Zero Pronominal Anaphora Resolution for the Romanian Language

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

[1] Faculty of Computer Science,
"Al.I. Cuza" University of Iaşi,
16 General Berthelot Street, Iaşi 700483, Romania
claudiu.mihaila@cs.man.ac.uk
[2] Research Institute in Information and Language Processing,
University of Wolverhampton,
Wulfruna Street, Wolverhampton WV1 1LY, UK
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. This paper presents a new study on the distribution, identification, and resolution of zero pronouns in Romanian. A Romanian corpus, including 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 were performed on the created corpus for the identification and resolution of zero pronouns.

Keywords: zero pronoun, ellipsis, anaphora resolution, 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 [1]. 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 extraction, question answering, automatic summarisation, machine translation, or generation of multiple-choice test items [2] depend on 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.

⋆ The author is now with the National Centre for Text Mining, School of Computer Science, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK.

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 resolution for several languages, including Romanian. In section 4, the corpora on which this work was performed are described, and in section 5 the method is presented and the results of the evaluation are analysed.

2 Zero subjects and 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 in [3]. Despite the existing controversy, in this work we adopt the theory that follows.

Two types of elliptic subjects are found in Romanian: zero subjects and implicit subjects. Although both these two types are missing from the text, the difference between them is that whilst the implicit subject can be lexically retrieved, such as in example 1, the zero subject cannot, as shown in example 2.

1. $_{zp}$*[Noi]*[1] mergem la şcoală.
   [We] are going to school.
2. ∅ Ninge.
   [It] is snowing.

In Romanian, clauses with zero subject are considered syntactically impersonal, whereas implicit or omitted subjects, which are not phonetically realised, can be retrieved lexically [4].

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 occurs when an anaphoric pronoun is omitted but nevertheless understood [1]. An anaphoric zero pronoun (AZP) results when the zero pronoun corefers to one or more overt nouns or noun phrases in the text.

The difficulty that arises in the task of identifying zero pronouns is to distinguish between personal and impersonal use of verbs. Whilst impersonally used verbs take zero subjects (and thus have no associated ZP), personally used verbs need a subject, which in turn can be explicit or implicit. The main classes of impersonal verbs are exemplified in what follows. The examples' translation into English may sound stilted, but this is in order to provide a better understanding of the phenomenon for non-Romanian speakers.

1. Meteorological phenomena:
   ∅ *S-a înnorat* azi. ∅ *Este* iarnă.
   [It] clouded over today. [It] is winter.

[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. Changes in the moments of the day:
   ⊘ *Se luminează* de ziuă la ora opt.
   [It] is dawning at eight o'clock.
3. Impersonal expressions with dative:
   ⊘ Îmi pare rău pentru tine. Azi ⊘ nu-mi arde de glumă.
   [It] feels sorry to me for you. Today [it] does not feel like joking to me.
4. Impersonal constructions with verbs *dicendi*:
   ⊘ Se vorbeşte despre ea.
   [People] are talking about her.
5. Romanian impersonal constructions with personal verbs when preceded by the reflexive pronoun "se":
   ⊘ *Se cântă* aici.
   [People] are singing here.

For the resolution step, a similar challenge exists. In this case, the main issue is to define a list of antecedent candidates, and to choose the correct one.

3 Related Research

Most of the studies developed for the task of coreference resolution were performed for the English language. Consequently, publicly available corpora that were created for this task are also available mostly for English, e.g., at the Message Understanding Conferences (MUC6 and MUC7[2]) [5].

In the context of machine translation, a hand-engineered rule-based approach to identify and resolve Spanish zero pronouns that are in the subject grammatical position is proposed by [6]. In their study, a slot unification grammar parser is used to produce either full parses or partial parses according to a runtime parameter. The parser produced "slot structures" that had empty slots for unfilled arguments. These were used to detect zero pronouns for verbs that were not imperatives or impersonal (e.g., "Llueve."/"[It] rains."). For testing the zero pronouns, the Lexesp corpus was used, which contains Spanish texts from different genres. It has 99 sentences, containing 2213 words, with an average of 21 words per sentence. The employed heuristics detected 181 verbs, of which 75% had a missing subject, and the system resolved 97% of those subjects correctly.

Furthermore, another Spanish corpus annotated with more than 1200 ZPs was created to complement the previous study by considering the detection of impersonal clauses using hand-built rules; the reported F-measure is 57% [7, 8].

Ching-Long Yeh tried detecting and resolving zero pronouns in Chinese [9] by POS-tagging followed by phrase-level chunking. Data structures called *triples* were created from the chunked sentence, to be used both in detecting zero pronouns and in resolving them. Yeh tested on a corpus of 150 news articles containing 4631 utterances and 41000 words. A precision of 80.5% in detecting zero pronouns was reported. For the resolution stage, a recall of 70% and a precision of 60.3% of the total zero anaphors were achieved.

[2] http://www-nlpir.nist.gov/related_ projects/muc/

Converse [10] developed a rule-based approach on data from the Penn Chinese Treebank. The heuristic used is based upon Hobbs' algorithm, which traverses the surface parse tree in a particular order looking for a noun phrase of the correct gender and number [11]. Converse defined rules to substitute the lack of gender and number verb markers in the corpus, and imposed selectional/semantic restrictions, both in order to reduce the number of candidates and obtain a better accuracy. The computed recency baseline is 35%, and the top score is 43%.

Another machine learning approach which identifies and resolves zero pronouns for Chinese is described in [12], and the results are comparable to the ones obtained in [10]. Making use of parse trees and simple rules to determine the ZP and NP candidates, two classifiers are built for the identification of actual ZP from the candidate list and for resolving each of the previously identified ZPs to one of the candidate NPs. The feature vectors were then computed for the ZP candidates and for their antecedents, and a value of 28.6% was obtained for the task of identifying zero pronouns, and 26% for resolution. However, the training data is highly disproportionate, with only one positive example for 29 negative examples. The best results were reported at a ratio of 1:8 positive:negative.

Other languages that have been more intensively and recently studied are Portuguese [13], Japanese [14] and Korean [15, 16].

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 in [17]. The authors create a bilingual collection by having the MUC-6 and MUC-7 coreference training texts translated into Romanian by native speakers, and using, wherever possible, the same coreference identifiers as the English data and incorporating additional tags as needed. By using an aligned English-Romanian corpus, they exploit natural language differences to reduce uncertainty regarding the antecedents and manage to correctly resolve coreferences. Furthermore, bilingual lexical resources are used, such as an English-Romanian dictionary and WordNets, to find translations of the antecedents for each of the language.

Another study on a rule-based Romanian anaphora resolution system relying on RARE [18] was reported in [19]. First, the input is analysed using a morphological parser and a nominal group identifier. Afterwards, by employing hand-written weighted rules, such as regarding agreement in person, gender, or number, the system manages to identify coreferential chains with a success rate of 70% and an MUC precision and recall of 25% and 60%, respectively.

However, it should be noted that none of these studies consider zero pronominal anaphora in their development.

4 Corpora

This section describes the corpora on which this study is based. 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 documents included in the corpus are classified in four genres, i.e., law (LT), newswire (NT), encyclopaedia (ET), and literature (ST). The newswire texts contain international news published in the beginning of 2009, while the law part of the corpus represents the Romanian constitution. The literary part is composed of children's short stories by Emil Gârleanu and Ion Creangă, whilst the encyclopaedic corpus comprises articles from the Romanian Wikipedia on various topics.

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

In what follows, the annotation setup is provided, 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 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.

The texts were afterwards manually annotated for zero pronouns by two authors, in order to create a golden standard. The inter-annotator agreement regarding the existence of zero pronouns is 90%.

A zero pronoun was 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" agreement="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 coresponding 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 unique identifier 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 pronoun appears, one more attribute (sentence type) provides information about the kind of sentence (main, coordinated, subordinated, etc.).

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

4.2 Statistics

The currently gathered corpus comprises over 55000 tokens and almost 1000 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 ZP per sentence.

Table 1. Description of the corpora.

Overview	ET	LT	ST	NT	Overall
No. of tokens	17191	13739	5141	19374	55445
No. of sentences	728	790	371	852	2741
No. of ZP	235	113	391	258	997
Avg. tokens/sentence	23.61	17.39	13.85	22.73	20.22
Avg. ZP/sentence	0.32	0.14	1.05	0.30	0.36

The distribution of the zero pronouns in the studied corpora is provided in Table 2. The distances from zero pronouns to their antecedents in the case of newswire and literature texts reveal unique values. This is due to the different writing styles, 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, which is on average 1.68 tokens away.

Table 2. Distances between the ZP and its antecedent and dependent verb.

Corpus	Antecedent (sentences)	Antecedent (tokens)	Dependent verb (tokens)
ET	0.68	23.87	1.77
LT	1.07	38.55	1.56
ST	2.92	46.63	1.62
NT	0.02	7.44	1.74
Overall	1.43	30.21	1.68

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 [7].

5 Evaluation

5.1 Identification of Zero Pronouns

The first goal is to classify the verbs into two distinct classes, either with or without a zero pronoun. The chosen method in this study is supervised machine

learning, using Weka[4] [20, 21]. Therefore, a feature vector was constructed for the verbs. The vector is composed of the following eleven elements:

- type – the type of the verb (i.e. main, auxiliary, copulative, or modal);
- mood – the mood of the verb (indicative, subjunctive, etc.);
- tense – the tense of the verb (present, imperfect, past, pluperfect);
- person – the person of the conjugation (first, second, or third);
- number – the number of the conjugation (singular or plural);
- gender – the gender of the conjugation (masculine, feminine, or neuter);
- clitic – whether the verb appears in a clitic form or not;
- impersonality – whether the verb is strictly impersonal or not (such as meteorological verbs);
- 'se' – whether the verb is preceded by the reflexive pronoun "se" or not;
- number_of_verbs_in_sentence – the number of verbs in the sentence where the candidate verb is located;
- 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, whilst the next three elements are computed automatically based on the annotated texts. The last item is the class whose values are true if the verb allows zero pronouns 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 1994 instances of the feature vector. Half of these instances correspond to the 997 verbs which have an associated ZP, whilst the other half contains randomly selected verbs without a ZP. As the baseline classifier employed, ZeroR, takes the majority class, the baseline to which we need to compare our accuracy is 50%.

Multiple classifiers pertaining to different categories were 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 3. SMO is the implementation of SVM, J48 is an implementation of decision trees, and Jrip is an implementation of decision rules. The Vote meta-classifier is configured to consider the three previous classifiers using a Majority Voting combination rule.

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.

In order to observe the rules according to which the decisions are made, the Jrip classifier was employed. The obtained output is included in Figure 1. The most used attribute is clearly the mode of the verb, whilst the gender and the clitic form do not appear at all.

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

Table 3. Scores from four classifiers for the classes of verbs.

Classifier	Accuracy	has zp			not zp		
		P	R	F_1	P	R	F_1
SMO	0.739	0.814	0.620	0.704	0.693	0.859	0.767
Jrip	0.734	0.722	0.764	0.742	0.749	0.705	0.727
J48	0.746	0.783	0.683	0.730	0.719	0.810	0.762
Vote	0.745	0.785	0.675	0.726	0.715	0.815	0.762

```
(MOOD = 1) => HASZP = true (308.0/33.0)
(MOOD = 0) and (TENSE = 2) => HASZP = true (200.0/58.0)
(MOOD = 0) and (VERBNUMBERINSENTENCE >= 6) => HASZP = true (140.0/44.0)
(MOOD = 0) and (TENSE = 3) => HASZP = true (28.0/2.0)
(MOOD = 0) and (PERSON = 0) => HASZP = true (36.0/6.0)
(PERSON = 2) and (VERBNUMBERINSENTENCE >= 5) and
                 (VERBNUMBERINSENTENCE >= 6) => HASZP = true (139.0/58.0)
(MOOD = 0) and (NUMBER = 0) and (TENSE = 1) => HASZP = true (52.0/14.0)
(MOOD = 0) and (NUMBER = 0) and (PERSON = 1) => HASZP = true (22.0/3.0)
 => HASZP = false (1069.0/290.0)
```

Figure 1: Rules output from the Jrip classifier.

Aiming at determining which features most influence classification, regardless of the classifying algorithm, two attribute evaluators have provided the results shown in Table 4.

Table 4. Attribute selection output from two attribute evaluators.

Attribute	ChiSquare	InfoGain
Mood	437.09	0.1728
Person	29.09	0.0108
Verb number in sentence	15.97	0.0058
Tense	15.42	0.0057
Type	14.65	0.0053
Impersonality	10.35	0.0044
Number	7.44	0.0026
'Se'	5.28	0.0019
Gender	1.95	7E-4
Clitic	0	0

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 that "se" is part of impersonal constructions which may or may not have zero pronouns. As a result, the system classifies the verbs incorrectly.

5.2 Resolution of Zero Pronouns

The second goal of our research is to find the correct antecedent to resolve the anaphor. The methodology employed in resolving zero pronouns is supervised machine learning, using the aforementioned gold corpus as training and test data.

The feature vector that was constructed for the verbs and antecedent candidates is composed of 21 elements, the first nine of which are the same as in the identification stage. The other are briefly described in what follows:

- number_of_verbs_in_sentence – the number of verbs in the sentence where the zero pronoun is located;
- candidate_pos – the part of speech of the candidate (i.e. noun or pronoun);
- candidate_type – the type of the candidate (i.e. main, auxiliary, copulative, or modal);
- candidate_case – the case of the candidate (direct, oblique, or vocative);
- candidate_person – the person of the candidate (first, second, or third);
- candidate_number – the number of the candidate (singular or plural);
- candidate_gender – the gender of the candidate (masculine, feminine, or neuter);
- candidate_definite – whether the candidate appears in a definite form or not;
- candidate_clitic – whether the candidate appears in a clitic form or not;
- distance_sentences – the distance in sentences between the verb and the candidate;
- distance_tokens – the distance in tokens between the verb and the candidate;
- isAnt – whether the candidate is the ZP's antecedent (verb's subject) or not.

Two baselines have been taken into account for this stage. Firstly, the ZeroR classifier takes the majority class as the class for the entire population. Due to the selection of the data, its accuracy is 50%.

The second baseline employed considers as antecedent the first previous noun, pronoun, or numeral which is in gender and number agreement with the verb. Its accuracy is really low, only 12.52%. Most of the cases that are correctly identified by this baseline are those in which the verb is in the subjunctive mood, and the antecedent precedes it and is declined in the oblique case. Such an example is included in the sentence below, where *it* refers to *Macedonia*.

- *[...]* a cerut Macedoniei $_{zp}$*[ea]* să stabilească relaţii diplomatice la Kosovo.
  [...] asked Macedonia $_{zp}$*[it] to establish diplomatic relations in Kosovo.*

The classifiers that were experimented with are the same as those in the prior identification stage. The SMO, JRip, and J48 classifiers and Vote meta-classifier were run with a 10-fold cross validation, and the results that were obtained are included in Table 5.

Due to the fact that only a subset of false candidates was considered in the training and test data, the results vary between various re-runs of the experiment. However, repeating the experiment several times with different data

Table 5. Classifier results for the classes of candidates.

Classifier	Accuracy	is Antecedent			not Antecedent		
		P	R	F_1	P	R	F_1
SMO	0.727	0.717	0.751	0.733	0.738	0.703	0.720
JRip	0.839	0.882	0.783	0.829	0.805	0.895	0.848
J48	0.864	0.852	0.882	0.867	0.877	0.847	0.862
Vote	0.865	0.867	0.862	0.865	0.863	0.868	0.865

proved that the variations are small and are not statistically significant. The SVM classifier is outperformed by the other two, decision trees and decision rules, and also by the Vote meta-classifier.

Figure 2 shows decision rules for the JRip classifier. The features that occur on higher levels, such as the distances, candidate case, POS, or definiteness, appear to help classify most of the given antecedents.

```
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 5) and
        (CANDIDATEDEFINITE = 1) => ISANTECEDENT = false (372.0/28.0)
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 5) and
        (VERBNUMBERINSENTENCE >= 4) => ISANTECEDENT = false (202.0/28.0)
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 5) and
        (CANDIDATECASE = 1) => ISANTECEDENT = false (68.0/8.0)
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 5) and
        (CANDIDATENUMBER = 1) => ISANTECEDENT = false (45.0/7.0)
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 5) and
        (DISTANCESENTENCES >= 2) and (CANDIDATEPOS = 0)
               => ISANTECEDENT = false (166.0/57.0)
(CANDIDATEPOS = 1) and (CANDIDATEDEFINITE = 1)
               => ISANTECEDENT = false (37.0/1.0)
(CANDIDATETYPE = 5) => ISANTECEDENT = false (13.0/3.0)
(DISTANCESENTENCES >= 1) and (DISTANCESENTENCES <= 3) and
        (CANDIDATETYPE = 0) => ISANTECEDENT = false (9.0/1.0)
(CANDIDATECASE = 1) and (DISTANCETOKENS >= 16)
               => ISANTECEDENT = false (4.0/0.0)
 => ISANTECEDENT = true (824.0/87.0)
```

Figure 2: Rules output from the Jrip classifier.

The attributes that are the most salient in this classification, according to Table 6, are the distances between the candidate and the verb, measured in both sentences and tokens. Other very important attributes are the definiteness, case, and type of the candidate, as can also be observed from the aforementioned decision rules.

It is important to note that the learning model relies more on candidate features than on verb features. While some of the candidate features have very high values, most of the verb features are given a null value by the two attribute

evaluators, ChiSquare and InfoGain. The features with null values have been omitted from the table.

Table 6. Resolution attribute selection output from two attribute evaluators.

Attribute	ChiSquare	InfoGain
Distance in sentences	770.282	0.3608
Distance in tokens	491.870	0.2212
Candidate definite	168.496	0.0714
Candidate case	154.011	0.0688
Candidate type	93.328	0.0480
Verb number in sentence	20.063	0.0083
Candidate person	7.825	0.0041
Candidate gender	5.542	0.0022
Verb mood	5.062	0.0021
Verb type	1.447	0.0006
Candidate PoS	1.100	0.0004
Verb tense	0.819	0.0003
Candidate number	0.224	0.0001

6 Conclusions and future work

This paper presents a study on the distribution, identification, and resolution of zero pronouns in Romanian. By creating and manually annotating a multiple-genre corpus, zero pronouns are identified and resolved using supervised machine learning algorithms. The accuracies of 74% for identification and 86% for resolution are comparable to those obtained for other languages for which such studies have been performed.

Concerning the usability of this study, applications include question answering and automatic summarisation. As a large number of ZPs are present in text, extracting the correct subject of important actions is vital. Furthermore, machine translation might benefit for pairs of languages with different rules regarding zero pronouns. Moreover, since the distribution depends largely on the genre, it might depend on the author as well, and thus automatic zero pronoun identification might be used in plagiarism and authorship detection.

References

1. Mitkov, R.: Anaphora Resolution. Longman, London (2002)
2. Mitkov, R., Ha, L.A., Karamanis, N.: A computer-aided environment for generating multiple-choice test items. Journal of Natural Language Engineering **12** (2006) 177–194

3. 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 Constanţa - Philology **16** (2005) 219–234
4. Institutul de Lingvistică "Iorgu Iordan - Al. Rosetti" Bucureşti: Gramatica limbii române. Editura Academiei Române, Bucureşti (2005)
5. Proceedings of the seventh Message Understanding Conference (MUC 7). (1998)
6. Ferrández, A., Peral, J.: A computational approach to zero-pronouns in Spanish. In: ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2000) 166–172
7. Rello, L., 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)
8. Rello, L., Ilisei, I.: A rule based approach to the identification of Spanish zero pronouns. In Temnikova, I., Nikolova, I., Konstantinova, N., eds.: Proceedings of the Student Workshop at RANLP 2009. (2009) 60–65
9. Yeh, C.L., Chen, Y.C.: Zero anaphora resolution in Chinese with shallow parsing. Journal of Chinese Language and Computing **17** (2007) 41–56
10. Converse, S.P.: Pronominal anaphora resolution in Chinese. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA (2006)
11. Hobbs, J.R.: Resolving pronoun references. Lingua **44** (1978) 311–338
12. Zhao, S., 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, Association for Computational Linguistics (2007) 541–550
13. Pereira, S.: ZAC.PB: An annotated corpus for zero anaphora resolution in Portuguese. In Temnikova, I., Nikolova, I., Konstantinova, N., eds.: Proceedings of the Student Workshop at RANLP 2009. (2009) 53–59
14. Iida, R., Inui, K., Matsumoto, Y.: 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, Morristown, NJ, USA, Association for Computational Linguistics (2006) 625–632
15. Kim, Y.J.: Subject/object drop in the acquisition of Korean: A cross-linguistic comparison. Journal of East Asian Linguistics **9** (2000) 325–351
16. Han, N.R.: Korean zero pronouns: analysis and resolution. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA (2006)
17. Harabagiu, S.M., Maiorano, S.J.: Multilingual coreference resolution. In: Proceedings of the sixth conference on Applied natural language processing, Morristown, NJ, USA, Association for Computational Linguistics (2000) 142–149
18. Cristea, D., Postolache, O.D., Dima, G.E., Barbu, C.: AR-Engine - a framework for unrestricted co-reference resolution. In: Proceedings of the LREC 2002 - Third International Conference on Language Resources and Evaluation. (2002) 2000–2007
19. Pavel, G., Postolache, O., Pistol, I., Cristea, D.: Rezoluţia anaforei pentru limba română. In Forăscu, C., Tufiş, D., Cristea, D., eds.: Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române, Iaşi (2006)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11** (2009)
21. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann (2005)