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

Hernani Costa, Gloria Corpas Pastor and Miriam Seghiri
{hercos,gcorpas,seghiri}@uma.es

LEXYTRAD, University of Malaga
Malaga, Spain

Introduction

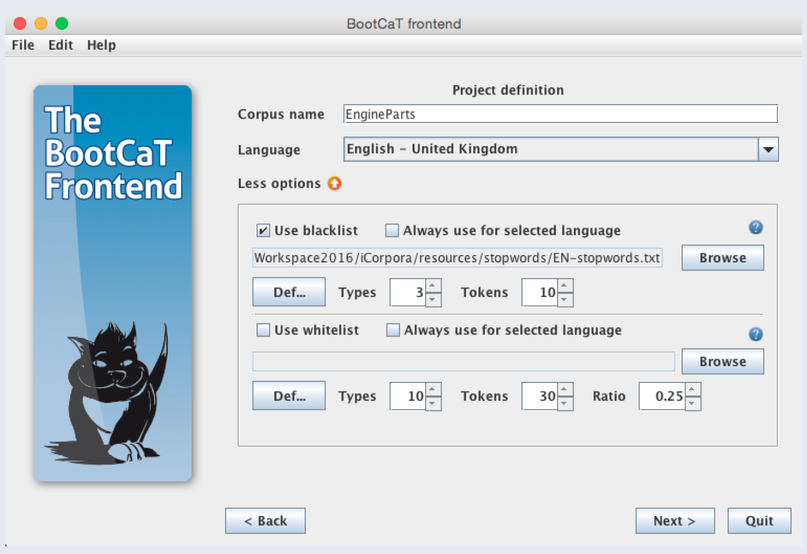
- The interest in mono-, bi- and multilingual corpora is vital in many research areas, such as:
  - terminology and specialised language
  - automatic and assisted translation
  - language teaching
  - natural language processing
  - amongst other research areas
- Particularly in translation, their benefits have been demonstrated by various authors [1, 2, 3, 4]

Comparable Corpora

- It is already a fact that using comparable corpora [5] is the solution for the lack of sufficient/up-to-date parallel corpora and linguistic resources, specially for narrow domains and poorly-resourced languages
- Some of the advantages of using comparable corpora are the following:
  - objectivity
  - reusability
  - multiplicity and applicability of uses
  - easy handling and quick access to large volume of data

Existing Corpora Compilation Solutions and their limitations

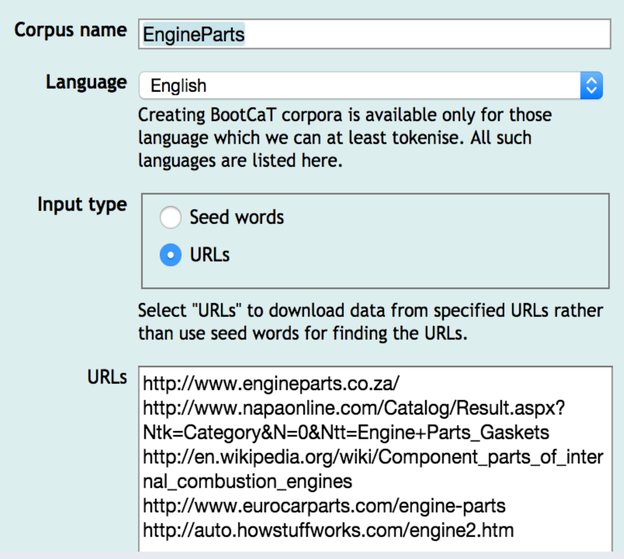
BootCaT [6]



Current limitations

- compilation tools are scarce or proprietary
- simplistic with limited features
- built to compile one monolingual corpus at a time
- or do not cover the entire compilation process (i.e. they do not allow managing and exploring both parallel and multilingual comparable corpora)

WebBootCaT [7]



iCompileCorpora

Manual

- Represents the option of compiling monolingual and multilingual corpora
- Allows for the manual upload of documents from a local or remote directory

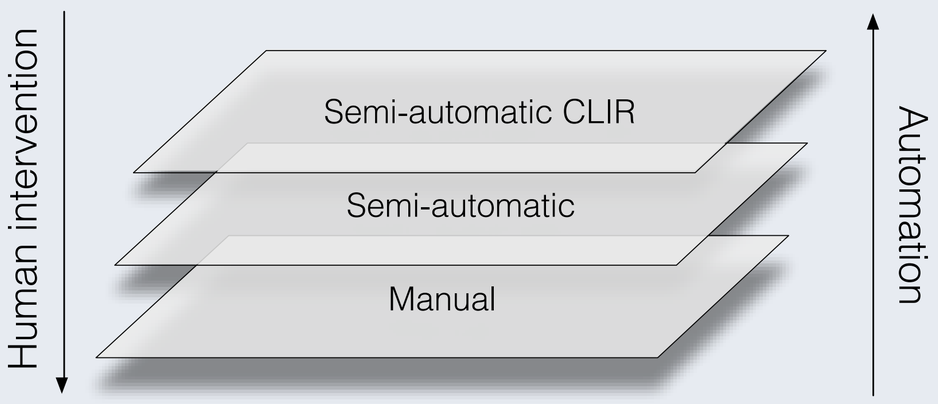
Semi-automatic

- Permits the exploitation of both mono- and multilingual corpora mined from the Internet
- Addresses some limitations in current solutions, such as: the use of more than one boolean operator when creating search query strings

Semi-automatic CLIR

- Address the demand for multilingual corpora by taking advantage of CLIR techniques
- Allows for the retrieval of relevant information written in a language different to the one semi-automatically retrieved by the *semi-automatic* layer

iCompileCorpora Layered Model



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 Excelence TERMITUR (ref. nº HUM2754, 2014-2017).

References

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

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

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

[4] G. Corpas Pastor and M. Seghiri, "Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish)," in *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* (A. Beeby, P. Inés, and P. Sánchez-Gijón, eds.), Benjamins translation library, ch. 5, pp. 75–107, John Benjamins Publishing Company, 2009.

[5] EAGLES, "Preliminary Recommendations on Corpus Typology," tech. rep., EAGLES Document EAG-TCWG-CTYP/P., May 1996. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

[6] M. Baroni and S. Bernardini, "BootCaT: Bootstrapping Corpora and Terms from the Web," in *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pp. 1313–1316, 2004.

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