Romanian Translational Corpora: Building Comparable Corpora for Translation Studies

Iustina Ilisei*, Diana Inkpen§, Gloria Corpas‡, Ruslan Mitkov*

*Research Institute in Information and Language Processing, University of Wolverhampton,
Wulfruna Street, Wolverhampton, United Kingdom,
iustina.ilisei, r.mitkov@wlv.ac.uk

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

‡Department of Translation and Interpreting, University of Málaga, Málaga, Spain
gcorpas@uma.es

Abstract

Building comparable corpora for the investigation of translational hypotheses is an important task within the translation studies domain. This paper describes the compilation of a translational comparable corpus for the Romanian language. The resource comprises translated and non-translated news articles and it is designed to be used in the investigation of translational language and translational hypotheses.

1. Introduction

Translational hypotheses proposed in the last two decades require certain resources. Most of these hypotheses (e.g., translation universals, laws or norms) imply the comparison between translated texts produced by professional translators to non-translated texts. As a consequence, there is a need of monolingual comparable corpora specifically designed for the study of translational language. These corpora need to contain two subcorpora: a subcorpus that comprises translated texts, and a comparable one which comprises non-translated, original texts.

This paper is structured as follows: first, several reasons are given as to why it is important to compile comparable corpora for translation studies, and then the definitions required for this study are described. In section 2., some other, similar resources built for other languages are highlighted, and furthermore the paper continues with the main section of the compilation of the current corpus. This main section, 3., comprises various details regarding the data collection, data preparation, and the statistics reported for the corpus. It also provides a short example of investigations which can be undertaken relying on this linguistic resource. Finally, the paper concludes with the highlights of the corpus.

1.1. Motivation

Compiling comparable corpora for the investigation of various hypotheses proposed within the area of translation studies is currently one of the main, time-consuming tasks within the domain. These hypotheses attempt to grasp and analyse certain features of the translational language and the lack of resources proves to be a serious obstacle for further refinement of the scholars' ideas and findings, and consequently for the advancement of translation theory.

The translationese effect, one of the assumptions of the discipline which considers translated language have certain specific, peculiar traits at various linguistic levels (Borin and Prütz, 2001; Hansen, 2003; Baroni and Bernardini, 2006; Puurtinen, 2003), has been a subject of debate for the last fifteen years, bringing together different perspectives on translational language. Translation universals are hypotheses that have also raised various questions among scholars; their validity is a continuous subject of debate (Corpas et al., 2008; Becher, 2011). More rigorous evidence of these claims would lead to a refinement of the theory, would raise awareness among translators about possible effects over translated texts (Laviosa, 2002, p. 77) and would facilitate further methodologies to more accurate translations with more "desired effects and fewer unwanted ones" (Chesterman, 2000). However, the lack of appropriate resources is a significant impediment to this end.

The exploitation of monolingual comparable corpora has been widely sustained among scholars, and the call for more developments of specific tools and resources for professional translators has had an impact on the domain. Even though a few translational corpora have been built (one well-known example is the English Translational Corpus), most languages still lack a proper resource for the investigation of the translational hypotheses. To the best of our knowledge, the Romanian language would be one of these languages. This work bridges this gap and reports on the compilation of the RoTC corpus, a monolingual comparable corpus that comprises newspaper articles.

Nevertheless, the exploitation of this type of resource is not restricted to translation researchers. It can also be used in other fields: for instance, for the improvement of statistical machine translation (SMT) systems. Scholars, such as (Kurokawa et al., 2009; Lembersky et al., 2011), found that making use of translation studies' main hypotheses and findings and training their SMT framework on translational corpora can result in an overall improvement of their system.

1.2. Translational Comparable Corpora

First, an attempt at defining comparable corpora is required. The key attributes of what constitutes comparable corpora are described as follows (McEnery, 2003): two corpora, A and B, are considered to be comparable if both A and B are found to have:

- the same *sampling frame* with *similar balance* and *representativeness*
- the same *proportions* of the same *genres* in the same *domains*
- the same *sampling period*

These requirements are imposed on the current resource and further details follow in section 3.

However, a definition of comparable corpora is not yet agreed by the scholars in the field. There is only a standard provided by EAGLES (1996) in which it is emphasised that a comparable corpus is *a corpus which comprises similar texts in more than one language or variety*. This standard describes the circumstances when a comparable corpus is needed: in a comparative analysis between two or more languages, or between two or more varieties of texts.

To prevent possible misinterpretations introduced by this definition (i.e., no translational corpus can be considered comparable since the resource only has texts in one language), Baker (1995) suggests that the concept of translational corpus to be seen as a new type of comparable corpus. The resource proposed includes two subcorpora in one and the same language: one subcorpus with originally produced texts in a given language, the other one with texts translated into the same language from one or more source languages. Baker (1995) proposes that both subcorpora should be similar in terms of domain, variety of language, time span, and to be of comparable length.

Considering these definitions, it seems to be a matter of how *similar* can be understood or modelled depending on the research question. The degree of comparability is "in the eye of the beholder", strictly depending on the requirements and the objectives of the research study (Maia, 2003). Although several scholars discuss this topic, the vagueness of the concept still continues, mainly because of its fuzzy notions from the definition.

Second, the concept of translational corpus is tackled. A translational corpus contains translated texts written by human translators, and it is usually exploited within the area of translation studies. Therefore, for the investigation of hypotheses which compare assumed features of translated texts to non-translated texts, a translational comparable corpus can be considered an appropriate resource for the given research question. If the translational hypothesis does not imply a comparison between translated and non-translated texts, then a translational corpus, comprising only translated texts, may suffice.

2. Related Work

As translated text is the focal point of the translation studies domain, compiling translational corpora (both comparable and parallel) is the vital resource for various investigations. As a result, several corpus-based approaches exploit monolingual comparable corpora, where comparability is between translated and non-translated texts in the same language. Despite the difficulties which arise in the compilation process, there are linguistic resources available for the following main languages: English (Baker, 1995), Portuguese (Frankenberg-Garcia, 2004), Spanish (Corpas, 2008), Dutch and German (De Sutter and Van de Velde, 2008), Chinese (Xiao et al., 2008).

The Translational English Corpus, TEC, is probably one of the first compiled corpora for translation studies in the mid-nineties (Baker, 1995). The ten-million-word corpus comprises four categories of texts: biography, fiction, newspaper texts and in-flight magazines, with translations into English from both European and non-European languages. The main experiments were employed manually and they show that corpus-based research allowed translation universals to be more clearly defined, to progress to large-scale, target-oriented research, and to consider a wider range of socio-cultural factors (Laviosa, 2002).

For Spanish, the statistical significance of various features proposed to stand for the simplification hypothesis[1] were tested using monolingual comparable corpora on medical and technical domains (Corpas, 2008; Corpas et al., 2008).

3. RoTC : Corpus Compilation

Regarding the comparability of corpora, all the definitions have in common the following parameter: *similarity between texts*. Furthermore, the definition narrows down the concept of similarity and is described in terms of genre, domain, sampling and time-frame, all of which are tackled in the compilation of process of the RoTC corpus.

Beyond the tricky notion of comparable corpora, there are also practical issues when compiling a corpus. Some of them are classical and some of them are specific to translational corpora. Fundamental aspects to consider are the *validity* and *reliability* of the research experiments based on the specific corpus, tailored to meet the intended purpose.

Representativeness is a challenging aspect for this type of linguistic resource, as it is difficult to assure that the data is representative of a particular language or genre. When considering which texts should be included in the corpus, the decision process can go beyond the text type or genre, text function or scope and how typical or influential the given text can be. Also, regional and temporal factors have to be taken into consideration, being part of the criteria of a corpus. Nationality, age, native language, ethnicity, etc. can all be decisive factors according to the research purpose, and more often than not this type of information cannot be accessed.

Sample size is another relevant consideration and may be the most important feature in achieving representativeness: how many texts should be included in the corpus and what the size of each of them should be. Representativeness depends on whether the sample includes the full range of language variability intended, so the researchers who use the corpus will be able to generalise their findings. In contrast, Kennedy (1998) argues that a bigger corpus is not necessarily more useful than a smaller one, as the data amount under

[1] Simplification hypothesis suggests that translated texts appear to be simpler than the non-translated ones (Baker, 1993).

investigation is always limited (Kennedy, 1998, p. 66-70). Nevertheless, a smaller corpus can be sufficient in some cases, for example, if the research lines have the grammar in focus (Hunston, 2002, p. 26) and, ultimately, the data availability factor of suitable texts should not be dismissed.

3.1. Corpus Design

Some scholars from the domain suggest that the best resource for the investigation of translationese is a monolingual translational comparable corpus (i.e., containing translated and non-translated texts in the same language) (Olohan, 2004), because in this manner the approach would avoid any foreign interference (Pym, 2008) and, consequently, it would fit well in the investigation of the nature of translated text.

The main objective of this resource, the Romanian Translational Corpus, is to allow the investigation of translationese and the related translational hypotheses, such as translation universals. As no study of the Romanian language has been done for translationese, to the best of our knowledge, a dedicated type of resource did not exist. For this reason a comparable corpus has been specially compiled for this task, consisting of newspaper articles published between 2005-2009.

The RoTC corpus comprises two subcorpora: a translated subcorpus and a non-translated subcorpus. The translated one is collected from the South-East European Times[2], a multilingual news portal translated into nine languages of the Balkans, one of them being Romanian. The translated subcorpus comprises 223 articles written between 2005-2009 to keep the same time frame as the non-translated subcorpus. The non-translated subcorpus comprises 416 documents in the same domain, from a well-known newspaper in Romania called 'Ziua'[3].

3.1.1. Data Preparation

The content of the South-East European website is realised as public domain, meaning it can be used and distributed without permission. The process of selecting the articles for the RoTC corpus is described in the following paragraphs.

All the articles were downloaded using various scripts which use the URL structure information. The link allows the selection of the articles to fit various needs, that in the given context are:

- to select articles after the language (i.e., the URL contains the string "www.setimes.com/ .../ro/... "for the Romanian language),
- to select articles after the date (i.e., the date can be easily extracted from the link as it appears in this format "www.setimes.com/ .../yyyy/mm/dd/... ").

The topic of the articles selected was the international news in order to be able to cover the same subjects over the same time-span, and hence obtain a comparable corpus between the texts selected from the South-East European Times website and the Ziua newspapers. Also, the number of texts between non-translated and translated texts have been balanced by randomly selecting 416 non-translations written between 2005-2007 versus the 224 translations written between 2005-2010. A ratio of 2:1 is kept.

The RoTC corpus has in total 341320 tokens (200211 for the translated subcorpus and 141109 tokens for the non-translated subcorpus). The selected articles are written by various translators, so the possibility of a specific style playing a role in the classification task is avoided. The main shortcoming of the translated subcorpus is that the portal, due to confidentiality issues, fails to provide precise information about the source language or the identity of the original author, nor the translator. Nevertheless, some of the articles do mention the source of their news information (e.g., Reuters) and it can thus be assumed the original source language of the given text. In addition, it is often stated that various information sources were used when the given article was produced.

The argument that the articles are translations and not original texts is inferred from two distinct sources: first, this portal was entirely harvested and used in a machine translation task, reporting the resource as having translations into languages of the Balkans, including the Romanian language (Tyers and Alperen, 2010). Second, it is inferred from the following rationale: one text can not be originally produced in ten languages and yet be perfectly aligned from one language to another (i.e., one Romanian article to have its source language Romanian, the corresponding, parallel Turkish article to have its source language Turkish, and at the same time, both the Romanian article and the Turkish one to be perfectly aligned to each other). The fact that all are aligned to each other leads to the assumption that, at least nine out of ten parallel articles are in fact translations. Consequently, it results in a high probability to have mostly translations, if not only translations, in the RoTC translated subcorpus. However, the attempt to clarify this issue from its source failed due to the portal's confidentiality policy.

The non-translated subcorpus does not present the same difficulty in assessing whether the texts are originally produced articles, since the newspaper is a national one having its texts written only in the Romanian language. Additionally, the articles do state their authors, and their full names indicate that they are Romanian natives. Thus, the subcorpus comprises non-translated texts, written by various authors.

3.1.2. Part of Speech Tagger

All the texts were tagged using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[4], the Romanian Academy (Tufiş et al., 2008b; Tufiş et al., 2008a), and its output transformed into XML[5] format to ease the access to the data representation of the document. A sample of the XML format is represented in figure 1. In the following section, a few statistics about the size of the RoTC corpus and its components are reported.

[2] http://www.setimes.com

[3] http://www.ziuaveche.ro

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

[5] Extensible Markup Language

```
<sentence id="w128">
<token id="w129"><text>Acestea</text>
<lemma>acesta</lemma><tags>
<morpho>Pd3fpr</morpho></tags>
</token>

<token id="w130"><text>au</text>
<lemma>avea</lemma><tags>
<morpho>Va--3p</morpho></tags>
</token>

<token id="w131"><text>fost</text>
<lemma>fi</lemma><tags>
<morpho>Vmp--sm</morpho></tags></token>

<token id="w132"><text>primele</text>
<lemma>prim</lemma><tags>
<morpho>Mofprly</morpho></tags></token>

<token id="w133"><text>alegeri</text>
<lemma>alegere</lemma><tags>
<morpho>Ncfp-n</morpho></tags></token>
... ... ...
</sentence>
```

Figure 1: Sample of the output provided from the POS tagger converted into XML format.

3.2. RoTC Corpus Statistics

Some fundamental statistics are computed for the RoTC corpus. In table 3.2. the size of the corpus is presented as the number of tokens for each subcorpus, and as a whole. It is noted that the RoTC corpus has a slight majority of non-translated texts, comprising 58.6578 % of the total number of articles. This happens as the amount of texts available for the same topic in the comparable translated corpus is slightly lower compared to the number of non-translated articles, and the intention is to obtain as many articles as possible to be able to use the resource in a machine learning framework. Obviously, the comparability aspects are considered, so it is settled to keep a ratio of 2:1 between the translated and non-translated texts to comply with the same sampling frame with similar balance factor.

RoTC Corpus			
Subcorpus	Tokens No.	Texts No.	Percentage
Non-Translated	200211	223	58.6578 %
Translated	141109	416	41.3421 %
Total	341320	639	100%

Table 1: RoTC Corpus Statistics.

To tackle the same proportions of the same genres in the same domain requirement, table 3.2. presents the average value for the number of tokens per text. The figures show that the RoTC corpus has an average number of tokens of 481 for the translated subcorpus, and 632 for the non-translated texts. These values are closely related (as expected since in this corpus there are only newspapers articles), and it remains to be investigated further whether the slight difference is due to some feature assumed to be specific to either translational language or to non-translational one (some hypotheses make references related to the size of translated texts in general). Nevertheless, the RoTC corpus also complies with the same proportion requirement for a comparable corpus.

RoTC Corpus	
Subcorpus	Average
Non-translated	632.7757848
Translated	481.2764423

Table 2: Average tokens per document.

Furthermore, a few details about the applicability of this linguistic resource in the investigation of translational hypotheses (Ilisei and Inkpen, 2011; Ilisei et al., 2011). In (Ilisei et al., 2011) the hypothesis targeted was the explicitation hypothesis, and brief details regarding their findings are summarised in the following subsection.

3.3. RoTC Corpus Applied in the Investigation of the Explicitation Hypothesis

The Explicitation hypothesis, also assumed to be a universal of translational language (Baker, 1996), states that additional background information which is found implicitly within the message of the source text appears explicitly spelled out in the equivalent translated text. Considering the opposite phenomenon resulting from this hypothesis, ellipsis would occur much more often within the non-translated texts than translational language. Therefore, investigating ellipsis within translated or non-translated texts can lead to findings regarding the explicitation hypothesis. A machine learning system was built for this analysis (Ilisei et al., 2011) and the following section provides brief details of these experiments and their results.

Ellipsis constitutes one of the attributes proposed for the investigation of the explicitation hypothesis. The correct understanding of ellipsis is absolutely essential in the translation process, and hence any type of linguistic resource labelled with this information would be highly appreciated within the domain. As the ellipsis of subjects is the most frequent type, the study focuses only on the anaphoric zero pronoun (hereafter noted as AZP). A tool which uses machine learning techniques is used to identify the verbs which have a zero pronoun in the subject position (Mihăilă et al., 2010; Mihăilă et al., 2011). The software used is known to have an accuracy of 74%.

Before presenting the results of the AZP impact on translational language, the notion of anaphoric zero pronoun is defined. As an agreement between scholars has not yet emerged, anaphora is still a controversial topic and there are thus different classifications of ellipsis (Mladin, 2005). The adopted definition is the following: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood (Mitkov, 2002), 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).

Their findings on the RoTC corpus show that a machine learning system is able to distinguish between translated and non-translated texts relying only on the anaphoric zero pronoun attribute. The accuracy obtained is between 71% and 75% (Ilisei et al., 2011). Therefore, once more it can be emphasised that the monolingual comparable corpus compiled for the Romanian language appears to be a reliable linguistic resource in the investigation of translational hypotheses, and most likely for other domains, such as translation technology. This linguistic resource will be made available online[6] once its documentation is complete.

4. Conclusion

Building comparable corpora for the investigation of translational hypotheses is an important task within the translation studies domain. This paper describes the compilation of a translational comparable corpus for the Romanian language. The resource comprises translated and non-translated news articles and is designed to be used in the investigation of translational language and translational hypotheses. Moreover, a few details about the applicability of this linguistic resource are mentioned: explicitation hypothesis is investigated by analysing the impact of the anaphoric zero pronouns in translational language compared to non-translational one.

5. References

M. Baker, 1993. *Text and Technology: In Honour of John Sinclair*, chapter Corpus Linguistics and Translation Studies Implications and Applications, pages 233–250. Amsterdam & Philadelphia: John Benjamins.

M. Baker. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2):223–43.

M. Baker, 1996. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, chapter Corpus-based Translation Studies:The Challenges that Lie Ahead, pages 175–186. Amsterdam & Philadelphia: John Benjamins.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21, 3:259–274.

V. Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis, University of Hamburg.

L. Borin and K. Prütz. 2001. Through a glass darkly: Part-of-speech distribution in original and translated text. *Language and Computers*, 37(1):30–44.

A. Chesterman, 2000. *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, chapter A Causal Model for Translation Studies, pages 15–27. St. Jerome.

G. Corpas, R. Mitkov, N. Afzal, and V. Pekar. 2008. Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the AMTA*, Waikiki, Hawaii.

G. Corpas. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main, Berlin & New York: Peter Lang.

G. De Sutter and M. Van de Velde. 2008. Do the mechanisms that govern syntactic choices differ between original and translated language? A corpus-based translation study of PP extraposition in Dutch and German. In R. Xiao, L. He, and M. Yue, editors, *Proceedings of the international symposium on using corpora in contrastive and translation studies (UCCTS 2008)*.

EAGLES. 1996. Expert Advisory Group on Language Engineering Standards Guidelines.

A. Frankenberg-Garcia. 2004. Are translations longer than source texts? A corpus-based study of explicitation. In *Third International Corpus Use and Learning to Translate Conference, Barcelona, Spain*, January.

S. Hansen. 2003. *The Nature of Translated Text. An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Ph.D. thesis, Saarbrücken: Saarland University.

S. Hunston. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

I. Ilisei and D. Inkpen. 2011. Translationese Traits in Romanian Newspapers: A Machine Learning Approach. *International Journal of Computational Linguistics and Applications*.

I. Ilisei, C. Mihaila, D. Inkpen, and R. Mitkov. 2011. The Impact of Zero Pronominal Anaphora on Translational Language: A Study on Romanian Newspapers. In *Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2011, Cluj-Napoca, Romania*, July 46.

G. Kennedy. 1998. *An Introduction to Corpus Linguistics*. Amsterdam: Rodopi.

D. Kurokawa, C. Goutte, and P. Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of the MT-Summit*.

S. Laviosa. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam & New York: Rodopi.

G. Lembersky, N. Ordan, and S. Wintner. 2011. Language Models for Machine Translation: Original vs. Translated Texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

B. Maia. 2003. What are Comparable Corpora? In *Proceedings of the Workshop on Multilingual Corpora: Linguistic Requirements and Technical perspectives, Corpus Linguistics*, pages 27–34, Lancaster, U.K.

A. McEnery, 2003. *Oxford Handbook of Computational Linguistics*, chapter Corpus Linguistics, pages 448–463. Oxford: Oxford University Press.

C. Mihăilă, I. Ilisei, and D. Inkpen. 2010. 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)*.

C. Mihăilă, I. Ilisei, and D. Inkpen. 2011. Zero Pronom-

[6] `http://clg.wlv.ac.uk`

inal Anaphora Resolution for the Romanian Language. *Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS"*, 42.

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

C. I. 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.

M. Olohan. 2004. *Introducing Corpora in Translation Studies*. Routledge.

T. Puurtinen, 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, chapter "Nonfinite constructions in Finnish children's literature: Features of translationese contradicting translation universals?", pages 141–154. Amsterdam & New York: Rodopi.

A. Pym, 2008. *Beyond Descriptive Translation Studies*, chapter On Toury's laws of how translators translate, pages 311–328. Benjamins.

D. Tufiş, D. Ştefănescu, R. Ion, and A. Ceauşu, 2008a. *Advances in Multilingual and Multimodal Information Retrieval (CLEF 2007), Lecture Notes in Computer Science*, volume 5152, chapter RACAI's Question Answering System at QA@CLEF 2007, pages 3284–3291. Springer-Verlag, September.

D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu. 2008b. 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, May.

F. M. Tyers and M. Alperen. 2010. South-East European Times: A parallel corpus of the Balkan languages. In *Proceedings of the 7th Language Resources and Evaluation Conference - LREC 2010, Valletta, Malta*.

R. Xiao, L. He, and M. Yue. 2008. In Pursuit of the 'Third Code': Using the ZJU Corpus of Translational Chinese in Translation Studies. In Lianzhen He Richard Xiao and Ming Yue, editors, *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*.