KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2011
Cluj-Napoca (Romania), July 4–6, 2011, pp. 25–35

THE IMPACT OF ZERO PRONOMINAL ANAPHORA ON TRANSLATIONAL LANGUAGE: A STUDY ON ROMANIAN NEWSPAPERS

IUSTINA ILISEI[(1)], CLAUDIU MIHĂILĂ[(2)], DIANA INKPEN[(3)], AND RUSLAN MITKOV[(4)]

Abstract. This study investigates the impact of zero pronominal anaphora for Romanian on a learning model able to distinguish between translated and non-translated texts. Even though the correct understanding of ellipsis from the source language and its mapping into the target language is essential in the translation process, zero pronominal anaphora has been scarcely investigated in the context of translation studies domain. This paper reports the results of a supervised learning system which exploits the anaphoric zero pronoun feature and its informativeness in the learning process. Moreover, ellipsis is one of the attributes proposed for the investigation of explicitation universal, and hence this study also brings an argument towards the existence of this hypothesis.

1. Introduction

Interest in studying translational language started a long time ago and certain theories and hypotheses have been proposed. It has been claimed that translated texts will always have certain particular features compared to non-translated ones, leaving them specific unnatural 'fingerprints'. This effect was named 'translationese' [1]. Furthermore, a set of various hypotheses were brought forward [2, 3], and some of them claimed to be universals of translations [4, 5]. The translation universals theory continues to be a highly debated issue within the translation studies domain. Some scholars disagree with these hypotheses or even argue the universality aspect of this theory

Received by the editors: 15th April 2011.

2010 *Mathematics Subject Classification.* Natural language processing, 68T50.

1998 *CR Categories and Descriptors.* I.2 [**Artificial Intelligence**]: Natural Language Processing – *Text Analysis*.

Key words and phrases. Anaphora, Zero Pronominal Anaphora, Machine Learning, Translationese, Translation Theory, Explicitation Universal.

[6, 7], while others emphasise the value brought by these assumptions in the practice of professional translation [8].

The reasons to investigate these hypotheses are multiple: first, to bring to light various tendencies of translational language [9], and hence, to pave the way for more accurate and natural translations [10]. Second, the automatic identification of these unconscious tendencies can improve automatic web-based parallel corpus extractors by enhancing their ability to correctly identify the candidate parallel text [11]. Also, according to recent studies ([12, 13]), the automatic detection of translationese can improve statistical machine translation frameworks.

The objective of the current study is to investigate to what extent zero pronominal anaphora appears in translational language. In the following paragraphs the main concepts and assumptions of this study are described.

1.1. **Explicitation.** One of these hypotheses is explicitation, first defined twenty-five years ago by Blum-Kulka [14]. She emphasised the concept that "explicitation is a universal strategy inherent in the process of language mediation "[14](p.21). In [15, 16] it is suggested that changes in function words, such as addition, deletion or replacement, can lead to a shift in the degree of explicitness through which cohesion is attained (p.81). As [17] points out that cohesion change is one of the syntactic strategies which "affects intra-textual reference, ellipsis, substitution, pronominalisation and repetition, or the use of connectors of various kinds"(p.98), then ellipsis can therefore be considered as one of the attributes through which the explicitation universal can be investigated. This universal states that professional translators prefer to "spell things out rather than leave them implicit"[5]. Also, various studies note an increased level of repetitions due to translators' tendency to be more precise and to disambiguate the message conveyed [9, 18]. Consequently, it can be concluded that ellipsis is expected to be avoided in translated language more than in non-translated language, and hence, it has the potential to be an important feature in the task of classifying between translated and non-translated texts. In this research study, the only type of ellipsis under investigation is the anaphoric zero pronoun explored in the Romanian language.

It is known that the typology of explicitation hypothesis can be divided into two categories: obligatory (ex.1) and voluntary (ex.2). There are classical examples in Portuguese used to clarify explicitation quoted from [19]. Obligatory explicitation appears when the target language forces translators to add information not present in the source text due to language restrictions, whilst voluntary manifests only if translators intentionally avoid any possible misinterpretations in their produced texts.

(1) *Source:* Frances liked her doctor.
Translation: Frances gostava dessa médica.
Back translation: Frances liked this *[female]* doctor.

(2) *Source:* Você também gosta dela?
Translation: So you like her too?
Back translation: You like her too?

Just like in almost all Romance languages, the anaphoric zero pronoun is entirely optional in Romanian (with the exception, however, of cases of emphasis, contrast and the like). Therefore, their presence in translated text is entirely dependent on the translators' decision. These experiments aim to analyse one potential characteristic of voluntary explicitation in Romanian. In the following subsection, an overview of the anaphoric zero pronoun for Romanian is presented.

1.2. **Zero Pronominal Anaphora.** Defining anaphora in the case of the Romanian language is a controversial topic, and complete agreement between the scholars has not yet emerged. As a consequence, there are different classifications of ellipsis [20]. This study exploits zero pronominal anaphora, and the definition adopted is as follows: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood [21], in which case the zero pronoun corefers to one or more overt nouns or noun phrases in the text (entities which provide the information for the correct understanding of the ellipsis). In this study we focus on the ellipsis of subjects, as it is the most frequent case.

Note that in the Romanian language there are two types of elliptic subjects: zero subjects and implicit subjects. The difference between them consists in the fact that implicit subjects can be lexically retrieved (ex. 3, example quoted from [22]), while zero subjects cannot[1] (ex. 4, example quoted from [22]).

(3) $_{zp}$*[Noi]* mergem la şcoală.
[We] are going to school.
(4) ⊘ Ninge.
[It] is snowing.

2. Research Methodology

2.1. **RoTC Corpus.** The corpus used for these experiments is a monolingual comparable corpus specifically designed for the investigation of translationese

[1]In the following examples, a zero pronoun is marked with $_{zp}$*[]*, while a zero subject is marked with the ⊘ sign.

and other translation hypotheses. The resource used is the Romanian Translational Comparable Corpus (RoTC corpus) that comprises several newspapers articles, translated and non-translated, written between 2005-2009. It has a subcorpus of 223 translated articles collected from the Southeast European Times website[2], and the comparable non-translated corpus which has 416 articles from the same time-span and in the same domain, documents collected from a well-known Romanian newspapers website, called 'Ziua'[3]. The RoTC corpus has a total of 341320 tokens, with 200211 for the translated subcorpus and 141109 tokens for the non-translated one. To avoid any type of source language interference or specific authorship style, the translated subcorpus comprises texts written by various authors and translated from various source languages.

This comparable corpus has been previously exploited in a similar experiment for the identification of translationese, except the ellipsis feature was not part of data representation and neither the scope of the study [23]. To the best of our knowledge, this is the first study which investigates the presence and impact of zero pronominal anaphora in translated texts compared to non-translated texts.

2.2. **Data Representation.** The approach undertaken is a supervised learning model which aims at learning to differentiate between translated and non-translated texts. Data representation considers the following language-independent features (suggested by various scholars in the field to stand in favour of simplification universal [5, 9, 24]): information load, lexical richness, sentence length, word length, and simple sentences.

In addition to this data representation, the learning model is enhanced with one more feature: the average number of anaphoric zero pronouns in the document. This attribute is automatically retrieved using the machine learning approach proposed by [25, 22], and it is computed as the number of verbs which have a zero pronoun in the subject position divided by the total number of verbs in the document. The learning model proposed achieves an accuracy of 74% using training and testing datasets from four domains: legal, encyclopaedic, literary, and news articles [22]. As the domain in the current experiments is also news texts, the learning model was used to identify the verbs which have a zero pronoun in the subject position. The assumption of this study is the following: if the addition of the anaphoric zero pronoun attribute improves the accuracy of the learning model, then this consequence may be considered as an argument in favour of the explicitation hypothesis.

[2]http://www.setimes.com
[3]http://www.ziuaveche.ro

The collected dataset was randomly divided into a training set of 639 texts and a test set of 148 texts. The same ratio of translated and non-translated class instances in the training and test set was maintained (49 translated class, 99 non-translated class). All attributes needed in the learning process were extracted using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[4], the Romanian Academy [26, 27]. The learning classifiers used for the experiments are: SVM, Naïve Bayes, JRip, and Decision Trees. These algorithms proved to be accurate in similar experiments on the identification task of translationese [28, 23].

An additional experiment constitutes the training of the learning model using only the anaphoric zero pronoun feature. The objective is to investigate to what extent the model is able to perform the same task relying only on this attribute. Because this study focuses only on anaphoric zero pronouns, the current data representation is not exploiting any other explicitation features, such as conjunctions, adverbs or sentence length [29, 24].

3. Evaluation

The baseline used is the ZeroR algorithm, which considers the majority class of the learning model. In our case, the baseline is 65.10% for the cross-validation and 66.89% for the randomly generated test dataset. By using the Weka tool[5] [30, 31], classifiers are trained by including and excluding the zero pronoun attribute from the learning model. In table 1, the results show that Naïve Bayes and SVM classifiers performed best: the addition of the AZP feature to the learning model improves the accuracy of Naïve Bayes algorithm from 88.58% to 89.67% for the 10-fold cross-validation evaluation, and from 85.81% to 89.19% for the test dataset. The SVM classifier is improved from 87.64% to 88.11% for the 10-fold cross-validation, and from 87.84% to 89.19% for the test dataset.

Even though the Naïve Bayes and SVM classifiers are improved by the addition of the AZP feature, the other two classifiers achieve interesting results. The decision trees classifier obtains an outstanding accuracy of 95.27% on the test dataset, and surprisingly, the addition of the AZP attribute decreases the accuracy of the classifier by 1.35% from 96.62% to 95.27%. The slight decreased accuracy is maintained also for the 10-fold cross-validation evaluation being hardly noticeable with a difference of only 0.63%. JRip is another classifier that achieves a slightly lower success rate when the learning model considers the AZP attribute. It presents similar behaviour to the decision trees classifier, but only for the 10-fold cross-validation decreasing

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

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

	Including AZP		**Excluding AZP**	
Classifier	*10-fold cross-validation*	*Test set*	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%	65.10%	66.89%
Naïve Bayes	89.67%	89.19%	88.58%	85.81%
SVM	88.11%	89.19%	87.64%	87.84%
JRip	*86.85%*	94.59%	88.42%	93.24%
J48	*88.26%*	*95.27%*	88.89%	96.62%

TABLE 1. Classification Results

from 88.42% to 86.85% accuracy. On the test dataset the addition of the AZP feature improves the learning model, from 93.24% to 94.59% success rate. A possible justification for such results on some of the classifiers is the fact that the other five features from the data representation are known to be among the most indicative attributes for the categorisation task between translated and non-translated texts [23]. Hence, the AZP feature appears to improve the learning model only for specific algorithms, such as SVM and Naïve Bayes.

Furthermore, in order to present the rules considered by the classifiers, the pruned tree output from the JRip algorithm is outlined in figure 1. This classifier is one of the algorithms which provide an intuitive output for a more detailed data analysis. To identify the translated text class, the algorithm uses the first four rules, and it frequently uses the first one considering the lexical richness and simple sentences features. The second most used rule relies on the AZP attribute among other three features: information load, lexical richness, and simple sentences. To note that neither sentence length nor word length appear to influence the JRip classifier.

To analyse the impact that the AZP feature has among the other features of the learning model, the chi-squared filter has been employed. In the table 2, all the attributes are ranked and the third most influential one is the AZP feature, having a score of 132.706. To obtain a deeper analysis of this attribute and to realise to what extent the AZP feature is able to distinguish the classes by itself, an additional experiment was performed and the results are outlined in the following subsection.

3.1. **Anaphoric Zero Pronoun Feature.** The experiment employs a single feature in the data representation of the system, the AZP feature, to verify if this feature is in fact relevant for the classification. The results of this experiment are presented in table 3. Among all the learning algorithms, the JRip classifier is the one which performs best: it achieves an accuracy of 72.46% on cross-validation, and 77.03% on the test dataset. The results hardly

```
Rule 1: (LexicalRichness <= 0.50) and (SimpleSentences >= 0.80)
=> class=translated (128.0/9.0)

Rule 2: (InformationLoad <= 0.001) and (VbHasZPavg <= 0.38) and
(InformationLoad >= 0.0007) and (LexicalRichness <= 0.49) and
(SimpleSentences >= 0.65) => class=translated (42.0/2.0)

Rule 3: (LexicalRichness <= 0.51) and (SimpleSentences >= 0.79) and
(LexicalRichness <= 0.50) => class=translated (15.0/0.0)

Rule 4: (InformationLoad <= 0.001) and (LexicalRichness <= 0.46) and
(InformationLoad >= 0.0006) and (SimpleSentences >= 0.53) and
(LexicalRichness <= 0.45) => class=translated (18.0/2.0)

Rule 5: => class=non-translated (436.0/33.0)
```

FIGURE 1. JRip classifier rules output.

Chi squared Ranking Filter	
321.455	InformationLoad
287.431	LexicalRichness
132.706	**VbHasZPavg**
130.871	SimpleSentences
34.758	SentenceLength
28.049	WordLength

TABLE 2. Attributes Filter Ranking.

vary between 71.05% (SVM) and 72.46% (JRip) for the 10-fold cross-validation evaluation. On the test dataset, the accuracy can reach up to 75% value for the decision tree classifier. The accuracies obtained for the learning system are outstanding, the model being able to effectively perform the same task relying only on this attribute, the anaphoric zero pronoun.

In order to compare the previous output provided by the JRip with the current learning model, the rules obtained for this experiment are presented in figure 2. It appears that the algorithm assigns the class as translated if the value of the AZP feature is lower than 0.27 or between 0.36 and 0.37. In the previous experiment, the value considered was 0.38, clearly close to the one achieved in the current experiment.

	Only AZP Feature	
Classifier	*10-fold cross-validation*	*Test set*
Baseline	65.10%	66.89%
Naïve Bayes	71.99%	72.30%
SVM	71.05%	70.27%
JRip	72.46%	77.03%
J48	72.14%	75.00%

TABLE 3. Classification accuracy results using only AZP Feature.

```
Rule 1: (VbHasZPavg <= 0.278351) => class=translated
Rule 2: (VbHasZPavg <= 0.372881) and (VbHasZPavg >= 0.366337)
=> class=translated
Rule 3: => class=non-translated
```

FIGURE 2. JRip classifier rules output.

4. Conclusions and Further Research

This study reports a learning model which aims at identifying to what extent anaphoric zero pronouns occur in translational language. The resource used is a Romanian comparable corpus of translated and non-translated newspaper articles. By studying zero pronominal anaphora, a type of ellipsis, the current experiments may shed light on the validation of explicitation hypothesis.

A learning model is employed for Romanian language to distinguish between translated and non-translated texts. The data representation used is enhanced with the anaphoric zero pronoun feature. The results show that the addition of this attribute can increase the success rate of the learning model by up to 3.38% for various classifiers such as Naïve Bayes and SVM. Moreover, an additional study exploiting only the anaphoric zero pronoun feature has been further performed and the results show that the learning system is in fact able to accomplish the same task on its own. The accuracy achieved in this case varies between 71% to 75% and it may be considered an argument for the existence of the explicitation universal. Further research for the analysis of the anaphoric zero pronouns in translational language can also consider various features from the resolution stage of zero pronouns.

References

[1] Gellerstam, M.: Translationese in Swedish novels translated from English. Translation Studies in Scandinavia. Lund: CWK Gleerup (1986)

[2] Toury, G.: Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins (1995)

[3] Teich, E.: Cross-linguistic Variation in System and Text. Berlin:Mouton de Gruyter (2003)

[4] Baker, M.: Corpus Linguistics and Translation Studies Implications and Applications. In: Text and Technology: In Honour of John Sinclair. Amsterdam & Philadelphia: John Benjamins (1993) 233–250

[5] Baker, M.: Corpus-based Translation Studies:The Challenges that Lie Ahead. In: Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager. Amsterdam & Philadelphia: John Benjamins (1996) 175–186

[6] Tymoczko, M.: Computerized corpora and the future of translation studies. Meta **43:4** (1998) 652–659

[7] Bernardini, S., Zanettin, F.: When is a Universal not a Universal? In: Translation Universals. Do they exist? Amsterdam: Benjamins (2004) 5162

[8] Toury, G.: Probabilistic explanations in translation studies. Welcome as they are, would they qualify as universals? In: Translation Universals: Do they exist? Amsterdam: John Benjamins (2004) 15–32

[9] Laviosa, S.: Corpus-based Translation Studies. Theory, Findings, Applications. Amsterdam & New York: Rodopi (2002)

[10] Chesterman, A.: A Causal Model for Translation Studies. In: Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects. St. Jerome (2000) 15–27

[11] Resnik, P., Smith, N.: The web as a parallel corpus. Computational Linguistics **29(3)** (2003) 349380 Motivation: web-based parallel corpus extractor by finding the candidate parallel texts.

[12] Goutte, C., Kurokawa, D., Isabelle, P.: Improving smt by learning translation direction. In: Statistical Multilingual Analysis for Retrieval and Translation, Barcelona, Spain (2009)

[13] Kurokawa, D., Goutte, C., Isabelle, P.: Automatic detection of translated text and its impact on machine translation. In: Proceedings of the MT-Summit. (2009)

[14] Blum-Kulka, S.: Shifts of cohesion and coherence in Translation. In: Interlingual and Intercultural Communication. Tubingen:Narr (1986) 17–35

[15] Leuven-Zwart, K.: Translation and original: similarities and dissimilarities i. Target **1:2** (1989) 151–181

[16] Leuven-Zwart, K.: Translation and original: similarities and dissimilarities ii. Target **2:1** (1990) 69–95

[17] Chesterman, A.: The Memes of Translation. The spread of ideas in translation theory. Amsterdam and Philadelphia: Benjamins (1997)

[18] Vanderauwera, R.: Dutch Novels Translated into English: The Transformation of a "Minority" Literature. Volume 6 of Approaches to translation studies. Amsterdam: Rodopi (1985)

[19] Pym, A.: Explaining Explicitation. In: New Trends in Translation Studies. In Honour of Kinga Klaudy. Budapest: Akadmia Kiad (2005) 29–34

[20] 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
[21] Mitkov, R.: Anaphora Resolution. Longman, London (2002)
[22] Mihăilă, C., Ilisei, I., Inkpen, D.: Zero Pronominal Anaphora Resolution for the Romanian Language. Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS" **42** (2011)
[23] Ilisei, I., Inkpen, D.: Translationese Traits in Romanian Newspapers: A Machine Learning Approach. International Journal of Computational Linguistics and Applications (2011)
[24] Corpas Pastor, G.: Investigar con corpus en traducción: los retos de un nuevo paradigma. Frankfurt am Main, Berlin & New York: Peter Lang (2008)
[25] Mihăilă, C., Ilisei, I., Inkpen, D.: To Be or Not to Be a Zero Pronoun: A Machine Learning Approach for Romanian. In: Proceedings of the Processing Romanian in Multilingual, Interoperational and Scalable Environments Workshop (PROMISE). (2010)
[26] Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: Racai's linguistic web services. In: Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco. Number ISBN 2-9517408-4-0, ELRA - European Language Ressources Association (2008)
[27] Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A.: RACAI's Question Answering System at QA@CLEF 2007. In: Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), Lecture Notes in Computer Science. Volume 5152. Springer-Verlag (2008) 3284–3291
[28] Ilisei, I., Inkpen, D., Corpas Pastor, G., Mitkov, R.: Identification of Translationese: A Machine Learning Approach. In Gelbukh, A.F., ed.: CICLing. Volume 6008 of Lecture Notes in Computer Science., Springer (2010) 503–511
[29] Becher, V.: The explicit marking of contingency relations in english and german texts: A contrastive analysis. In: Societas Linguistica Europaea - 42nd Annual Meeting, Workshop: Connectives across Languages, University of Lisbon (2009)
[30] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11** (2009) 10–18
[31] Witten, I.H., Frank, E.: Data Mining : Practical Machine Learning Tools and Techniques. Second edition edn. Morgan Kaufmann. Morgan Kaufman (2005)

(1)Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom
E-mail address: iustina.ilisei@wlv.ac.uk

(2) National Centre for Text Mining, School of Computer Science, University of Manchester, United Kingdom
E-mail address: claudiu.mihaila@cs.man.ac.uk

(3) School of Information Technology and Engineering, University of Ottawa, 800, King Edward Street, Ottawa, Canada
E-mail address: diana@site.uOttawa.ca

(4) Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom
E-mail address: r.mitkov@wlv.ac.uk