in corpus annotation, participated in the task. The experiment was run in two sessions: one training session and one testing session. First, each of the annotators was trained through a sixty minutes seminar which explained the annotation guidelines and, afterwards, the volunteer was supervised trough a testing annotation process. A program was written in Python to extract all occurrences of finite verbs from the parsed documents. The annotators were presented the sentences in which a verb or a group of verbs appeared and prompted to classify the verb into one of the thirteen classes shown in Figure 1. Each of the tags were grouped in one of the three main categories. Table 3 presents the number of instances found by category and genre in the corpus.

4.1. Reliability

To measure inter-annotator reliability we chose Fleiss' Kappa statistical measure (Fleiss, 1971).

We extracted 10% of the instances of each of the texts covering the two genres and two languages. From these instances, we discarded the examples considered ambiguous. This might overestimate the reliability; however, we only found two ambiguous instances.

[2] Explicit subjects in the examples are presented in *italics.*

[3] Zero subjects are presented by the symbol Ø.

[4] Impersonal constructions in the examples are not explicitly indicated using a symbol.

[5] `http://www.connexor.eu/technology/machinese/`.

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

Linguistic information		Phonetic Realization		Syntactic category	Verbal Diathesis	Semantic interpretation	Discourse
Annotation Categories	**Annotation Tags**	Elliptic noun phrase	Elliptic noun phrase head	Nominal subject	Active	Active participant	Referential subject
Explicit subject	Explicit subject	–	–	+	+	+	+
	Reflex passive subject	–	–	+	+	–	+
	Passive subject	–	–	+	–	–	+
Zero pronoun	Omitted subject	+	–	+	+	+	+
	Omitted subject head	–	+	+	+	+	+
	Non-nominal subject	–	–	–	+	+	+
	Reflex passive omitted subject	+	–	+	+	–	+
	Reflex passive omitted subject head	–	+	+	+	–	+
	Reflex passive non-nominal subject	–	–	–	+	–	+
	Passive omitted subject	+	–	+	–	–	+
	Passive non-nominal subject	–	–	–	–	–	+
Impersonal construction	Reflex impersonal clause (with *se*)	–	–	n/a	–	n/a	–
	Impersonal construction (without *se*)	–	–	n/a	+	n/a	–

Table 1: ESZIC Annotation Tags.

Collection	ESZIC_es			ESZIC_pt		
Texts	Tokens	Sentences	Clauses	Tokens	Sentences	Clauses
Legal	56,453	3,510	3,556	57,269	3,011	2,523
Health	37,058	1,702	3,530	45,018	2,045	3,554
Total	**93,511**	**5,212**	**7,086**	**102,287**	**5,056**	**6,077**

Table 2: ESZIC: Number of Tokens, Sentences and Clauses.

No. of Instances	ESZIC_es	ESZIC_pt
Explicit subjects	4,855	4,353
Zero subjects	1,793	1,202
Impersonal constructions	179	110
Total	**6,827**	**5,665**

Table 3: ESZIC: Number of Instances per Category.

Collection	Two Annotators	Three Annotators
ESZIC_es Legal	0.945	0.934
ESZIC_pt Legal	0.826	0.826
ESZIC_es Health	0.949	0.870
ESZIC_pt Health	0.958	0.857

Table 4: ESZIC's Fleiss Kappa Inter-annotator Agreement Coefficient.

Our results indicate that the ESZIC annotation is reliable to an acceptable degree. There is a small number of categories but the Fleiss Kappa value is high. Therefore, the ESZIC corpus can provide a reliable resource to study subject ellipsis in Portuguese and Spanish.

In Table 5 we show the coincidence matrix of the corpus. This coincidence matrix reports in the diagonal the perfect agreements where in the rest reports the number of cases where two annotators (rows) disagreed with one annotator (columns). Disagreements outside the diagonal are not frequent.

We noticed that coincidence matrixes of legal texts show more annotation consistency.

ESZIC	Explicit Subject	Zero Subject	Impersonal Construction
Subject	**945**	49	3
Zero	32	**249**	1
Impersonal	6	3	**22**

Table 5: Coincidence Matrix of ESZIC.

Class	P	R	F	Acc.
Explicit subj.	90.1%	92.3%	91.2%	87.3%
Zero subj.	77.2%	74.0%	75.5%	87.4%
Impersonals	85.6%	63.1%	72.7%	98.8%

Table 6: K* performance (87.6% accuracy for ten-fold cross validation).

4.2. Usefulness

This resource was found to be useful to improve subject ellipsis recognition (Rello et al., 2010; Rello et al., 2012) as well as to carry out linguistic descriptions of ellipsis (Rello and Illisei, 2009a; Gayo and Rello, 2011).

ESZIC_es was used as a training corpus to solve the problem of the identification of zero subjects and impersonal constructions in Spanish. The analyses and results presented in (Rello et al., 2012) show the usefulness of this resource since the machine learning method trained on this corpus was found to be more accurate than the other approaches for identifying explicit subjects and impersonal constructions in Spanish. In Table 6 we show the performance of the machine learning method using the KStar algorithm.

In (Gayo and Rello, 2011) ESZIC_pt was used to validate the linguistic hypothesis of being Brazilian Portuguese a partial pro-drop language

5. Conclusions

In this paper we have presented a linguistic classification and a free resource to study subject ellipsis in Portuguese and Spanish. We have described the specific characteristics of ESZIC Corpus and concisely discussed the linguistic criteria behind the annotation categories and the sources of disagreement.

The reliability and usefulness of this resource is proved by: a relatively high inter-annotator agreement; and the possibility of training and testing learning-based algorithms for automatic subject ellipsis detection in Spanish as well as to carry out a linguistic description of Portuguese ellipsis occurrence in real data.

Further explorations of this resource related to anaphora resolution, cross-lingual ellipsis identification and genre analysis are expected for future work.

6. References

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

J. M. Brucart. 1999. La elipsis. In I. Bosque and V. Demonte, editors, *Gramática descriptiva de la lengua española*, volume 2, pages 2787–2863. Espasa-Calpe, Madrid.

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

A. Ferrández 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.

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

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

G. Matos. 2003. Construções elípticas. In Mateus et. al., editor, *Gramática da Língua Portuguesa*, page Cap. 21. Caminho, Lisboa.

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

V. Ng 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.

S. Pereira. 2009. ZAC. PB: An annotated corpus for zero anaphora resolution in Portuguese. In *Student Research Workshop. International Conference on Recent Advances in Natural Language Processing (RANLP-09)*, pages 209–214.

M. Recasens and M.A. Martí. 2010. Ancora-co: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, 44(4):315–345.

L. Rello and I. Illisei. 2009a. A comparative study of Spanish zero pronoun distribution. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains (ISMTCL-09)*, pages 209–214. Presses Universitaires de Franche-Comté, Besançon.

L. Rello and I. Illisei. 2009b. 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.

L. Rello, P. Suárez, and R. Mitkov. 2010. A machine learning method for identifying non-referential impersonal sentences and zero pronouns in Spanish. *Procesamiento del Lenguaje Natural*, 45:281–287.

L. Rello, R. Baeza-Yates, and R. Mitkov. 2012. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in Spanish. In *Proceedings of the 13th European chapter of the Association for Computational Linguistics (EACL 2012)*.