TO BE OR NOT TO BE A ZERO PRONOUN: A MACHINE LEARNING APPROACH FOR ROMANIAN

Claudiu MIHĂILĂ*, Iustina ILISEI**, and Diana INKPEN***

*Faculty of Computer Science, "Al.I. Cuza" University of Iaşi
**Research Institute in Information and Language Processing, University of Wolverhampton
***School of Information Technology and Engineering, University of Ottawa
E-mail: claudiu.mihaila@info.uaic.ro, iustina.ilisei@gmail.com, diana@site.uOttawa.ca

Abstract: This paper presents a new study on the distribution and identification of zero pronouns in Romanian. A Romanian corpus that includes legal, encyclopaedic, literary, and news texts has been created and manually annotated for zero pronouns. Using a morphological parser for Romanian and machine learning methods, experiments have been performed on the created corpus for the identification of verbs which have a zero pronoun in the subject position.

The evaluation results highlight that zero pronouns appear frequently in Romanian, and their distribution depends largely on the genre. Additionally, a search scope for the antecedent has been determined, increasing the chances of correct resolution. Furthermore, more than 70% of the zero pronouns have been accurately identified by various machine learning algorithms. The strong similarity between our results and those obtained for other Romance languages support our conclusions.

Key words: zero pronoun, ellipsis, anaphora, Romanian, machine learning.

1. INTRODUCTION

In natural language processing (NLP), coreference resolution is the task of determining whether two or more noun phrases have the same referent in the real world (Mitkov, 2002). This task is extremely important in discourse analysis, since many natural language applications benefit from a successful coreference resolution. NLP sub-fields such as information and terminology extraction (Mihăilă & Mekhaldi, 2009), question answering, automatic summarisation, machine translation, or generation of multiple-choice test items (Mitkov et al., 2006) are conditioned by the correct identification of coreferents.

Zero pronoun identification is one of the first steps towards coreference resolution and, moreover, a fundamental task for the development of pre-processing tools in NLP. Furthermore, the resolution of zero pronouns improves significantly the performance of more complex systems, which rely on anaphora resolution.

This paper is structured as follows: section 2 contains a description of subject ellipsis occurring in Romanian. Section 3 highlights some of the recent works in zero pronoun identification for several languages, including Romanian. In section 4, the corpora on which this

work was performed are described, and in section 5 the results of the evaluation are presented and discussed. Section 6 includes some type of errors that influence negatively the classifiers' performance.

2. ZERO SUBJECTS vs. ZERO PRONOUNS

The definition of ellipsis in the case of Romanian is not very clear and a consensus has not yet emerged. Many different opinions and classifications of ellipsis types exist, as is reported by Mladin (2005). Despite the existing controversy, in this work we adopt the theory that follows.

Two types of elliptic subjects are found in Romanian: implicit subjects and zero subjects. The difference between these two types is that whilst the former can be lexically retrieved, such as in example (a), the latter cannot, as shown in example (b).

(a) $_{zp}$*[Noi]*[1] mergem la şcoală.
[We] are going to school.

(b) ∅ Ninge.
[It] is snowing.

In Romanian, the clauses with zero subjects are considered syntactically impersonal, whereas the implicit or omitted subjects, which are not phonetically realised, can be lexically recovered from the inflection of the verb (Popescu, 2009).

The main classes of impersonal and impersonally used verbs which take zero subjects are exemplified in what follows. The examples' translation into English may sound forced, but was performed as such in order to provide a better understanding of the phenomenon. Some of these classes refer to weather and temporal changes, and are found in other Romance languages, such as Spanish. Other classes are similar to Slavic languages (e.g., Russian), where the logical subject is expressed as a dative construction, leaving the verb phrase without a subject.

- Meteorological phenomena:
  ∅ S-a înnorat dimineaţă.
  [It] clouded over this morning.
- Changes in the moments of the day:
  ∅ Se luminează de ziuă la ora cinci.
  [It] is dawning at five o'clock.
- Impersonal expressions with dative:
  ∅ Îmi pare rău de tine. Azi ∅ nu-mi arde de glumă.
  [It] feels sorry to me for you. Today [it] does not feel like joking to me.
- Impersonal constructions with verbs *dicendi*[2]:
  ∅ Se vorbeşte despre el.
  [People] are talking about him.
- Romanian impersonal constructions with personal verbs when preceded by the reflexive pronoun *se*:
  ∅ Se lucrează aici.
  [People] are working here.

On the other hand, an implicit subject does not need to be overtly expressed, due to the fact that the inflection of the verb and the context provide all the necessary information for its understanding. The gap (or zero anaphor) in the sentence that refers to the entity which provides the necessary

[1] From this point forward, we denote by $_{zp}$*[]* a zero pronoun (e.g., implicit subject), whereas a zero subject will be marked using the ∅ sign.

[2] Latin for verbs of saying.

information for the correct understanding of the gap is also known as a zero pronoun (ZP) (Mitkov, 2002). Although many different forms of zero anaphora (or ellipsis) have been identified (e.g., noun anaphora, verb anaphora), this study focusses only on zero pronominal anaphora, which occurs when an anaphoric pronoun is omitted but nevertheless understood.

An anaphoric zero pronoun (AZP) results when the zero pronoun corefers to one or more overt nouns, noun phrases, or clauses in the text. In a similar manner to coreferential noun phrases, coreferential zero pronouns can be divided in anaphoric or cataphoric, depending on the position of its referred noun phrase. Furthermore, zero pronouns may be exophoric, meaning that the referent is not found in the text, but in the real world.

The difficulty that arises in the task of identifying zero pronouns is to distinguish between the personal and impersonal use of verbs, since in Romanian personal verbs can also be used impersonally. Whilst the impersonally used verbs may take zero subjects (thus having no associated ZP), personally used verbs need a subject, which in turn can be explicit or implicit. Moreover, there are cases when impersonal verbs have subjects, fact which increases the complexity. For instance, the two examples below reveal a type of ambiguity which may appear in the case of a verbal expression with the same meaning but in different contexts.

(a) ∅ E greu pentru tine.
[It] is difficult for you.
(b) E greu [să scrii versuri].
[It] is difficult to write lyrics.
(c) Tu nu poți ridica patul. $_{zp}$*[El]/* ∅ E greu pentru tine.
You cannot lift the bed up. [It] is difficult/heavy for you.

Example (a) contains an impersonal verbal expression which has a zero subject. In contrast, example (b) shows the same expression having the nominal clause between brackets as its subject. Nevertheless, when in another context, the same expression in example (c) has a zero pronoun referring to either the bed (case when the translation would be *heavy*), or the action of lifting the bed up (case when the translation would be *difficult*). Moreover, it may be possible to interpret the second sentence as impersonal, being expressed as a general conclusion drawn from the impossibility of lifting the bed up. Therefore, the system may encounter classifying problems because of the ambiguity.

3. RELATED RESEARCH

In the existing literature, a large part of the studies on coreference resolution is dedicated to English. Even publicly available corpora created especially for this task are available mostly for English, e.g., at the Message Understanding Conferences[3] (Chinchor, 1998).

A hand-engineered rule-based approach to identify and resolve zero pronouns that are in the subject grammatical position in Spanish is proposed by Ferrández & Peral (2000). In their study, the verbs tagged with a ZP are identified as those not having a noun phrase or pronoun on the left-hand side, provided that they are not imperative or impersonal. Furthermore, Rello & Ilisei (2009a, 2009b) create a Spanish corpus annotated with more than 1200 ZPs and complement the previous studies by considering the detection of impersonal clauses using hand-built rules; the reported F-measure is 57%.

For Chinese, a machine learning approach which automatically identifies and resolves zero pronouns is described by Zhao & Ng (2007), and their results are comparable to the ones obtained by a heuristic rule-based approach by Converse (2006). The authors make use of parse trees to compute the feature vectors for the ZP candidates and for their antecedents, and obtain a value of

[3] MUC6 and MUC7, http://www-nlpir.nist.gov/related_projects/muc/

26% for the F-measure. Other languages that have been more intensively and recently studied are Portuguese (Pereira, 2009), Japanese (Iida et al., 2006), and Korean (Kim, 2000; Han, 2006).

In contrast, fewer studies have been performed for the coreference resolution in Romanian. A data-driven SWIZZLE-based system for multilingual coreference resolution is presented by Harabagiu & Maiorano (2000). They use an aligned English-Romanian corpus to resolve coreferences, exploiting language differences to reduce uncertainty regarding the antecedents. The obtained results have a precision of 76% and a recall of 70%, which are better than using a monolingual corpus. Another study on a rule-based Romanian anaphora resolution system relying on RARE (Cristea et al., 2002) has been reported by Pavel et al. (2006).

4. CORPORA

This section describes the corpora compiled for this study. In the first subsection, details about the annotation are provided, whilst in the second subsection some statistics regarding the distribution of zero pronouns in the corpora are included.

The genres of the documents which were included in the study are law (LT), newswire (NT), encyclopaedia (ET), and literature (ST). The law part of the corpus contains the Romanian constitution, and the newswire texts represent international news published in the beginning of 2009. The encyclopaedic corpus comprises articles from the Romanian Wikipedia[4] on various topics, whilst the literary part is composed of children's short stories by Emil Gârleanu and Ion Creangă.

The important contribution of this study is two-fold: the selection of genres which are likely to be subject for several NLP applications (e.g., multiple choice tests generation, question answering), and all four genres are manually annotated with the anaphoric zero pronouns information.

The following subsection provides the annotation setup, and some statistics regarding the distribution of zero pronouns are presented in the second subsection.

4.1. Annotation

The documents comprised in the corpora were parsed automatically using the web service[5] published by the Research Institute for Artificial Intelligence, part of the Romanian Academy. This parser provides the lemma and the morphological characteristics regarding the tokens.

The texts were afterwards manually annotated for zero pronouns by two authors, in order to create a gold standard. Zero pronouns were manually identified by the addition of the following empty XML tag containing the necessary information as attributes into the parsed text:

```
<ZERO_PRONOUN id="w152.5" ant="w136" depend_head="w153"
confidence="high" sentence_type="main" />
```

Each `ZERO_PRONOUN` tag includes various pieces of information regarding its antecedent (the `ant` attribute), the verb it depends on (the `depend_head` attribute), and the type of sentence it appears in (the `sentence_type` attribute). The attribute corresponding to the antecedent may have one of three types of values:

(i) *elliptic*, if the ZP has no antecedent in the text,
(ii) *non_nominal*, if the antecedent is a clause, or
(iii) a reference which points back to the antecedent, in the case of an AZP.

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

[4] http://ro.wikipedia.org/
[5] http://www.racai.ro/webservices/

pronoun appears, one more attribute (sentence type) provides the information of the kind of sentence (main, coordinated, subordinated, etc.).

4.2. Statistics

The currently gathered corpus comprises almost 50000 tokens and almost 800 zero pronouns, as shown in Table 1. Nevertheless, it can be noticed from the table that the legal and literary texts have a very low and a very high, respectively, density of ZPs per sentence. This is due to the style of the writings, in which either to avoid possible misinterpretations, or to increase the fluency of narrative sequences, the authors adjust the use of zero pronouns.

Table 1: Description of the corpora.

Corpus	ET	LT	ST	NT	Overall
No. of tokens	12963	13739	3391	18690	48783
No. of sentences	574	790	253	816	2433
No. of ZP	172	113	251	245	781
Avg. tokens/sentence	22.58	17.39	13.40	22.90	20.05
Avg. ZP/sentence	0.30	0.14	0.99	0.30	0.31

Table 2 offers the number of zero pronouns as they appear in four different clause types, main clauses (MC), juxtaposed clauses (JC), coordinated clauses (CC), and subordinated clauses (SC). Most of the ZPs are found in subordinated clauses, whilst juxtaposed clauses contain the least number of ZPs. This fact is easily explained by considering that there is no need to repeat the subject of the main clause in the secondary clause, provided that the two clauses have the same subject. However, exceptions occur when the author desires to emphasise the subject more than the action.

Moreover, it can be observed that the newswire texts contain a significant number of ZPs in subordinated clauses, whilst the majority of ZPs in the main clause are found in the children's stories. This use of zero pronouns is specific to the types of writings, whether to create more complex sentences, showing causes, effects, or explanations, or to express the facts in a simple manner, using simple sentences. The zero pronouns in legal texts are contained mostly in coordinated clauses, since it is usual for the same subject to perform multiple actions, linked together by coordinating conjunctions. The encyclopaedic genre is not as specific as the other three, and does not have, in consequence, outstanding values.

Table 2: Number of ZPs s by clause type.

Clause type	ET	LT	ST	NT	Overall
Main	48	19	103	28	198
Juxtaposed	8	6	26	3	43
Coordinated	44	50	42	40	176
Subordinated	72	38	80	174	364

The distribution of distances from the zero pronouns to their antecedents for each of the genres in the studied corpora is provided in Table 3. The distances from the zero pronouns to their antecedent in the case of newswire and literature texts reveal unique values. Longer distances are specific to narrative sequences, where multiple sentences have the same subject. On the other hand, a short distance is specific to a rapid change in the subject of the sentences, such as in news articles. However, the distance to the dependent verb is constant throughout the corpora, which is on

average at 1.69 tokens away. This distance is due to the existence of pronouns, conjunctions, adverbs, or combinations of these which precede the verb.

Table 3: Distances between the ZP and its antecedent and dependent verb for each genre.

Corpus	ET	LT	ST	NT	Overall
Antecedent (sentences)	1.32	1.07	3.77	0.02	1.54
Antecedent (tokens)	34.08	38.64	58.02	7.79	34.63
Dependent verb (tokens)	1.47	1.54	1.98	1.77	1.69

(a) Pronouns:

[...] Napoleon rămâne cu armata [...] şi $_{zp}$*[el]* **îşi** concentrează [...]
[...] Napoleon remains with the army [...] and $_{zp}$*[he] concentrates [...]*

[...] pe care $_{zp}$*[ei]* **l**-au denumit "fat-man factor A".
[...] which $_{zp}$*[they] named "fat-man factor A".*

(b) Conjunctions:

[...] Gruevski a cerut tuturor preşedinților [...] $_{zp}$*[ei]* **să** acționeze [...]
[...] Gruevski asked all the presidents [...] $_{zp}$*[they] to act [...]*

(c) Adverbs:

[...] francezii [...] lansează o violentă ofensivă, dar $_{zp}$*[ei]* **nu** pot disloca [...]
[...] the French [...] launch a violent offensive [...] but $_{zp}$*[they] cannot dislocate [...]*

Considering that no previous study has been undertaken for the Romanian language, we note that the results for the encyclopaedic and legal texts can be compared to the ones obtained for another Romance language, Spanish, in Rello & Ilisei (2009a). The differences are not considerably significant and prove the consistence of the distribution within the same language family.

Furthermore, in Table 4, the distances to the antecedent and dependent verb in the various clause types are included. In subordinated clauses, zero pronoun antecedents tend to be fairly close – they are rarely found outside the same sentence, whilst zero pronouns in main sentences are longer-distance anaphors, whose antecedents tend to be the subject of some of the previous sentences.

Table 4: Distances between the ZP and its antecedent and dependent verb for each clause type.

Distance	MC	JC	CC	SC	Overall
Antecedent (sentences)	1.32	1.07	3.77	0.02	1.54
Antecedent (tokens)	34.08	38.64	58.02	7.79	34.63
Dependent verb (tokens)	1.47	1.54	1.98	1.77	1.69

Table 5 contains the distribution over the part of speech for the ZPs' antecedents. As it can be observed, most of the antecedents are nouns (over 81%), whilst a small fraction are pronouns (around 19%); antecedents with other parts of speech have not been found. The difference between the overall numbers for each corpus genre here and the number of zero pronouns in Table 1 is represented by zero pronouns which have an elliptical or non-nominal antecedent.

Table 5: Part of speech distribution of ZP antecedents.

Corpus	ET	LT	ST	NT	Overall
Noun	157	108	166	212	643
Pronoun	15	5	85	33	138
Overall	172	113	251	245	781

This part-of-speech distribution is important due to the fact that it gives an insight into which are the most interesting candidates for the resolution of zero pronouns that need to be considered. Furthermore, the significantly restricts the search space from all the words in the sentences part of the search scope to only a few.

5. EVALUATION

The goal is to classify the verbs into two distinct classes: having or not a zero pronoun. The chosen method in this study is supervised machine learning, by using the Weka[6] application (Hall et al., 2009; Witten & Frank, 2005). Therefore, a feature vector was constructed for the verbs. The vector is composed of the following ten elements:

1. *type* – the type of the verb (i.e., main, auxiliary, copulative, or modal);
2. *mood* – the mood of the verb (indicative, subjunctive, etc.);
3. *tense* – the tense of the verb (present, imperfect, past, pluperfect);
4. *person* – the person of the conjugation (first, second, or third);
5. *number* – the number of the conjugation (singular or plural);
6. *gender* – the gender of the conjugation (masculine, feminine, or neuter);
7. *clitic* – whether the verb appears in a clitic form or not;
8. *impersonality* – whether the verb is strictly impersonal or not (such as meteorological verbs);
9. *'se'* – whether the verb is preceded by the reflexive pronoun “se” or not;
10. *hasZP* – whether the verb has a ZP or not.

The first seven elements of the feature vector are extracted from the morphological parser's output and represent all the morphological characteristics of a Romanian verb, whilst the next two elements are computed automatically based on the annotated texts. The last item is the class whose values are true if the verb allows zero pronoun and false otherwise, and it is used only for training purposes. When in test mode the class is not used, except when computing the evaluation measures.

The data set on which the experiments were performed includes 1572 instances of the feature vector. Half of these instances correspond to the 781 verbs which have an associated ZP, whilst the other half contains randomly selected verbs without a ZP. As the baseline classifier employed takes the majority class, the baseline to which we need to compare our accuracy is 50%.

Multiple classifiers pertaining to different categories have been experimented with. The results that follow are obtained by 10-fold cross validation on the data. Precision, Recall and F-measure for each of the classes of verbs and the accuracy for three classifiers (SMO, JRip, and J48) and one meta-classifier (Vote) are included in Table 6. The SMO (acronym for Sequential Minimal Optimisation) classifier is the implementation of support vector machines (SVM), J48 is an implementation of decision trees, and JRip is an implementation of a propositional rule learner, namely Repeated Incremental Pruning to Produce Error Reduction (RIPPER). The Vote meta-classifier is configured to consider the three previous classifiers, using a Majority Voting combination rule.

Table 6: Classifier results for the classes of verbs.

[6] http://www.cs.waikato.ac.nz/ml/weka/

Classifier	Accuracy	has ZP			not ZP		
		P	R	F_1	P	R	F_1
SMO	0.739	0.684	0.889	0.773	0.841	0.590	0.694
JRip	0.733	0.709	0.793	0.748	0.765	0.675	0.717
J48	0.720	0.698	0.777	0.735	0.749	0.663	0.703
Vote	0.733	0.705	0.802	0.750	0.770	0.665	0.713

The results may vary slightly, since only a subset of verbs with no ZP was selected. Nevertheless, repetitions of the experiment with different test datasets produced similar values. As observed, the Vote meta-classifier does not improve the results, which leads us to the conclusion that the three classifiers make relatively the same decisions.

```
mood = indicative
|   person = 1st: true (54.75/4.0)
|   person = 2nd: true (9.03/1.0)
|   person = 3rd
|   |   tense = present
|   |   |   se = false: false (387.16/181.3)
|   |   |   se = true: true (57.19/26.4)
|   |   tense = imperfect
|   |   |   number = singular: true (71.54/25.47)
|   |   |   number = plural: false (18.09/7.02)
|   |   tense = past: true (175.47/57.66)
|   |   tense = pluperfect
|   |   |   type = main: true (26.0/3.0)
|   |   |   type = auxiliary: false (8.08/3.46)
|   |   |   type = modal: true (0.0)
|   |   |   type = copulative: true (0.0)
mood = subjunctive: true (315.91/58.94)
mood = imperative: true (10.03/2.98)
mood = infinitive: false (119.09/15.43)
mood = participle
|   type = main: false (210.0/1.0)
|   type = auxiliary
|   |   person = 1st: true (3.24/0.57)
|   |   person = 2nd: true (0.43/0.08)
|   |   person = 3rd
|   |   |   se = false: false (58.75/22.09)
|   |   |   se = true: true (5.88/1.96)
|   type = modal: false (0.0)
```

```
|   type = copulative: false (0.0)
mood = gerund: false (31.34/3.27)
```

Fig. 1 The J48 pruned decision tree.

In order to observe the rules according to which the decisions are made to classify the verbs, we show in Fig. 1 the pruned decision tree produced from the J48 classifier. The most important attribute is clearly the mode of the verb, followed by the person, tense, and type attributes. Oppositely, the gender and the clitic form do not appear at all. The numbers displayed on the leaf nodes represent the number of instances which pertain to that specific path, and the number of instances which do not, respectively. The decimal values instead of integer ones are explained by the fact that some instances in the dataset have missing values for some features.

Furthermore, in order to have a deeper analysis, we looked at the output of the JRip classifier. The features obtained in the rules are similar to those appearing in the decision tree from J48, fact that is supported also by the similar values of the various scores.

Fig. 2 depicts the receiver operating characteristic (ROC) curves of the J48 classifier, for the two classes of verbs, i.e. TRUE for having a zero pronoun, and FALSE for not having a zero pronoun. The two curves are very similar in shape, fact that tells us that the classifier behaves comparably with regard to the two distinct classes. The true positive (TP) rate augments rapidly for a small false positive (FP) rate, but the increase is diminished at TP rates of over 0.6-0.7. The curves for the other classifiers are similar.

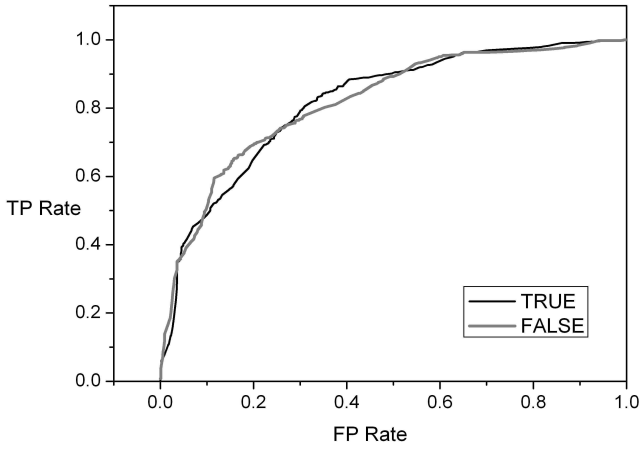

Fig. 2 ROC curves of the J48 classifier for verbs having a ZP (TRUE), and verbs not having a ZP (FALSE).

Aiming at determining which features influence the most the verb classification, regardless of the classifying algorithm, two attribute evaluators have provided the results shown in Table 4. As it can be observed, the most informative feature is the mode of the verb, followed by the “se” reflexive pronoun and the person of the verb. At the other end, the number and gender of the verb influence the classification the least. Whether the verb is in a clitic form or not does not seem to have any effect on the classification. The most relevant attributes are those used also by the JRip and J48 classifiers.

As expected, the most problematic case is that of the present indicative verbs in the third person and preceded by the reflexive pronoun “se”. A reason for this effect is the fact that “se” is part of the impersonal constructions which may or may not have zero pronouns. As a result, the system classifies incorrectly the verbs.

Table 4: Attribute selection output from two attribute evaluators.

Attribute	ChiSquare	InfoGain
Mood	402.546	0.206
'Se'	25.719	0.012
Person	21.217	0.01
Impersonality	12.092	0.007
Tense	9.371	0.004
Type	2.577	0.001
Number	0.354	1E-4
Gender	7E-4	3E-7
Clitic	0	0

6. ERROR ANALYSIS

The two main types of factors involved in the misclassified instances of the learning system that have been identified are the following: errors which appear due to the complexity posed by the task itself, and errors which appear at the pre-processing stage.

The language model poses certain difficulties when dealing with language ambiguities due to the fact that, for instance, one expression may have either a zero subject or an explicit one, depending on the context. Compound subjects, the verb-subject number disagreement, or the unexpected verb inflections are other causes for the misclassifications of the learning system.

- Same expression with zero subject and explicit subject:
  ∅ Este greu pentru tine.
  [It] is difficult for you.
  Este greu **a scrie** versuri.
  [It] is difficult to write lyrics.
- Verb-subject number disagreement:
  O **mulţime** de fete v**in** şi $_{zp}$*[ea]* cânt**ă**.
  A multitude of girls come and $_{zp}$*[it] sing.*
- Compound subjects:
  Bărbatul şi femeia vin şi $_{zp}$*[ei]* pleacă.
  The man and the woman come and $_{zp}$*[they] go.*
- Ambiguous verb inflections:
  Maria şi Ion cânt**ă** şi $_{zp}$*[ei]* dans**ează**.
  Mary and John sing and $_{zp}$*[they] dance.*

Moreover, at the pre-processing stage, there are several cases when morphological classes are confused: some nouns and pronouns (e.g., *bosniac*, *retinal*, *etnic*, *ce*) are parsed as adjectives. In addition, some proper nouns (such as *Elmer* or *Mussolini*) can be found annotated in various ways, either as a proper noun, a common noun or even as an adjective.

7. CONCLUSIONS AND FUTURE WORK

This paper presents a new study on the distribution and identification of zero pronouns in Romanian. By compiling and manually annotating a multiple-genre corpus, zero pronouns are identified using supervised machine learning algorithms, and the results are comparable to those obtained for other Romance languages. Moreover, the position of the ZP has been investigated, as well as the scope in which the antecedent may be found.

After the identification stage, the future research direction that will be undertaken is the resolution of the anaphoric zero pronouns. A successful completion of this task will improve significantly the results of other NLP applications that rely on correct anaphora resolution. For instance, the investigation of translation universals may benefit from the correct identification of the zero pronouns in translated texts. Moreover, since the distribution of zero pronouns depends largely on the genre, it is possible for it to depend on the author as well. Therefore, automatic zero pronoun identification may be used in plagiarism and authorship detection.

REFERENCES

1. CHINCHOR, N. *Proceedings of the Seventh Message Understanding Conference*, Science Applications International Corporation (SAIC), 1998.

2. CONVERSE, S. *Pronominal anaphora resolution in Chinese*, PhD Thesis, Philadelphia, PA, USA, 2006.

3. CRISTEA, D., POSTOLACHE, O., DIMA, G., BARBU, C. AR-Engine – a framework for unrestricted co-reference resolution. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 2000–2007, 2002.

4. FERRÁNDEZ, A., and PERAL, J. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL '00)*, pp. 166–172, 2000.

5. HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., and WITTEN, I.H. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, 11(1), pp. 10-18, 2009.

6. HAN, N.-R. *Korean zero pronouns: analysis and resolution*, PhD Thesis, Philadelphia, PA, USA, 2006.

7. HARABAGIU, S., and MAIORANO, S. Multilingual coreference resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP 2000)*, pp. 142–149, 2000.

8. IIDA, R., INUI, K., and MATSUMOTO, Y. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL 2006)*, pp. 625–632, 2006.

9. KIM, Y.-J. Subject/object drop in the acquisition of Korean: A cross-linguistic comparison. *Journal of East Asian Linguistics*, 9(4), pp. 325–351, 2000.

10. MIHĂILĂ, C., and MEKHALDI, D. Bimodal corpora terminology extraction: another brick in the wall. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pp. 236–240, 2009.

11. MITKOV, R., HA, L.A., and KARAMANIS, N. A computer-aided environment for generating multiple-choice test items. *Journal of Natural Language Engineering*, 12(2), pp. 177–194, 2006.

12. MITKOV, R. *Anaphora resolution*, Longman, 2002.

13. MLADIN, C.I. Procese şi structuri sintactice "marginalizate" în sintaxa românească actuală. Consideraţii terminologice din perspectivă diacronică asupra contragerii - construcţiilor - elipsei, *The Annals of Ovidius University Constanța - Philology*, 16, pp. 219–234, 2005.

14. PAVEL, G., POSTOLACHE, O., PISTOL, I., and CRISTEA, D. Rezoluţia anaforei pentru limba română, *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, 2006.

15. PEREIRA, S. ZAC.PB: An annotated corpus for zero anaphora resolution in Portuguese. In *Proceedings of the Student Workshop at RANLP 2009*, pp. 53–59, 2009.
16. POPESCU, ȘȘ. *Gramatica practică a limbii române – ediția a XV-a*, TEDIT FZH, 2009.
17. RELLO, L., and ILISEI, I. A comparative study of Spanish zero pronoun distribution. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, 2009.
18. RELLO, L., and ILISEI, I. A rule based approach to the identification of Spanish zero pronouns. In *Proceedings of the Student Workshop at RANLP 2009*, pp. 60–65, 2009.
19. WITTEN, I.H. and FRANK, E. *Data Mining: practical machine learning tools and techniques (second edition)*, Morgan Kaufmann, 2005.
20. ZHAO, S. and NG, H.T. 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*, pp. 541–550, 2007.