iCorpora: Compiling, Managing and Exploring Multilingual Data

Hernani Costa*[a], Gloria Corpas Pastor[a],
Miriam Seghiri[a] and Ruslan Mitkov[b]
[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas,seghiri}@uma.es, r.mitkov@wlv.ac.uk

Abstract

In the last decade, there has been a growing interest in bilingual and multilingual corpora. Particularly, in translation their benefits have been demonstrated by several authors (cf. Bowker and Pearson (2002); Bowker (2002); Zanettin et al. (2003); Corpas Pastor and Seghiri (2009)). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data are just an example of their advantages. Thus, it is not surprising that the use of corpora has been considered an essential resource in several research domains such as translation, language learning, stylistics, sociolinguistics, terminology, language teaching, automatic and assisted translation, amongst others. Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement on these areas. One potential solution to the insufficient parallel translation data is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES (1996); Corpas Pastor, 2001:158). Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains for example, the problem of data collection presupposes a significant technical challenge. The solution proposed in iCorpora project and presented in this article is to exploit the fact that comparable corpora are much more widely available than parallel translation data. This ongoing project aims to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora by creating a new web-based application from scratch. iCorpora intends to fulfil not only translators' and interpreters' needs (Costa et al. (2014b;a)), but also professionals' and ordinary people's, either by breaking some of the usability problems found in the current compilation tools available on the market (e.g. BootCaT (Baroni and Bernardini (2004)) and WebBootCat (Baroni et al. (2006)) or by improving their limitations and performance issues. iCorpora will aggregate three applications: iCompileCorpora, iManageCorpora and iExploreCorpora. The first application, iCompileCorpora (Costa et al. (2014c)), can be seen as a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer. This design option will permit not only to increase the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer (i.e. the CLIR layer will automatically translate the queries to other languages (Talvensaari et al. (2007))). iManageCorpora will be specially designed to: manage (i.e. it will allow to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and to explore the representativeness of the corpora (cf. Corpas Pastor and Seghiri (2009)). Finally, iExploreCorpora intends to offer a set of concordance features, such as search for words in context, automatic extraction of the most frequent words and multi-words, amongst other.

* Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. n FFI2012-38881, 2012-2015), and the R&D Project for Excelence TERMITUR (ref. n HUM2754, 2014-2017).

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014c). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems*, 25(1).

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.