Elliphant: Improved Automatic Detection of Zero Subjects and Impersonal Constructions in Spanish

Luz Rello*
NLP and Web Research Groups
Univ. Pompeu Fabra
Barcelona, Spain

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain

Ruslan Mitkov
Research Group in
Computational Linguistics
Univ. of Wolverhampton, UK

Abstract

In pro-drop languages, the detection of explicit subjects, zero subjects and non-referential impersonal constructions is crucial for anaphora and co-reference resolution. While the identification of explicit and zero subjects has attracted the attention of researchers in the past, the automatic identification of impersonal constructions in Spanish has not been addressed yet and this work is the first such study. In this paper we present a corpus to underpin research on the automatic detection of these linguistic phenomena in Spanish and a novel machine learning-based methodology for their computational treatment. This study also provides an analysis of the features, discusses performance across two different genres and offers error analysis. The evaluation results show that our system performs better in detecting explicit subjects than alternative systems.

1 Introduction

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.

Various natural language processing (NLP) tasks benefit from the identification of elliptical subjects, primarily anaphora resolution (Mitkov, 2002) and co-reference resolution (Ng and Cardie, 2002). The difficulty in detecting missing subjects and non-referential pronouns has been acknowledged since the first studies on the computational treatment of anaphora (Hobbs, 1977; Hirst, 1981). However, this task is of crucial importance when processing pro-drop languages since subject ellipsis is a pervasive phenomenon in these languages (Chomsky, 1981). For instance, in our Spanish corpus, 29% of the subjects are elided.

Our method is based on classification of all expressions in subject position, including the recognition of Spanish non-referential impersonal constructions which, to the best of our knowledge, has not yet been addressed. The necessity of identifying such kind of elliptical constructions has been specifically highlighted in work about Spanish zero pronouns (Ferrández and Peral, 2000) and co-reference resolution (Recasens and Hovy, 2009).

The main contributions of this study are:

- A public annotated corpus in Spanish to compare different strategies for detecting explicit subjects, zero subjects and impersonal constructions.
- The first ML based approach to this problem in Spanish and a thorough analysis regarding features, learnability, genre and errors.
- The best performing algorithms to automatically detect explicit subjects and impersonal constructions in Spanish.

The remainder of the paper is organized as follows. Section 2 describes the classes of Spanish subjects, while Section 3 provides a literature review. Section 4 describes the creation and the annotation of the corpus and in Section 5 the machine learning (ML) method is presented. The analysis of the features, the learning curves, the

[*] This work was partially funded by a 'La Caixa' grant for master students.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 706–715,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

genre impact and the error analysis are all detailed in Section 6. Finally, in Section 7, conclusions are drawn and plans for future work are discussed. This work is an extension of the first author master's thesis (Rello, 2010) and a preliminary version of the algorithm was presented in Rello et al. (2010).

2 Classes of Spanish Subjects

Literature related to ellipsis in NLP (Ferrández and Peral, 2000; Rello and Illisei, 2009a; Mitkov, 2010) and linguistic theory (Bosque, 1989; Brucart, 1999; Real Academia Española, 2009) has served as a basis for establishing the classes of this work.

Explicit subjects are phonetically realized and their syntactic position can be pre-verbal or post-verbal. In the case of post-verbal subjects (a), the syntactic position is restricted by some conditions (Real Academia Española, 2009).

(a) Carecerán de validez *las disposiciones que contradigan otra de rango superior.*[1]
The dispositions which contradict higher range ones will not be valid.

Zero subjects (b) appear as the result of a nominal ellipsis. That is, a lexical element –the elliptic subject–, which is needed for the interpretation of the meaning and the structure of the sentence, is elided; therefore, it can be retrieved from its context. The elision of the subject can affect the entire noun phrase and not just the noun head when a definite article occurs (Brucart, 1999).

(b) Ø Fue refrendada por el pueblo español.
(It) was countersigned by the people of Spain.

The class of *impersonal constructions* is formed by impersonal clauses (c) and reflexive impersonal clauses with particle *se* (d) (Real Academia Española, 2009).

(c) No hay matrimonio sin consentimiento.
(There is) no marriage without consent.

(d) Se estará a lo que establece el apartado siguiente.
(It) will be what is established in the next section.

3 Related Work

Identification of non-referential pronouns, although a crucial step in co-reference and anaphora resolution systems (Mitkov, 2010),[2] has been applied only to the pleonastic *it* in English (Evans, 2001; Boyd et al., 2005; Bergsma et al., 2008) and expletive pronouns in French (Danlos, 2005). Machine learning methods are known to perform better than rule-based techniques for identifying non-referential expressions (Boyd et al., 2005). However, there is some debate as to which approach may be optimal in anaphora resolution systems (Mitkov and Hallett, 2007).

Both English and French texts use an explicit word, with some grammatical information (a third person pronoun), which is non-referential (Mitkov, 2010). By contrast, in Spanish, non-referential expressions are not realized by expletive or pleonastic pronouns but rather by a certain kind of ellipsis. For this reason, it is easy to mistake them for zero pronouns, which are, in fact, referential.

Previous work on detecting Spanish subject ellipsis focused on distinguishing verbs with explicit subjects and verbs with zero subjects (zero pronouns), using rule-based methods (Ferrández and Peral, 2000; Rello and Illisei, 2009b). The Ferrández and Peral algorithm (2000) outperforms the (Rello and Illisei, 2009b) approach with 57% accuracy in identifying zero subjects. In (Ferrández and Peral, 2000), the implementation of a zero subject identification and resolution module forms part of an anaphora resolution system.

ML based studies on the identification of explicit non-referential constructions in English present accuracies of 71% (Evans, 2001), 87.5% (Bergsma et al., 2008) and 88% (Boyd et al., 2005), while 97.5% is achieved for French (Danlos, 2005). However, in these languages, non-referential constructions are explicit and not omitted which makes this task more challenging for Spanish.

4 Corpus

We created and annotated a corpus composed of legal texts (law) and health texts (psychiatric

[1] All the examples provided are taken from our corpus. In the examples, explicit subjects are presented in *italics*. Zero subjects are presented by the symbol Ø and in the English translations the subjects which are elided in Spanish are marked with parentheses. Impersonal constructions are not explicitly indicated.

[2] In zero anaphora resolution, the identification of zero anaphors first requires that they be distinguished from non-referential impersonal constructions (Mitkov, 2010).

papers) originally written in peninsular Spanish. The corpus is named after its annotated content "Explicit Subjects, Zero Subjects and Impersonal Constructions" (ESZIC_es Corpus).

To the best of our knowledge, the existing corpora annotated with elliptical subjects belong to other genres. *The Blue Book* (handbook) and *Lexesp* (journalistic texts) used in (Ferrández and Peral, 2000) contain zero subjects but not impersonal constructions. On the other hand, the Spanish AnCora corpus based on journalistic texts includes zero pronouns and impersonal constructions (Recasens and Martí, 2010) while the Z-corpus (Rello and Illisei, 2009b) comprises legal, instructional and encyclopedic texts but has no annotated impersonal constructions.

The ESZIC corpus contains a total of 6,827 verbs including 1,793 zero subjects. Except for AnCora-ES, with 10,791 elliptic pronouns, our corpus is larger than the ones used in previous approaches: about 1,830 verbs including zero and explicit subjects in (Ferrández and Peral, 2000) (the exact number is not mentioned in the paper) and 1,202 zero subjects in (Rello and Illisei, 2009b).

The corpus was parsed by Connexor's Machinese Syntax (Connexor Oy, 2006), which returns lexical and morphological information as well as the dependency relations between words by employing a functional dependency grammar (Tapanainen and Järvinen, 1997).

To annotate our corpus we created an annotation tool that extracts the finite clauses and the annotators assign to each example one of the defined annotation tags. Two volunteer graduate students of linguistics annotated the verbs after one training session. The annotations of a third volunteer with the same profile were used to compute the inter-annotator agreement. During the annotation phase, we evaluated the adequacy and clarity of the annotation guidelines and established a typology of the rising borderline cases, which is included in the annotation guidelines.

Table 1 shows the linguistic and formal criteria used to identify the chosen categories that served as the basis for the corpus annotation. For each tag, in addition to the two criteria that are crucial for identifying subject ellipsis ([± elliptic] and [± referential]) a combination of syntactic, semantic and discourse knowledge is also encoded during the annotation. The linguistic motivation for each of the three categories is shown against the thirteen annotation tags to which they belong (Table 1).

Afterwards, each of the tags are grouped in one of the three main classes.

- Explicit subjects: [- elliptic, + referential].
- Zero subjects: [+ elliptic, + referential].
- Impersonal constructions: [+ elliptic, - referential].

Of these annotated verbs, 71% have an explicit subject, 26% have a zero subject and 3% belong to an impersonal construction (see Table 2).

Number of instances	Legal	Health	All
Explicit subjects	2,739	2,116	4,855
Zero subjects	619	1,174	1,793
Impersonals	71	108	179
Total	3,429	3,398	6,827

Table 2: Instances per class in ESZIC Corpus.

To measure inter-annotator reliability we use Fleiss' Kappa statistical measure (Fleiss, 1971). We extracted 10% of the instances of each of the texts of the corpus covering the two genres.

Fleiss' Kappa	Legal	Health	All
Two Annotators	0.934	0.870	0.902
Three Annotators	0.925	0.857	0.891

Table 3: Inter-annotator Agreement.

In Table 3 we present the Fleiss kappa inter-annotator agreement for two and three annotators. These results suggest that the annotation is reliable since it is common practice among researchers in computational linguistics to consider 0.8 as a minimum value of acceptance (Artstein and Poesio, 2008).

5 Machine Learning Approach

We opted for an ML approach given that our previous rule-based methodology improved only 0.02 over the 0.55 F-measure of a simple baseline (Rello and Illisei, 2009b). Besides, ML based methods for the identification of explicit non-referential constructions in English appear to perform better than than rule-based ones (Boyd et al., 2005).

LINGUISTIC INFORMATION		PHONETIC REALIZATION		SYNTACTIC CATEGORY	VERBAL DIATHESIS	SEMANTIC INTERPR.	DISCOURSE
Annotation Categories	**Annotation Tags**	Elliptic noun phrase	Ell. noun phrase head	Nominal subject	Active	Active participant	Referential subject
Explicit subject	Explicit subject	–	–	+	+	+	+
	Reflex passive subject	–	–	+	+	–	+
	Passive subject	–	–	+	–	–	+
Zero subject	Omitted subject	+	–	+	+	+	+
	Omitted subject head	–	+	+	+	+	+
	Non-nominal subject	–	–	–	+	+	+
	Reflex passive omitted subject	+	–	+	+	–	+
	Reflex pass. omitted subject head	–	+	+	+	–	+
	Reflex pass. non-nominal subject	–	–	–	+	–	+
	Passive omitted subject	+	–	+	–	–	+
	Pass. non-nominal subject	–	–	–	–	–	+
Impersonal construction	Reflex imp. clause (with *se*)	–	–	n/a	–	n/a	–
	Imp. construction (without *se*)	–	–	n/a	+	n/a	–

Table 1: ESZIC Corpus Annotation Tags.

5.1 Features

We built the training data from the annotated corpus and defined fourteen features. The linguistically motivated features are inspired by previous ML approaches in Chinese (Zhao and Ng, 2007) and English (Evans, 2001). The values for the features (see Table 4) were derived from information provided both by Connexor's Machinese Syntax parser and a set of lists.

We can describe each of the features as broadly belonging to one of ten classes, as follows:

1 PARSER: the presence or absence of a subject in the clause, as identified by the parser. We are not aware of a formal evaluation of Connexor's accuracy. It presents an accuracy of 74.9% evaluated against our corpus and we used it as a simple baseline.

2 CLAUSE: the clause types considered are: main clauses, relative clauses starting with a complex conjunction, clauses starting with a simple conjunction, and clauses introduced using punctuation marks (commas, semicolons, etc). We implemented a method to identify these different types of clauses, as the parser does not explicitly mark the boundaries of clauses within sentences. The method took into account the existence of a finite verb, its dependencies, the existence of conjunctions and punctuation marks.

3 LEMMA: lexical information extracted from the parser, the lemma of the finite verb.

4-5 NUMBER, PERSON: morphological information of the verb, its grammatical number and its person.

6 AGREE: feature which encodes the tense, mood, person, and number of the verb in the clause, and its agreement in person, number,

Feature	Definition	Value
1 PARSER	Parsed subject	True, False
2 CLAUSE	Clause type	Main, Rel, Imp, Prop, Punct
3 LEMMA	Verb lemma	Parser's lemma tag
4 NUMBER	Verb morphological number	SG, PL
5 PERSON	Verb morphological person	P1, P2, P3
6 AGREE	Agreement in person, number, tense and mood	FTFF, TTTT, FFFF, TFTF, TTFF, FTFT, FTTF, TFTT, FFFT, TTTF, FFTF, TFFT, FFTT, FTTT, TFFF, TTFT
7 NHPREV	Previous noun phrases	Number of noun phrases previous to the verb
8 NHTOT	Total noun phrases	Number of noun phrases in the clause
9 INF	Infinitive	Number of infinitives in the clause
10 SE	Spanish particle *se*	True, False
11 A	Spanish preposition *a*	True, False
12 POS$_{pre}$	Four parts of the speech previous to the verb	292 different values combining the parser's POS tags
14 POS$_{pos}$	Four parts of the speech following the verb	280 different values combining the parser's POS tags
14 VERB$_{type}$	Type of verb: copulative, impersonal pronominal, transitive and intransitive	CIPX, XIXX, XXXT, XXPX, XXXI, CIXX, XXPT, XIPX, XIPT, XXXX, XIXI, CXPI, XXPI, XIPI, CXPX

Table 4: Features, definitions and values.

tense, and mood with the preceding verb in the sentence and also with the main verb of the sentence.[3]

7-9 NHPREV, NHTOT, INF: the candidates for the subject of the clause are represented by the number of noun phrases in the clause that precede the verb, the total number of noun phrases in the clause, and the number of infinitive verbs in the clause.

10 SE: a binary feature encoding the presence or absence of the Spanish particle *se* when it occurs immediately before or after the verb or with a maximum of one token lying between the verb and itself. Particle *se* occurs in passive reflex clauses with zero subjects and in some impersonal constructions.

11 A: a binary feature encoding the presence or absence of the Spanish preposition *a* in the clause. Since the distinction between passive reflex clauses with zero subjects and impersonal constructions sometimes relies on the appearance of preposition *a* (to, for, etc.). For instance, example (e) is a passive reflex clause containing a zero subject while example (s) is an impersonal construction.

(e) Se admiten los alumnos que reúnan los requisitos.
Ø (They) accept the students who fulfill the requirements.

(f) Se admite a los alumnos que reúnan los requisitos.
(It) is accepted for the students who fulfill the requirements.

12-3 POS$_{pre}$, POS$_{pos}$: the part of the speech (POS) of eight tokens, that is, the 4-grams preceding and the 4-grams following the instance.

14 VERB$_{type}$: the verb is classified as copulative, pronominal, transitive, or with an impersonal use.[4] Verbs belonging to more than one class are also accommodated with different feature values for each of the possible combinations of verb type.

5.2 Evaluation

To determine the most accurate algorithm for our classification task, two comparisons of learning algorithms implemented in WEKA (Witten and Frank, 2005) were carried out. Firstly, the classification was performed using 20% of the training instances. Secondly, the seven highest performing classifiers were compared using 100% of the

[3] In Spanish, when a finite verb appears in a subordinate clause, its tense and mood can assist in recognition of these features in the verb of the main clause and help to enforce some restrictions required by this verb, especially when both verbs share the same referent as subject.

[4] We used four lists provided by Molino de Ideas s.a. containing 11,060 different verb lemmas belonging to the Royal Spanish Academy Dictionary (Real Academia Española, 2001).

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

training data and ten-fold cross-validation. The corpus was partitioned into training and tested using ten-fold cross-validation for randomly ordered instances in both cases. The lazy learning classifier K* (Cleary and Trigg, 1995), using a blending parameter of 40%, was the best performing one, with an accuracy of 87.6% for ten-fold cross-validation. K* differs from other instance-based learners in that it computes the distance between two instances using a method motivated by information theory, where a maximum entropy-based distance function is used (Cleary and Trigg, 1995). Table 5 shows the results for each class using ten-fold cross-validation. In contrast to previous work, the K* algorithm (Cleary and Trigg, 1995) was found to provide the most accurate classification in the current study. Other approaches have employed various classification algorithms, including JRip in WEKA (Müller, 2006), with precision of 74% and recall of 60%, and K-nearest neighbors in TiMBL: both in (Evans, 2001) with precision of 73% and recall of 69%, and in (Boyd et al., 2005) with precision of 82% and recall of 71%.

Since there is no previous ML approach for this task in Spanish, our baselines for the explicit subjects and the zero subjects are the parser output and the previous rule-based work with the highest performance (Ferrández and Peral, 2000). For the impersonal constructions the baseline is a simple greedy algorithm that classifies as an impersonal construction every verb whose lemma is categorized as a verb with impersonal use according to the RAE dictionary (Real Academia Española, 2001).

Our method outperforms the Connexor parser which identifies the explicit subjects but makes no distinction between zero subjects and impersonal constructions. Connexor yields 74.9% overall accuracy and 80.2% and 65.6% F-measure for explicit and elliptic subjects, respectively.

To compare with Ferrández and Peral (Ferrández and Peral, 2000) we do consider it without impersonal constructions. We achieve a precision of 87% for explicit subjects compared to 80%, and a precision of 87% for zero subjects compared to their 98%. The overall accuracy is the same for both techniques, 87.5%, but our results are more balanced. Nevertheless, the approaches and corpora used in both studies are different, and hence it is not possible to do a fair comparison. For example, their corpus has 46% of zero subjects while ours has only 26%.

Algorithm	Explicit subjects	Zero subjects	Impersonals
RAE	–	–	70.4%
Connexor	71.7%	83.0%	
Ferr./Peral	79.7%	**98.4%**	–
Elliphant	**87.3%**	87.4%	**98.8%**

Table 6: Summary of accuracy comparison with previous work.

For impersonal constructions our method outperforms the RAE baseline (precision 6.5%, recall 77.7%, F-measure 12.0% and accuracy 70.4%). Table 6 summarizes the comparison. The low performance of the RAE baseline is due to the fact that verbs with impersonal use are often ambiguous. For these cases, we first tagged them as ambiguous and then, we defined additional criteria after analyzing then manually. The resulting annotated criteria are stated in Table 1.

6 Analysis

Through these analyses we aim to extract the most effective features and the information that would complement the output of an standard parser to achieve this task. We also examine the learning process of the algorithm to find out how many instances are needed to train it efficiently and determine how much Elliphant is genre dependent. The analyses indicate that our approach is robust: it performs nearly as well with just six features, has a steep learning curve, and seems to generalize well to other text collections.

6.1 Best Features

We carried out three different experiments to evaluate the most effective group of features, and the features themselves considering the individual predictive ability of each one along with their degree of redundancy.

Based on the following three feature selection methods we can state that there is a complex and balanced interaction between the features.

6.1.1 Grouping Features

In the first experiment we considered the 11 groups of relevant ordered features from the training data, which were selected using each WEKA attribute selection algorithm and performed the classifications over the complete training data, using only the different groups features selected.

The most effective group of six features (NHPREV, PARSER, NHTOT, POS$_{pos}$, PERSON, LEMMA) was the one selected by WEKA's SymmetricalUncertAttribute technique, which gives an accuracy of 83.5%. The most frequently selected features by all methods are PARSER, POS$_{pos}$, and NHTOT, and they alone get an accuracy of 83.6% together. As expected, the two pairs of features that perform best (both 74.8% accuracy) are PARSER with either POS$_{pos}$ or NHTOT. Based on how frequent each feature is selected by WEKA's attribute selection algorithms, we can rank the features as following: (1) PARSER, (2) NHTOT, (3) POS$_{pos}$, (4) NHPREV and (5) LEMMA.

6.1.2 "Complex" vs. "Simple" Features

Second, a set of experiments was conducted in which features were selected on the basis of the degree of computational effort needed to generate them. We propose two sets of features. One group corresponds to "simple" features, whose values can be obtained by trivial exploitation of the tags produced in the parser's output (PARSER, LEMMA, PERSON, POS$_{pos}$, POS$_{pre}$). The second group of features, "complex" features (CLAUSE, AGREE, NHPREV, NHTOT, VERB$_{type}$) have values that required the implementation of more sophisticated modules to identify the boundaries of syntactic constituents such as clauses and noun phrases. The accuracy obtained when the classifier exclusively exploits "complex" features is 82.6% while for "simple" features is 79.9%. No impersonal constructions are identified when only "complex" features are used.

6.1.3 One-left-out Feature

In the third experiment, to estimate the weight of each feature, classifications were made in which each feature was omitted from the training instances that were presented to the classifier. Omission of all but one of the "simple" features led to a reduction in accuracy, justifying their inclusion in the training instances. Nevertheless, the majority of features present low informativeness except for feature A which does not make any meaningful contribution to the classification. The feature PARSER presents the greatest difference in performance (86.3% total accuracy); however, this is no big loss, considering it is the main feature. Hence, as most features do not bring a significant loss in accuracy, the features need to be combined to improve the performance.

6.2 Learning Analysis

The learning curve of Figure 1 (left) presents the increase of the performance obtained by Elliphant using the training data randomly ordered. The performance reaches its plateau using 90% of the training instances. Using different ordering of the training set we obtain the same result.

Figure 1 (right) presents the precision for each class and overall in relation to the number of training instances for each one of them. Recall grows similarly to precision. Under all conditions, subjects are classified with a high precision since the information given by the parser (collected in the features) achieves an accuracy of 74.9% for the identification of explicit subjects.

The impersonal construction class has the fastest learning curve. When utilizing a training set of only 163 instances (90% of the training data), it reaches a precision of 63.2%. The unstable behaviour for impersonal constructions can be attributed to not having enough training data for that class, since impersonals are not frequent in Spanish. On the other hand, the zero subject class is learned more gradually.

The learning curve for the explicit subject class is almost flat due to the great variety of subjects occurring in the training data. In addition, reaching a precision of 92.0% for explicit subjects using just 20% of the training data is far more expensive in terms of the number of training instances (978) as seen in Figure 1 (right). Actually, with just 20% of the training data we can already achieve a precision of 85.9%.

This demonstrates that Elliphant does not need very large sets of expensive training data and is able to reach adequate levels of performance when exploiting far fewer training instances. In fact, we see that we only need a modest set of

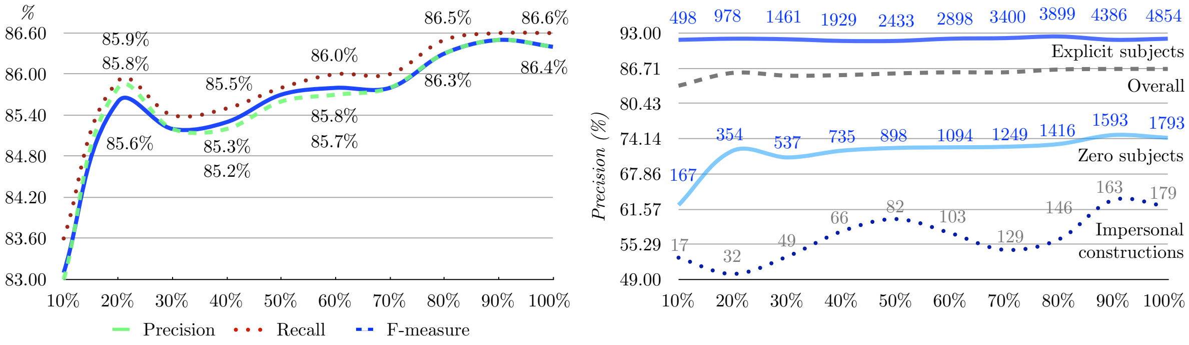

Figure 1: Learning curve for precision, recall and F-measure (left) and with respect to the number of instances of each class (right) for a given percentage of training data.

annotated instances (fewer than 1,500) to achieve good results.

6.3 Impact of Genre

To examine the influence of the different text genres on this method, we divided our training data into two subgroups belonging to different genres (legal and health) and analyze the differences.

A comparative evaluation using ten-fold cross-validation over the two subgroups shows that Elliphant is more successful when classifying instances of explicit subjects in legal texts (89.8% accuracy) than health texts (85.4% accuracy). This may be explained by the greater uniformity of the sentences in the legal genre compared to ones from the health genre, as well as the fact that there are a larger number of explicit subjects in the legal training data (2,739 compared with 2,116 in the health texts). Further, texts from the health genre present the additional complication of specialized named entities and acronyms, which are used quite frequently. Similarly, better performance in the detection of zero subjects and impersonal sentences in the health texts may be due to their more frequent occurrence and hence greater learnability.

Training/Testing	Legal	Health	All
Legal	90.0%	86.8%	89.3%
Health	86.8%	85.9%	88.7%
All	92.5%	93.7%	87.6%

Table 7: Accuracy of cross-genre training and testing evaluation (ten-fold evaluation).

We have also studied the effect of training the classifier on data derived from one genre and testing on instances derived from a different genre. Table 7 shows that instances from legal texts are more homogeneous, as the classifier obtains higher accuracy when testing and training only on legal instances (90.0%). In addition, legal texts are also more informative, because when both legal and health genres are combined as training data, only instances from the health genre show a significant increased accuracy (93.7%). These results reveal that the health texts are the most heterogeneous ones. In fact, we also found subsets of the legal documents where our method achieves an accuracy of 94.6%, implying more homogeneous texts.

6.4 Error Analysis

Since the features of the system are linguistically motivated, we performed a linguistic analysis of the erroneously classified instances to find out which patterns are more difficult to classify and which type of information would improve the method (Rello et al., 2011).

We extract the erroneously classified instances of our training data and classify the errors. According to the distribution of the errors per class (Table 8) we take into account the following four classes of errors for the analysis: (a) impersonal constructions classified as zero subjects, (b) impersonal constructions classified as explicit subjects, (c) zero subjects classified as explicit subjects, and (d) explicit subjects classified as zero subjects. The diagonal numbers are the true predicted cases. The classification of impersonal constructions is less balanced than the ones for explicit subjects and zero subjects. Most of the wrongly identified instances are classified as explicit subject, given that this class is the largest one. On the other hand, 25% of the zero subjects are classified as explicit subject, while only 8% of the explicit subjects are identified as zero subjects.

Class	**Zero subjects**	**Explicit subjects**	**Impers.**
Zero subj.	1327	453 **(c)**	13
Explicit subj.	368 **(d)**	4481	6
Impersonals	25 **(a)**	41 **(b)**	113

Table 8: Confusion Matrix (ten-fold validation).

For the analysis we first performed an exploration of the feature values which allows us to generate smaller samples of the groups of errors for the further linguistic analyses. Then, we explore the linguistic characteristics of the instances by examining the clause in which the instance appears in our corpus. A great variety of different patterns are found. We mention only the linguistic characteristics in the errors which at least double the corpus general trends.

In all groups (a-d) there is a tendency of using the following elements: post-verbal prepositions, auxiliary verbs, future verbal tenses, subjunctive verbal mode, negation, punctuation marks appearing before the verb and the preceding noun phrases, concessive and adverbial subordinate clauses. In groups (a) and (b) the lemma of the verb may play a relevant role, for instance verb *haber ('there is/are')* appears in the errors seven times more than in the training while verb *tratar ('to be about', 'to deal with')* appears 12 times more. Finally, in groups (c) and (d) we notice the frequent occurrence of idioms which include verbs with impersonal uses, such as *es decir ('that is to say')* and words which can be subject on their own *i.e. ambos ('both')* or *todo ('all')*.

7 Conclusions and Future Work

In this study we learn which is the most accurate approach for identifying explicit subjects and impersonal constructions in Spanish and which are the linguistic characteristics and features that help to perform this task. The corpus created is freely available online.[5] Our method complements previous work on Spanish anaphora resolution by addressing the identification of non-referential constructions. It outperforms current approaches in explicit subject detection and impersonal constructions, doing better than the parser for every class.

A possible future avenue to explore could be to combine our approach with Ferrández and Peral (Ferrández and Peral, 2000) by employing both algorithms in sequence: first Ferrández and Peral's algorithm to detect all zero subjects and then ours to identify explicit subjects and impersonals. Assuming that the same accuracy could be maintained, on our data set the combined performance could potentially be in the range of 95%.

Future research goals are the extrinsic evaluation of our system by integrating our system in NLP tasks and its adaptation to other Romance pro-drop languages. Finally, we believe that our ML approach could be improved as it is the first attempt of this kind.

Acknowledgements

We thank Richard Evans, Julio Gonzalo and the anonymous reviewers for their wise comments.

References

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

S. Bergsma, 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.

I. Bosque. 1989. Clases de sujetos tácitos. In Julio Borrego Nieto, editor, *Philologica: homenaje a Antonio Llorente*, volume 2, pages 91–112. Servicio de Publicaciones, Universidad Pontificia de Salamanca, Salamanca.

A. Boyd, 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.

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.

J.G. Cleary and L.E. Trigg. 1995. K*: an instance-based learner using an entropic distance measure. In *Proceedings of the 12th International Conference on Machine Learning (ICML-95)*, pages 108–114.

[5] ESZIC_es Corpus is available at: `http://luzrello.com/Projects.html`.

Connexor Oy, 2006. *Machinese language model.*

L. Danlos. 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.

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

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.

G. Hirst. 1981. *Anaphora in natural language understanding: a survey*. Springer-Verlag.

J. Hobbs. 1977. Resolving pronoun references. *Lingua*, 44:311–338.

R. Mitkov and C. Hallett. 2007. Comparing pronoun resolution algorithms. *Computational Intelligence*, 23(2):262–297.

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

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

C. Müller. 2006. Automatic detection of nonreferential *it* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 49–56.

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.

Real Academia Española. 2001. *Diccionario de la lengua española*. Espasa-Calpe, Madrid, 22 edition.

Real Academia Española. 2009. *Nueva gramática de la lengua española*. Espasa-Calpe, Madrid.

M. Recasens and E. Hovy. 2009. A deeper look into features for coreference resolution. In Lalitha Devi Sobha, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications. Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-09)*, pages 29–42. Springer, Berlin, Heidelberg, New York. Lecture Notes in Computer Science, Vol. 5847.

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, G. Ferraro, and A. Burga. 2011. Error analysis for the improvement of subject ellipsis detection. *Procesamiento de Lenguaje Natural*, 47:223–230.

L. Rello. 2010. Elliphant: A machine learning method for identifying subject ellipsis and impersonal constructions in Spanish. Master's thesis, Erasmus Mundus, University of Wolverhampton & Universitat Autònoma de Barcelona.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 64–71.

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

S. Zhao 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.