STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

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

The interest of 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' [9]. Furthermore, a set of various hypotheses were brought forward [24, 23], and some of them claimed to be universals of translations [1, 2]. The translation universals theory continues to be a highly debated issue within translation studies domain. Some scholars disagree with these hypotheses or even argue the universality aspect of this theory [28, 4],

Received by the editors: April 15, 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.

while others emphasise the value brought by these assumptions in the practice of professional translation [25].

The reasons to investigate these hypotheses are multiple: first, to bring to light various tendencies of translational language [14], and hence, to pave the way for more accurate and natural translations [7]. Second, the automatic identification of these unconscious tendencies can improve the automatic web-based parallel corpus extractors by enhancing the ability to correctly identify the candidate parallel text [22]. Also, according to [10], the automatic detection of translationese can improve statistical machine translation frameworks.

The objective of the current study is to investigate to what extent the 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 [5]. She emphasised the concept that "explicitation is a universal strategy inherent in the process of language mediation "[5](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 [6] 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 explicitation universal can be investigated. This universal states that professional translators prefer to "spell things out rather than leave them implicit"[2]. Also, various studies note an increased level of repetitions due to translators' tendency to be more precise and to disambiguate the message conveyed [14, 29]. Consequently, it can be concluded that ellipsis is expected to be avoided in translated language than in non-translated language, and hence, it has the potential to be an important feature in the classification task 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: the obligatory one (ex.1), and the voluntary one (ex.2). There are classical examples in Portuguese used to clarify explicitation quoted from [21]. The obligatory explicitation appears when the target language

forces translators to add information not present in the source text due to language restrictions, whilst the voluntary one manifests only because the 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 language. In the following subsection, an overview of the anaphoric zero pronoun for Romanian language is presented.

1.2. **Zero Pronominal Anaphora.** Defining anaphora in the case of the Romanian language is a controversial topic, and a complete agreement between the scholars has not yet emerged. As a consequence, there are different classifications of ellipsis [20]. This study exploits the zero pronominal anaphora, and the definition adopted is as follows: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood [19], 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 [18]), while zero subjects cannot[1] (ex. 4, example quoted from [18]).

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

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

(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 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 [12]. 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 [2, 14, 8]): 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 [17, 18], and it is computed as the number of

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

verbs which have zero pronouns divided by the total numbers of verbs in the document. 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.

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. 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 [27, 26]. 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 [13, 12].

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 [3, 8].

2.3. **Main Results.** 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] [11, 30], classifiers are trained by including and excluding the zero pronoun attribute from the learning model. The results show that Naïve Bayes and JRip 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 89.91% for the test dataset. To note that JRip classifier obtains an outstanding accuracy of 95.27% on the test dataset. For the additional experiment, when the learning model uses only the AZP feature, 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. Interestingly,

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

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

the results prove that the model is able to effectively perform the same task relying only on this attribute, the anaphoric zero pronoun.

3. 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 the zero pronominal anaphora, a type of ellipsis, the current experiments may shed light on the validation of explicitation hypothesis. Further research can also consider the investigation of zero pronominal resolution in translational language.

References

1. M. Baker, *Text and Technology: In Honour of John Sinclair*, ch. Corpus Linguistics and Translation Studies Implications and Applications, pp. 233–250, Amsterdam & Philadelphia: John Benjamins, 1993.
2. ———, *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, ch. Corpus-based Translation Studies:The Challenges that Lie Ahead, pp. 175–186, Amsterdam & Philadelphia: John Benjamins, 1996.
3. V. Becher, *The explicit marking of contingency relations in english and german texts: A contrastive analysis*, Societas Linguistica Europaea - 42nd Annual Meeting, Workshop: Connectives across Languages (University of Lisbon), September 9-12 2009.
4. S. Bernardini and F. Zanettin, *Translation universals. do they exist?*, ch. When is a Universal not a Universal?, p. 5162, Amsterdam: Benjamins, 2004.
5. S. Blum-Kulka, *Interlingual and Intercultural Communication*, ch. Shifts of cohesion and coherence in Translation, pp. 17–35, Tubingen:Narr, 1986.
6. A. Chesterman, *The Memes of Translation. The spread of ideas in translation theory*, Amsterdam and Philadelphia: Benjamins, 1997.
7. A. Chesterman, *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, ch. A Causal Model for Translation Studies, pp. 15–27, St. Jerome, 2000.
8. G. Corpas Pastor, *Investigar con corpus en traducción: los retos de un nuevo paradigma*, Frankfurt am Main, Berlin & New York: Peter Lang, 2008.
9. M. Gellerstam, *Translationese in Swedish novels translated from English*, Translation Studies in Scandinavia. Lund: CWK Gleerup, 1986.
10. C. Goutte, D. Kurokawa, and P. Isabelle, *Improving smt by learning translation direction*, Statistical Multilingual Analysis for Retrieval and Translation (Barcelona, Spain), May 2009.
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explor. Newsl. **11** (2009), 10–18.

12. I. Ilisei and D. Inkpen, *Translationese Traits in Romanian Newspapers: A Machine Learning Approach*, International Journal of Computational Linguistics and Applications (2011).
13. I. Ilisei, D. Inkpen, G. Corpas Pastor, and R. Mitkov, *Identification of Translationese: A Machine Learning Approach*, CICLing (Alexander F. Gelbukh, ed.), Lecture Notes in Computer Science, vol. 6008, Springer, 2010, pp. 503–511.
14. S. Laviosa, *Corpus-based translation studies. theory, findings, applications*, Amsterdam & New York: Rodopi, 2002.
15. K. Leuven-Zwart, *Translation and original: similarities and dissimilarities i*, Target **1:2** (1989), 151–181.
16. ———, *Translation and original: similarities and dissimilarities ii*, Target **2:1** (1990), 69–95.
17. C. Mihăilă, I. Ilisei, and D. Inkpen, *To Be or Not to Be a Zero Pronoun: A Machine Learning Approach for Romanian*, Proceedings of the Processing Romanian in Multilingual, Interoperational and Scalable Environments Workshop (PROMISE), 2010 (english).
18. C. Mihăilă, I. Ilisei, and D. Inkpen, *Zero Pronominal Anaphora Resolution for the Romanian Language*, Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS" **42** (2011).
19. R. Mitkov, *Anaphora Resolution*, Longman, London, 2002.
20. C. I. Mladin, *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 (Romanian).
21. A. Pym, *New Trends in Translation Studies. In Honour of Kinga Klaudy*, ch. Explaining Explicitation, pp. 29–34, Budapest: Akadmia Kiad, 2005.
22. P. Resnik and N. Smith, *The web as a parallel corpus*, Computational Linguistics **29(3)** (2003), 349380, Motivation: web-based parallel corpus extractor by finding the candidate parallel texts.
23. E Teich, *Cross-linguistic variation in system and text*, Berlin:Mouton de Gruyter, 2003.
24. G. Toury, *Descriptive translation studies and beyond*, Amsterdam: John Benjamins, 1995.
25. G. Toury, *Translation universals: Do they exist?*, ch. Probabilistic explanations in translation studies. Welcome as they are, would they qualify as universals?, pp. 15–32, Amsterdam: John Benjamins, 2004.
26. D. Tufiş, D. Ştefănescu, R. Ion, and A. Ceauşu, *Advances in multilingual and multimodal information retrieval (clef 2007), lecture notes in computer science*, vol. 5152, ch. RACAI's Question Answering System at QA@CLEF 2007, pp. 3284–3291, Springer-Verlag, September 2008.
27. D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu, *Racai's linguistic web services*, Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, no. ISBN 2-9517408-4-0, ELRA - European Language Ressources Association, May 2008.

28. M. Tymoczko, *Computerized corpora and the future of translation studies*, Meta **43:4** (1998), 652–659.
29. R. Vanderauwera, *Dutch novels translated into english: The transformation of a "minority" literature*, Approaches to translation studies, vol. 6, Amsterdam: Rodopi, 1985.
30. I. H. Witten and E. Frank, *Data mining : Practical machine learning tools and techniques*, second edition ed., Morgan Kaufmann, Morgan Kaufman, June 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