iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora

Hernani Costa
University of Malaga

Gloria Corpas Pastor
University of Malaga

Miriam Seghiri
University of Malaga

ABSTRACT

This article presents an ongoing project that aims to design and develop a robust and agile web-based application capable of semi-automatically compiling monolingual and multilingual comparable corpora, which we named iCompileCorpora. The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic and a Cross-Language Information Retrieval (CLIR) layer. This design option will not only permit to increase the flexibility of the compilation process, but also to hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer. The manual layer presents the option of compiling monolingual or multilingual corpora. It will allow the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of either monolingual or multilingual corpora mined from the Internet. As nowadays there is an increasing demand for systems that can somehow cross the language boundaries by retrieving information of various languages with just one query, the third layer aims to answer this demand by taking advantage of CLIR techniques to find relevant information written in a language different from the one semi-automatically retrieved by the methodology used in the previous layer.

1. Introduction

The interest in mono-, bi- and multilingual corpora is vital in many research areas such as language learning, stylistics, sociolinguistics, translation studies, amongst other research areas. Particularly in translation, their benefits have been demonstrated by various authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor and Seghiri, 2009). The main advantages of its usage are their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data. In detail, corpus linguistics:

- Empowers the study of the foreign language: the study of the foreign language with the use of corpora allows the foreign language learners to get a better “feeling” about

that language and learn the language through "real world" texts rather than "controlled" texts (cf. Gries, 2008).

- Simplifies the study of naturalistic linguistic information: as previously mentioned, a corpus assembles "real world" text, mostly a product of real life situations, which results in a valuable research source for dialectology (cf. Hollmann and Siewierska, 2006), sociolinguistics (cf. Baker, 2010) and stylistics (cf. Wynne, 2006), for example.

- Helps linguistic research: as the time needed to find particular words or phrases has been dramatically reduced with the use of electronically readable corpora, a research that would take days or even weeks to be manually performed can be done in a couple of seconds with an high degree of accuracy.

- Enables the study of wider patterns and collocation of words: before the advent of computers, corpus linguistics was studying only single words and their frequency. More recently, the emergence of modern technology allowed the study of wider patterns and collocation of words (cf. Roland et al., 2007).

- Allows simultaneous analysis of multiple parameters: in the last decades, the development of corpus linguistic software tools helped the researchers to analyse a wider number of parameters simultaneously, such as determine how the usage of a particular word and its syntactic function varies.

Moreover, they are a suitable tool for translators, as they can easily determine how specific words and their synonyms collocate and vary in practical use or even help interpreters speeding up the research for unfamiliar terminology (cf. Costa et al., 2014). Furthermore, in the last decade, a growing interest in bi- and multilingual corpora has been shown by researchers working in other fields, such as terminology and specialised language, automatic and assisted translation, language teaching, Natural Language Processing, 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 corpora is the exploitation of non-parallel bi- 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. Moreover, the difficulty of retrieving and classifying such data is considered a complex issue as there is no unique notion of what it really covers and how it can be truly exploited (cf. Skadina et al., 2010:12).

2. Existing Corpora Compilation Solutions

Although this compilation process could be manually performed, nowadays specialised tools can be used to automate this tedious task. By a way of example, BootCaT (Baroni and Bernardini,

2004) was built to exploit specialised monolingual corpora from the Web. It is capable of compiling a corpus through automated search queries, and only requires a small set of seed words as input. This tool has been used, for example, to create specialised comparable corpora for travel insurance (Corpas Pastor and Seghiri, 2009), medical treatments (Gutiérrez Florido et al., 2013), among other narrow-domains. WebBootCat (Baroni et al., 2006) is similar to BootCaT, but instead of having to download and install the application, WebBootCat can be used online. Despite being designed for other purposes, Terminus and Corpográfo should also be mentioned as examples of web-based compilation tools.

As we can see, several semi-automatic compilation tools have been proposed so far. Nevertheless, these compilation tools are scarce or proprietary, simplistic with limited features, built to compile one monolingual corpus at a time and do not cover the entire compilation process (i.e. apart from compiling monolingual comparable corpora, they do not allow managing and exploring both parallel and multilingual comparable corpora). Thus, their simplicity, lack of features, performance issues and usability problems result in a pressing need to design new compilation tools tailored to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's.

3. iCompileCorpora

Departing from a careful analysis of the weaknesses and strengths of the current compilation solutions, we started by designing and developing a robust and agile web-based application prototype to semi- automatically compile mono- and multilingual comparable corpora, which we named iCompileCorpora. iCompileCorpora can be simply described as a Web graphical interface that will guide the user through the entire corpus compilation process. Designed and implemented from scratch, this application aims to cater to both novice and experts in the field. It will not only provide a simple interface with simplified steps, but also will permit experienced users to set advanced compilation options during the process.

The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer (see Figure 1). 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. In detail, the manual layer represents the option of compiling monolingual and multilingual corpora. It will allow for the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of both mono- and multilingual corpora mined from the Internet. Although this layer can be considered similar to the approaches used by BootCaT and WebBootCat, it has been designed to address some of their limitations (e.g. allow the use of more than one boolean operator when creating search query strings), and to improve the User Experience (UX) with this type of software. As nowadays there is an increasing demand for systems that can somehow cross the language boundaries by retrieving information in various languages with just one query, the third layer aims to answer this demand by taking advantage of CLIR techniques to find relevant information written in a language different to the one semi-automatically retrieved by the methodology used in the previous layer.

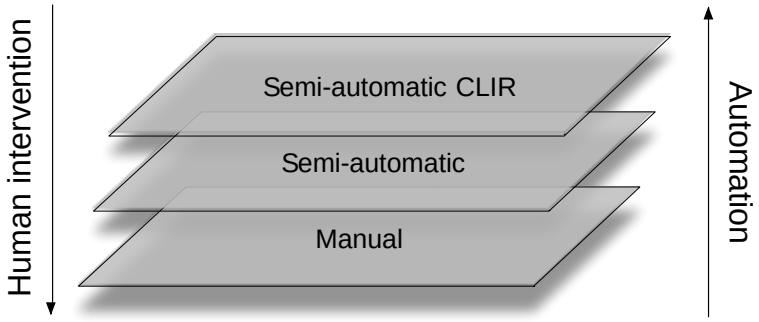

Figure 1: iCompileCorpora layered model.

4. Conclusion

This article presents an ongoing project that aims to increase the flexibility and robustness of the compilation of monolingual and multilingual comparable corpora by creating a new web-based application from scratch. iCompileCorpora intends to fulfil not only translators' and interpreters' needs, 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 or by improving their limitations and performance issues. By the end of this project, we intend to make this compilation tool publicly available, both in a research or in a commercial setting.

Acknowledgements

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 no 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. no FFI2012-38881, 2012-2015), and the R&D Project for Excellence TERMITUR (ref. no HUM2754, 2014-2017).

References

BAKER, P. (2010). Sociolinguistic and Corpus Linguistics. Edinburgh University Press, Edinburgh, UK.

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., POMIKALEK, J., AND RYCHLY, 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 ensenanza 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, John Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

COSTA, H., CORPAS PASTOR, G., AND DURÁN MUNOZ, I. (2014). Technology-assisted Interpreting. MultiLingual #143, 25(3):27–32.

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

GRIES, S. T. (2008). Corpus-based methods in analyses of SLA data, pages 406–431. Routledge, NY, USA.

GUTIÉRREZ FLORIDO, R., CORPAS PASTOR, G., AND SEGHIRI, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In 10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounter, Paris, France.

HOLLMANN, W. AND SIEWIERSKA, A. (2006). Corpora and (the need for) other methods in a study of Lancashire dialect. Zeitschrift fur Anglistik und Amerikanistik, 1(54):203–216.

ROLAND, D., DICK, F., AND ELMAN, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. Journal of Memory and Language, 57(3):348–379.

SKADINA, I., VASILJEVS, A., SKADINS, R., GAIZAUSKAS, R., TUfiS¸, D., AND GORNOSTAY, T. (2010). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In 3rd Workshop on Building and Using Comparable Corpora (BUCC'10), pages 6–14, Valletta, Malta.

WYNNE, M. (2006). Stylistics: corpus approaches. Encyclopedia of Language and Linguistics, 12(2):223–226.

ZANETTIN, F., BERNARDINI, S., AND STEWART, D. (2003). Corpora in Translator Education. Manchester: St. Jerome Publishing.