Romanian Zero Pronoun Distribution:
A Comparative Study

Claudiu Mihăilă[1], **Iustina Ilisei**[2], **Diana Inkpen**[3]

[1] Faculty of Computer Science, ”Al.I. Cuza” University of Iaşi,
16, General Berthelot Street, Iaşi 700483, Romania
claudiu.mihaila@info.uaic.ro

[2] Research Institute in Information and Language Processing, University of Wolverhampton,
Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom
iustina.ilisei@gmail.com

[3] School of Information Technology and Engineering, University of Ottawa,
800, King Edward Street, Ottawa ON K1N 6N5, Canada
diana@site.uOttawa.ca

Abstract

Anaphora resolution is still a challenging research field in natural language processing, lacking an algorithm that correctly resolves anaphoric pronouns. Anaphoric zero pronouns pose an even greater challenge, since this category is not lexically realised. Thus, their resolution is conditioned by their prior identification stage. This paper reports on the distribution of zero pronouns in Romanian in various genres: encyclopaedic, legal, literary, and news-wire texts. For this purpose, the RoZP corpus has been created, containing almost 50000 tokens and 800 zero pronouns which are manually annotated. The distribution patterns are compared across genres, and exceptional cases are presented in order to facilitate the methodological process of developing a future zero pronoun identification and resolution algorithm. The evaluation results emphasise that zero pronouns appear frequently in Romanian, and their distribution depends largely on the genre. Additionally, possible features are revealed for their identification, and a search scope for the antecedent has been determined, increasing the chances of correct resolution.

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 and 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 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. For instance, in the case of multiple-choice test item generation for Romanian, language specific techniques are required, additional to the ones used in English. This is due to the flexibility of Romanian grammar, which allows verbs to take zero pronouns. Since the choices of the test items are usually the subjects of sentences, it is necessary to correctly identify and resolve the zero pronouns.

This study offers an insight into the distribution patterns of zero pronouns in Romanian. Based on this data, it becomes easier to develop an algorithm for zero pronoun identification and resolution in Romanian.

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 and considerations about zero pronouns and their importance in Romanian. In section 4, the corpus created for the analysis of the distribution of zero pronouns is described, and in section 5 statistics are presented. Finally, issues that might arise in the resolution of zero pronouns are presented and discussed in section 6.

2. Zero subjects and zero pronouns

The definition of ellipsis in the case of the Romanian language is not very clear and a consensus has not yet emerged. Several different opinions and classifications of ellipsis types exist, as is reported by Mladin (2005). In spite of 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 in example (b).

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

(b) ⊘ Plouă.
[It] is raining.

[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.

In Romanian, clauses with zero subject are considered syntactically impersonal, whereas implicit or omitted subjects, which are not phonetically realised, can be retrieved lexically (Popescu, 2009).

A zero subject is found in clauses whose verbs do not require a subject. Despite the fact that this phenomenon is not present in English, it is found frequently in Romanian and many other languages, such as Spanish, Chinese, and Japanese. Nevertheless, it is often the case that the subject is present, but not explicitly realised. However, this implicit subject is understood from the context, and it is usually encoded in the inflection of the verb.

A zero pronoun (ZP) is the gap (or zero anaphor) in the sentence that refers to an entity which provides the necessary information for the gap's correct understanding. 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 occur when an anaphoric pronoun is omitted but nevertheless understood (Mitkov, 2002).

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

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., the Message Understanding Conferences[2] (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 and 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, in (Rello and Ilisei, 2009a; Rello and Ilisei, 2009b), the authors 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 0.57.

For Chinese, a machine learning approach which automatically identifies and resolves zero pronouns is described by Zhao and 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 26% for the F-measure.

Other languages that have been more intensively 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 and Maiorano (2000). They use an aligned English-Romanian corpus to resolve coreferences and the obtained results have a precision of 76% and a recall of 70%. 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. RoZP Corpora: Description and Annotation Scheme

The genres of the documents which were included in the study are newswire (NT), encyclopaedia (ET), law (LT) and literature (ST). The newswire texts represent international news published in the beginning of 2009, and a section of the Romanian Constitution forms the legal part of the corpus. The encyclopaedic corpus comprises articles from the Romanian Wikipedia 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 of several NLP applications (e.g., multiple choice text generation, question answering), and all four genres are manually annotated with the anaphoric zero pronouns information.

The documents contained in the corpora were parsed automatically using the web service published by the Research Institute for Artificial Intelligence[3], part of the Romanian Academy. This parser provides the lemma and the morphological characteristics regarding the tokens.

A zero pronoun was afterwards manually identified by the addition of an empty XML tag containing the necessary information as attributes into the parsed text. Such a tag is exemplified in Figure 1.

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

Figure 1: Empty XML tag marking a zero pronoun.

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 there is no antecedent,

(ii) *non_nominal*, if the antecedent is a clause, or

(iii) a number which points back to the antecedent.

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

[3] http://www.racai.ro/webservices/

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

The `confidence` attribute represents the annotator's confidence regarding that specific positioning of the zero pronoun in the text; it can have two values, high and low.

The texts were manually annotated for zero pronouns by two human judges, in order to create a gold standard. The agreement between the annotations of the two judges is high, reaching up to over 98% in the case of determining whether a verb has or does not have a zero pronoun. Moreover, to exclude the possibility that the two judges annotate similarly by chance, Cohen's kappa coefficient was computed; the obtained value is of over 90%. Regarding the position of the ZP in the text, the agreement is slightly lower, of 90%. This is due to the flexibility of Romanian grammar, which allows various word orderings. However, this latter agreement is not significant, since the position of the ZP in the text is neither relevant for its resolution, nor for the semantics of the sentence.

5. Zero Pronoun Distribution

The currently gathered corpus comprises almost 50000 tokens and almost 800 zero pronouns, as shown in Table 1. Furthermore, the table proves that zero pronouns are found relatively frequently in Romanian, with 0.32 ZPs per sentence. Nevertheless, it can be noticed 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 2 offers the number of zero pronouns as they appear in four different clause types. 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.

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

Table 2: Number of ZPs by clause type

The distribution of distances from the zero pronouns to their antecedents in the studied corpora is provided in Table 3. The distances from the zero pronoun to its antecedent in the case of newswire and literature texts reveal unique values. 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. However, the distance to the dependent verb is constant throughout the corpora; on average 1.60 tokens away. This distance is due to the existence of pronouns (example (a)), conjunctions (example (b)), adverbs (example (c)), or combinations of these, which must precede the verb.

(a) Pronoun:
[...] 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) Conjunction:
[...] Gruevski a cerut tuturor preşedinţilor [...] $_{zp}$*[ei]* **să** acţioneze [...]
[...] Gruevski asked all the presidents [...] $_{zp}$*[they] to act [...]*

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

Corpus	Antecedent (sentences)	Antecedent (tokens)	Dependent verb (tokens)
NT	0.02	7.79	1.77
ET	1.17	32.60	1.56
LT	1.07	38.55	1.53
ST	5.37	67.88	1.55
Overall	1.90	36.70	1.60

Table 3: Distances between the ZP and its antecedent and dependent verb

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

Considering that no previous study has been undertaken for the Romanian language, the results for the encyclopaedic

Overview	NT	ET	LT	ST	Overall
No. of tokens	18690	12963	13739	3391	48783
No. of sentences	816	574	790	253	2433
No. of ZP	245	172	113	251	781
Avg. tokens/sentence	22.90	22.58	17.39	13.40	20.05
Avg. ZP/sentence	0.30	0.30	0.14	0.99	0.32

Table 1: Description of the corpora

and legal texts can be compared to the ones obtained for another Romance language, Spanish. Rello and Ilisei (2009a) report similar values and conclusions. The differences are not considerably significant and prove the constancy of the distribution within the same language family.

6. Constraints for Future Zero Pronoun Identification

One possible baseline for the identification of ZP could be gathering the set of all potential anaphoric zero pronouns. This set could be compiled by selecting the clauses in which the verb does not have a subject depending on it. The lack of subject in a clause makes it a likely candidate to contain an anaphoric zero pronoun.

However, this baseline rule introduces a set of false positives candidates, which will result in a high recall but in a low precision. To discard a part of these false positives, the set of candidates is refined by applying some constraints. These constraints, exemplified in what follows, exclude verbs and verbal expressions which actually take zero subjects instead of zero pronouns.

(a) Meteorological phenomena:
⊘ *S-a înnorat* dimineaţă.
[It] clouded over this morning.

(b) Changes in the moments of the day:
⊘ *Se luminează* de ziuă la ora cinci.
[It] dawns at five o'clock.

(c) Impersonal expressions:
⊘ *E bine* pentru noi. Azi ⊘ *nu-mi arde* de glumă.
[It] is good for us. Today [it] doesn't feel like joking to me.

(d) Impersonal constructions with verbs *dicendi*:
⊘ *Se vorbeşte* despre el.
[People] are talking about him.

(e) Romanian impersonal constructions with the reflexive pronoun "se":
⊘ *Se lucrează* aici.
[People] are working here.

However, ambiguous cases will still exist, and they can confuse the rules and classifiers. For instance, the two examples below reveal a type of ambiguity which may appear in the case of a verbal expression, which has the same meaning, but is found in different contexts.

(a) ⊘ *Este greu* pentru tine.
[It] is difficult for you.

(b) *Este greu* să scrii versuri.
[It] is difficult to write lyrics.

Example (a) contains an impersonal verbal expression which has a zero subject. In contrast, example (b) shows the same expression having a nominal clause as its subject. Therefore, a zero pronoun identification and resolution system may encounter classifying problems because of the ambiguity.

Although these constraints cover the majority of false-positive cases, there are still various infrequent constructions, which need be filtered from the candidate list. Furthermore, it becomes difficult to distinguish between the reflexive and impersonal use of verbs when they are preceded by "se" and do not have an explicit subject.

Moreover, problems may be caused because of the lack of semantic information. For example, number disagreement between the antecedents of zero pronouns and the dependent verbs is a frequently occurring pattern. Whether the subject is a singular collective noun (example (a)) or a sequence of coordinated singular nouns (example (b)), the verb will have the plural number due to the semantics of the sentence. Thus, an important marker for a zero pronoun will produce false negatives.

(a) *O sumedenie* de copii au venit şi $_{zp}$*[ea]* au cântat.
A multitude of children came and $_{zp}$*[it] sang.*

(b) *Olli Rehn şi liderii Albaniei* au condamnat crima şi $_{zp}$*[ei]* au cerut pedepsirea infractorilor.
Olli Rehn and Albania's leaders condemned the murder and $_{zp}$*[they] requested the perpetrators' punishment.*

Additionally, another error source is the parser itself. It encounters difficulties when detecting the case of nouns and the mode, time, or person for verbs, since sometimes the same form of the word is used for multiple declensions and conjugations, respectively.

7. Conclusions

This paper introduces a new linguistic resource for the Romanian language, RoZP, which contains texts from four different genres, manually annotated for zero pronouns. A quantitative analysis has been reported, whilst the qualitative description provides an insight useful for future resolution methodologies.

As future development of RoZP, an enlargement of the number of genres and annotated ZPs is planned. Furthermore, an useful extension could be the inclusion of annotated texts which are translated into Romanian. Thus, a comparison between native and translated-into-Romanian texts regarding the use of zero pronouns and, more generally, translation universals could be perfomed.

8. References

Nancy Chinchor. 1998. *Proceedings of the Seventh Message Understanding Conference*. Science Applications International Corporation (SAIC), San Francisco, CA.

Susan P. Converse. 2006. *Pronominal anaphora resolution in Chinese*. Ph.D. thesis, Philadelphia, PA, USA.

Dan Cristea, Oana Postolache, Gabriela Dima, and Cătălina Barbu. 2002. AR-Engine - a framework for unrestricted co-reference resolution. In *Proceedings of the LREC 2002 - Third International Conference on Language Resources and Evaluation*, pages 2000–2007.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in Spanish. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 166–172, Morristown, NJ, USA. Association for Computational Linguistics.

Na-Rae Han. 2006. *Korean zero pronouns: analysis and resolution*. Ph.D. thesis, Philadelphia, PA, USA.

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the sixth conference on Applied natural language processing*, pages 142–149, Morristown, NJ, USA. Association for Computational Linguistics.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 625–632, Morristown, NJ, USA. Association for Computational Linguistics.

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

Claudiu Mihăilă and Dalila Mekhaldi. 2009. Bimodal Corpora Terminology Extraction: Another Brick in the Wall. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 236–240.

Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2006. A Computer-Aided Environment for Generating Multiple-Choice Test Items. *Journal of Natural Language Engineering*, 12(2):177–194.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Constantin Ioan Mladin. 2005. Procese şi structuri sintactice "marginalizate" în sintaxa românească actuală. consideraţii terminologice din perspectivă diacronică asupra contragerii - construcţiilor - elipsei. *The Annals of Ovidius University Constanţa - Philology*, 16:219–234.

Gabriela Pavel, Oana Postolache, Ionuţ Pistol, and Dan Cristea. 2006. Rezoluţia anaforei pentru limba română. In Corina Forăscu, Dan Tufiş, and Dan Cristea, editors, *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Iaşi, 3 November.

Simone Pereira. 2009. ZAC.PB: An Annotated Corpus for Zero Anaphora Resolution in Portguese. In Irina Temnikova, Ivelina Nikolova, and Natalia Konstantinova, editors, *Proceedings of the Student Workshop at RANLP 2009*, pages 53–59.

Ştefania Popescu. 2009. *Gramatica practică a limbii române*. TEDIT FZH, Bucureşti, 15th edition.

Luz Rello and Iustina Ilisei. 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 (ISMTCL)*.

Luz Rello and Iustina Ilisei. 2009b. A Rule Based Approach to the Identification of Spanish Zero Pronouns. In Irina Temnikova, Ivelina Nikolova, and Natalia Konstantinova, editors, *Proceedings of the Student Workshop at RANLP 2009*, pages 60–65.

Shanheng Zhao and Hwee Tou 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*, pages 541–550. Association for Computational Linguistics.