A Comparative Study of Spanish Zero Pronoun Distribution

Luz Rello and Iustina Ilisei
Research Group in Computational Linguistics
University of Wolverhampton

Wolverhampton, United Kingdom
{*luzrello, iustina.ilisei*}*@gmail.com*

***Abstract*—The aim of this paper is to report the distribution of Spanish zero pronouns in three different genres: legal, encyclopaedic and instructional. The Z-corpora were created for this purpose and a sample of 1043 zero pronouns was annotated. The most salient patterns of distribution are compared for each genre, and some relevant issues concerning the use of zero pronouns are described in relation to a zero pronoun identification algorithm that is to be implemented.**

***Keywords*-zero-pronoun; anaphora; ellipsis; corpus annotation; corpus analysis**

I. Introduction

1) Background: Study of the distribution of zero pronouns in Spanish is a first step research towards the development of pre-processing tools useful in NLP fields where anaphora resolution in Spanish is necessary, such as the generation of multiple choice tests, automatic summarisation, machine translation, question answering, etc.

In particular, the generation of multiple choice tests in Spanish requires specific techniques in addition to the ones used for the English language [1], which is due to the flexible syntactic structure of the Spanish language. These specific techniques need methods such as recognition and resolution of the widespread occurrence of zero pronouns, and methods to identify simple declarative sentences. The development of these methods could facilitate the creation of specific rules for multiple choice test question generation for the Spanish language because the main process to generate a question starts from the detection of the subject in declarative sentences. However, when the subject is elliptic, resolution of zero anaphora is necessary in order to identify it.

2) Methodology: This preparatory research provides preliminary information about the patterns of distribution for zero pronouns in Spanish. Based on these findings, useful data for the design of the zero pronoun identification algorithm are outlined. This data will be used in subsequent research to develop an application for zero pronoun identification and resolution in Spanish.

This study required deep parsing techniques and the compilation and manual annotation of corpora. The aim of our approach is to offer an overview of the distribution of zero pronouns according to a set of parameters, such as their position in the sentence in relation to their antecedent and the verb on which they depend and the type of sentence in which they occur amongst others.

3) Outline of the paper: The paper is structured as follows: section II outlines previous studies and considerations about the zero pronoun and its importance in Spanish; section III describes the Z-corpora; section IV discloses the methodology used for the annotation of the corpus and the criteria applied; section V analyses the results of the distribution patterns taking into consideration the genre, the types of sentences, the kind of sub-clauses and the locations of antecedent and verb; and section VI presents some issues for the future zero pronoun identification method, especially relating to the difficulty that impersonal verbs in Spanish entail as well as some other concerns.

II. What is a zero-pronoun?

A zero pronoun[1] is the resultant "gap" (zero anaphor), where zero anaphora or ellipsis occurs. Among the different forms of zero anaphora, this study focuses on zero pronominal anaphora, which occurs when an anaphoric pronoun is omitted but nevertheless understood [4].

There are two different views about the distribution of zero pronouns. While some approaches consider that Spanish zero pronouns only appear in the position of the subject [5], others [3] defend that zero pronouns can be noticed in the object position[2] as well. Nevertheless, with regard to the final objective of this research, only zero pronouns in subject position are taken into consideration.

Two related phenomena occur when the Spanish subject is omitted: anaphora and ellipsis [6]. The zero pronoun is anaphoric (b) when it points back to its antecedent

[1] Notice that the term zero pronoun used in this paper is not equivalent to the concept of zero pronoun from the *Zero Hypothesis* [2], where *Zero pronoun* can have phonetic content (full pronoun) or not (null pronoun). In this theory the concept of Zero pronoun has to do only with its lack of lexical content in opposition to *Lexical pronouns* [3].

[2] "Spanish allows null arguments in subject and object position, in these positions Spanish Zero pronouns are null." [3] It follows an example (a) of zero pronoun in object position taken from [3]:

(a)
Napoleón$_i$ *le*$_j$ ordenó *pro*$_j$ *pro*$_j$ atacar.
Napoleón him order:3SPST pro pro to attack.
Napoleon ordered him to attack.

(coreferential relation), while the zero pronoun is elliptical[3] (c) when there is no linguistic entity to which it refers [9]. Both phenomena are closely related and occur in ambiguous examples especially when omitting the subject [9]. However, both are examples of zero anaphora [4].

(b)

La costumbre$_i$ sólo regirá en defecto de ley aplicable, siempre que *ZP[ella]*$_i$ no sea contraria a la moral o al orden público y que *ZP[ella]*$_i$ resulte probada.[4]

The custom$_i$ will only be valid by default on the applicable law, whenever *ZP[it]*$_i$ is not opposite to the moral a the public order and *ZP[it]*$_i$ is passed.

(c)

Los sistemas químicos que *ZP[nosotros]*$_i$ podemos estudiar por vía experimental son más complejos.

The chemistry systems that *ZP[we]*$_i$ can study through experimental way are more complex.

The Z-corpora contain instances of both cases in its annotation.

III. Description of the Z-Corpora

One contribution of this research is that the corpora are composed of different genres (legal, instructional and encyclopaedic) than those compiled in previous works[5] in relation with the zero pronoun resolution[6] in Spanish [11]. As this study proves, the genre determines the distribution and frequency of zero pronouns.

The specific orientation of this study has led to the selection of the genres which are likely to be the subject of multiple choice text generation in the real world (Highway Code, textbooks, legal texts used in state competition exams, etc.). They are mainly composed of declarative sentences from which questions can be generated.

The legal corpus contains the Spanish Constitution (from the beginning to article 110), the first book of the Civil Law Code (to article 10 in Chapter III) and the Highway Code (to Chapter II, article 27). The instructional corpus is composed of three handbooks taken from the open source wikibooks: Chemistry, Sewage Engineering and Relativity Theory. The encyclopaedic corpus is made up of 127 Wikipedia articles about mammals, medicine and linguistics[7]. They all together constitute a sample of 1043 zero pronouns. Sample containing roughly equivalent numbers of zero pronouns were collected in each genre.

Table I
Z-Corpora Description

Overview	LT	IT	ET	LT+IT+ET
No. of tokens	21961	16843	13386	52190
No. of sentences	1316	698	592	2606
No. of finite verbs	1765	1753	1314	4832
No. of zero pronouns	342	351	349	1042
Avg. tokens/sentence	19.47	26.25	24.67	23.46
Avg. zero pronoun/sentence	0.33	0.68	0.59	0.54

IV. Annotation Criteria

The Z-corpora were parsed with the dependency parser for Spanish, Connexor's Machinese [8]. This parser tokenises the text, providing the parts of speech, the morphological and syntactic information of each token, its lemma and dependencies. As this parser does not identify zero pronouns, each zero pronoun was annotated manually by adding an empty xml tag (d) containing informative attributes in the parsed text. Each tag was included preferably at the beginning of the clause, unless this position produced an ungrammatical result.

(d)

<ZERO_PRONOUN id=“w2440.5” ant=“w2419” depend_head=“w2441” agreement=“high” sentence_type=“sub” title=“yes”/ >

The criteria in annotating the corpora take into account the long term objective of this research, Spanish zero pronoun resolution. Therefore, each zero pronoun tag includes the following information in its attributes: the position of the zero pronoun in the sentence, its antecedent, its dependency head (the clause verb), the kind of sentence in which it appears and the clause where it stands. The morphological information of each zero pronoun was not annotated as it can be extracted from the verb on which the zero pronoun depends (included in the dependency head attribute). Furthermore, the syntactic role of the zero pronoun is always a subject.

The antecedent attribute has three different possible values: (i)“elliptic” —when there is no antecedent—, (ii) “non_nominal’, when the antecedent is a clause, or as in the majority of the cases, (iii) a number which points the existing antecedent in the text. The antecedent to which the zero pronoun tag refers is the immediately previous one in the coreferential chain and when the antecedent is a complex noun or noun phrase, it points to its head[9] .

The dependency head attribute points to the verb on which the zero pronoun depends. If the verb is complex, it points to the auxiliary verb. In order to cover the possible clauses where the zero pronoun appears, two more attributes (sentence type and subtype) provide the information of the kind of sentence (main, coordinated, subordinated, etc.) and

[3] While some linguists consider these examples as an ellipsis of the subject [7], it has been claimed that the subject is not elliptic as it is present in the verb morphology [8].

[4] Unless otherwise specified, all the examples are taken from the Z-corpora.

[5] The Spanish AnCora corpus includes the annotation of the zero pronoun (10,791 examples tagged as elliptic pronouns in subject position). Although it is based on journalistic texts [10], its exploitation is valuable for future work.

[6] The corpus used were the handbook Blue Book corpus (15,571 words) and Lexesp (9,745 words), which contains mainly texts taken from newspapers [11].

[7] Table I follows the order in which the texts were mentioned in the text. The encyclopaedic articles were categorised into three sets according to their topics.

[8] http://www.connexor.eu/technology/machinese/

[9] Following the Functional Discourse Grammar of Connexor Machinese Language Model, so the complete antecedent can be retrieved [12].

the clause type (simple, copulative, relative, etc.) to which the zero pronoun belongs.

Moreover, there are two other optional attributes: one for the cases where the zero pronoun is cataphoric and another one which displays whether the antecedent of the zero pronoun corresponds with the "title" of an encyclopaedic entry or legal article. This last phenomenon turned out to be fairly significant: 171 out of 1043 zero pronouns find their antecedent in the title.

The annotation of the corpora was carried out by one author alone, hence no inter-annotator agreement measure could be performed. Nevertheless, there was an attempt to estimate the possible points of disagreement as there were some unclear examples while annotating. An additional attribute was added offering the annotator the possibility of including their grade of confidence for each example. This difficulty in annotating the zero pronouns is derived mainly from the ambiguity of language itself and from the subtle difference between anaphora and ellipsis in subject omission in Spanish[10]. Some uncertain examples will be detailed in Section VI.

V. Distribution of the zero pronoun in Spanish

The frequency of the zero pronoun in Spanish is high, for instance, in our Z-corpora there is an average of 0.59 zero pronouns per sentence. Moreover, in [11] from the 1,599 verbs classified in their corpora, 734 of them have zero pronouns, whilst in AnCora [10] for every 49 words a zero pronoun is found.

The annotated corpora has a sample data of 52,190 tokens within 2,606 sentences, legal texts having the highest number of tokens and sentences. There is an expected correlation between the number of finite verbs and the number of zero pronouns as can be infered from Table I. As it can be observed, the number of zero pronouns is proportional to the number of finite verbs with an average of 21.98%.

However, the mean value of zero pronouns per sentence emphasises the fact that legal texts contain a lower average of zero pronouns per sentence with only 0.33, comparing to number 0.68 and 0.59 for the instructional and encyclopaedic texts respectively. One reason for this lower average in legal texts can be the aim to avoid ambiguity and misinterpretations that may occur.

The fact that the sentences in instructional texts reach the highest length could be a reason for the highest proportion of zero pronouns found in this genre, with an average of 0.68 zero pronouns compared to 0.33 zero pronouns average for legal texts. On the other hand, the legal corpus contains the shortest sentences, and thus, the lowest proportion of zero pronouns.

In contrast to the striking differences found in the zero pronoun distribution between the diverse texts of the legal and instructional genres (with standard deviations of 0.15 and 0.33 respectively), the encyclopaedic texts reveal a more constant distribution, with a standard deviation of only 0.04. The percentage of zero pronouns out of the finite verbs is a constant feature throughout all texts with an average of 0.24 and a standard deviation of 0.08.

The results provided in Table II show that the distribution of elliptic subjects is approximately constant throughout all three genres, while the zero pronoun with elliptic antecedent is highly dependent on the genre. In Z-corpora a percentage of 18.71% has been noticed for the zero pronouns which have elliptic antecedent and 61.80% for elliptic subjects. Comparing the results, the legal texts have the lowest percentage of zero pronouns with elliptic antecedent, in contrast to the instructional texts, for which a higher frequency is outlined. The influence of the genre is obvious, and thus, a system for zero pronoun resolution could consider this variation.

Table II
Elliptic Subjects and Zero pronoun with Elliptic Antecedents

Percentage (%)	Elliptical Subjects	Zero Pronoun with Elliptical Antecedents
LT	61.42	5.26
IT	63.15	40.46
ET	60.5	10.03
LT+IT+ET	61.8	18.71

From all the subordinated clauses which contain zero pronouns, more than half are relative clauses.

For the legal genre, zero pronouns located in the relative clauses are considerably more frequent compared to the simple clauses, having a percentage of 53.22%, while the presence of a zero pronoun in a simple clause is only 0.29% (Table III). In contrast, the number of simple sentences containing zero pronouns is high for the encyclopaedic genre, reaching up to 15.47% in the main clauses.

Table III
Simple and Relative Clauses

Percentage (%)	Simple	Relative
LT	0.29	53.22
IT	5.7	39.6
ET	15.47	40.11

Encyclopaedic texts are the ones which have the greatest number of zero pronouns in main sentences with a percentage of 29.51%, whilst legal text subordinated clauses have the frequency of 69.88%. Table IV proves that most of the sentences containing zero pronouns are subordinated

[10] "Podemos encontrar dos problemas lingüísticos no tan claros de encuadrar o bien dentro de la elipsis o bien dentro de la anáfora; hablamos del problema del núcleo del sintagma nominal (elipsis o anáfora) y de la omisión del sujeto nominal" [9]

"There are problems in the attempt to classify some instances as elliptic or anaphoric: this is due to the problematic examples that bring the omission of the noun phrase head or the nominal subject."

sentences, their frequency varying between 50.43% for encyclopaedic texts and 70.66% for instructional texts.

Table IV
Sentence Types

Percentage(%)	Main Clauses	Subordinate Clauses	Coordinate Clauses	Juxtaposed Clauses
LT	5.56	69.88	19.01	5.56
IT	17.95	70.66	7.12	4.27
ET	29.51	50.43	10.03	10.03

In Table V, the first column shows the antecedent distance in terms of number of sentences. From the statisitics included in the table it can be seen that zero pronoun anaphora operates over the longest distances in the legal genre pointing to an antecedent which is on average 2.97 sentences back, with the highest standard deviation of the three genres: 7.15 sentences. In contrast, for the other two genres, the antecedent tends to be in the same or in the previous sentence, according to the averages presented: 0.42 and 0.11 for encyclopaedic and instructional genres, with standard deviations of 0.81 and 0.43 sentences respectively.

Additionally, the second column displays the antecedent distance measured in tokens, which indicates that legal and encyclopaedic texts have the longest distance anaphors, an average of 36.41 and 12.34 tokens in between the zero pronoun and their antecedents, respectively, while in the instructional genre this average is considerable lower: 5.67 tokens. Nevertheless, he overall Z-corpora mean value is 17.92 tokens between the zero pronoun and its antecedent, with a standard deviation of 19.48 tokens.

The third column of the table represents the average distance between the zero pronoun and its dependent verb, measured in number of tokens. The furthermost position of the verb on which the zero pronoun depends occurs in legal texts with a distance of 1.58 tokens. However, a significant difference between the various genres in the Z-corpora has not been noticed, as the average distance is 1.42 tokens. The verb distance from the zero pronoun has a standard deviation of 0.64 tokens distance in the instructional texts, 1.63 for legal genre and 1.004 for the encyclopaedic texts.

Table V
Antecedent distance and dependend verb distance

	Antecedent distance: Avg. of sentences	Antecedent distance: Avg. of tokens	Dependent verb distance: Avg. of tokens
LT	2.97	36.41	1.58
IT	0.11	5.67	1.25
ET	0.42	12.34	1.42
LT+IT+ET	1.15	17.92	1.42

VI. Issues for the future zero pronoun identification

Qualitatively, some repetitive patterns in the zero pronoun behaviour were observed, which can be useful in the future identification and resolution method.

In subordinated clauses, the zero pronoun antecedent tends to be fairly close —it is rarely found outside the same sentence—, while zero pronouns in main sentences are longer-distance anaphors whose antecedents tend to be in the subject of some of the previous sentences.

Moreover, in the legal and more frequently in encyclopaedic texts, a tendency has been observed (e) to state the antecedent of the zero pronoun in the title of the article or entry.

(e)

Monotremas$_i$

ZP[ellos]$_i$ Son los mamíferos más primitivos. *ZP[ellos]*$_i$ Son ovíparos y *ZP[ellos]*$_i$ poseen glándulas mamarias sin [...]

Monotremes$_i$

ZP[They]$_i$ are the most primitive mammals. *ZP[They]*$_i$ are oviparous and *ZP[They]*$_i$ have mammary glands

The future zero pronoun identifier method will have to address some difficulties due to the nature of the Spanish language itself and the ones related to the parser output. The main issue concerns Spanish impersonal verbs (f), which are not detected by the parser. Moreover, in some cases, these verbs can have a subject without any changes in form (g), which adds to the complexity of the task.

(f)

Se distingue un rasgo característico [...]

ZP[It] is distinguished a characteristic feature [...]

(g)

El objeto de estudio de *la química*$_i$ es la materia y sus transformaciones, aunque *ZP[ella]*$_i$ se distingue de la física [...]

Chemistry$_i$ subject of study is the matter and its transformations, although *ZP[it]*$_i$ is distinguished from Physics [...]

The lack of semantic information in addition to the unavoidable language ambiguity causes other problems. Example (h) shows frequently occurring patterns of number disagreement between the zero pronoun and its antecedent, which are seen to occur in some encyclopaedia entries. Real world knowledge is needed to infer that the plural zero pronoun points back to the singular antecedent in the title. In example (i), the subject is postponed and it is composed by the sentences stated after the colon.

(h)

Oposum$_i$

(45 cm) Cola larga y escamosa, orejas en forma de cucurucho, patas prensiles. *ZP[ellos]*$_i$ Se alimentan de materias vegetales y de animales vivos o muertos.

Oposum$_i$

(45 cm) Long and scaly tail, cone shaped ears, prehensile legs. *ZP[they]*$_i$ feed on vegetable matter and living or death animals

(i)

[...] corresponderá a la Administración del Estado:
a) La facultad de determinar [...]
b) La previa homologación [...]
c) La publicacin de las normas básicas y mínimas [...]
[it] will correspond to the State Administration:
a) The faculty to determine [...]
b) The previous recognition [...]
c) The publication of basic and minimum regulations [...]

Moreover, instructional texts often include explanatory examples which are interpreted as part of the sentence by the parser (j):

(j)

Podría decidirse, así, que "perdió", "repartió", "exportó", etc. pertenecen a una clase [...]

[It] could be stated that "[he]lost", "[he]shared", "[he]exported", etc. belongs to a class [...]

Finally, some other inevitable problems are generated by the parser itself. It encounters difficulties in detecting the subject when it comes after the verb, and imperative verbs are often parsed as subjunctive ones, as they have the same form in some cases.

VII. Conclusions and future work

In this study, we have presented new Spanish corpora composed of legal, encyclopaedic and instructional texts, annotated with zero pronouns. We have reported quantitative analysis of the distribution of zero pronouns and outlined some qualitative observations concerning their behaviour that will facilitate the design of a methodology for their identification and resolution. Analysis of the legal genre led to important discoveries, especially regarding the search scope required for a resolution algorithm. Concerning the Z-Corpora data, the measurement of inter-annotator agreement is an important aspect of the corpus validation process and so in future work, annotators will be involved in order to undertake such assessments.

Acknowledgment

We would like to express our gratitude to Prof. Ruslan Mitkov for his supervision and useful suggestions throughout this study. We are also especially thankful to Richard Evans, Jose M. Gómez, Elena Lloret and Claudiu Mihăilă for their invaluable comments and contributions.

References

[1] R. Mitkov, L. An, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Journal of Natural Language Engineering*, vol. 12, no. 2, pp. 177–194, 2006.

[2] A. Kratzer, "More structural analogies between pronouns and tenses," in *The Proceedings of Semantics and Linguistic Theory VIII (SALT VIII)*. Cornell University: CLC Publications, 1998.

[3] L. Alonso-Ovalle and F. D'Introno, *Hispanic Linguistics at the Turn of the Myllenium*. Sommerville, MA: Cascadilla Press, 2001, ch. Full and Null Pronouns in Spanish: the Zero Pronoun Hypothesis, pp. 189–210.

[4] R. Mitkov, *Anaphora resolution*. London: Longman, 2002.

[5] J. Peral and A. Ferrández, *Lecture Notes In Computer Science*. London, UK: Springer-Verlag, 2000, vol. 1835, ch. Generation of Spanish Zero-Pronouns into English, pp. 252–260.

[6] J. M. Brucart, *La elisión sintáctica en español*. Barcelona: Universitat Autónoma de Barcelona, 1987.

[7] ——, *Gramática descriptiva de la lengua española, 2*. Madrid: Espasa-Calpe, 1999, ch. "La elipsis", pp. 2787–2863.

[8] RAE, *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa Calpe, 1977.

[9] A. Ferrández, A. Palomar, and L. Moreno, "El problema del núcleo del sintagma nominal: elipsis o anáfora?" *Procesamiento del lenguaje natural*, vol. 20, pp. 13–26, 1997.

[10] M. Taulé, M. Martí, and M. Recasens, "Ancora: Multilevel annotated corpora for catalan and spanish," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008.

[11] A. Ferrández and J. Peral, "A computational approach to zero-pronouns in spanish," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, 2000, pp. 166–172.

[12] Connexor, *Machinese Language Model*, 1997-2006.