iCorpora: Compiling, Managing and Exploring Multilingual Data

Hernani Costa[1a], Gloria Corpas Pastor[1], Miriam Seghiri[1] and Ruslan Mitkov[2]
{hercos,gcorpas,seghiri}@uma.es, r.mitkov@wlv.ac.uk

[1]LEXYTRAD, University of Malaga, Spain
[2]RIILP, University of Wolverhampton, UK

[a]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).

Introduction

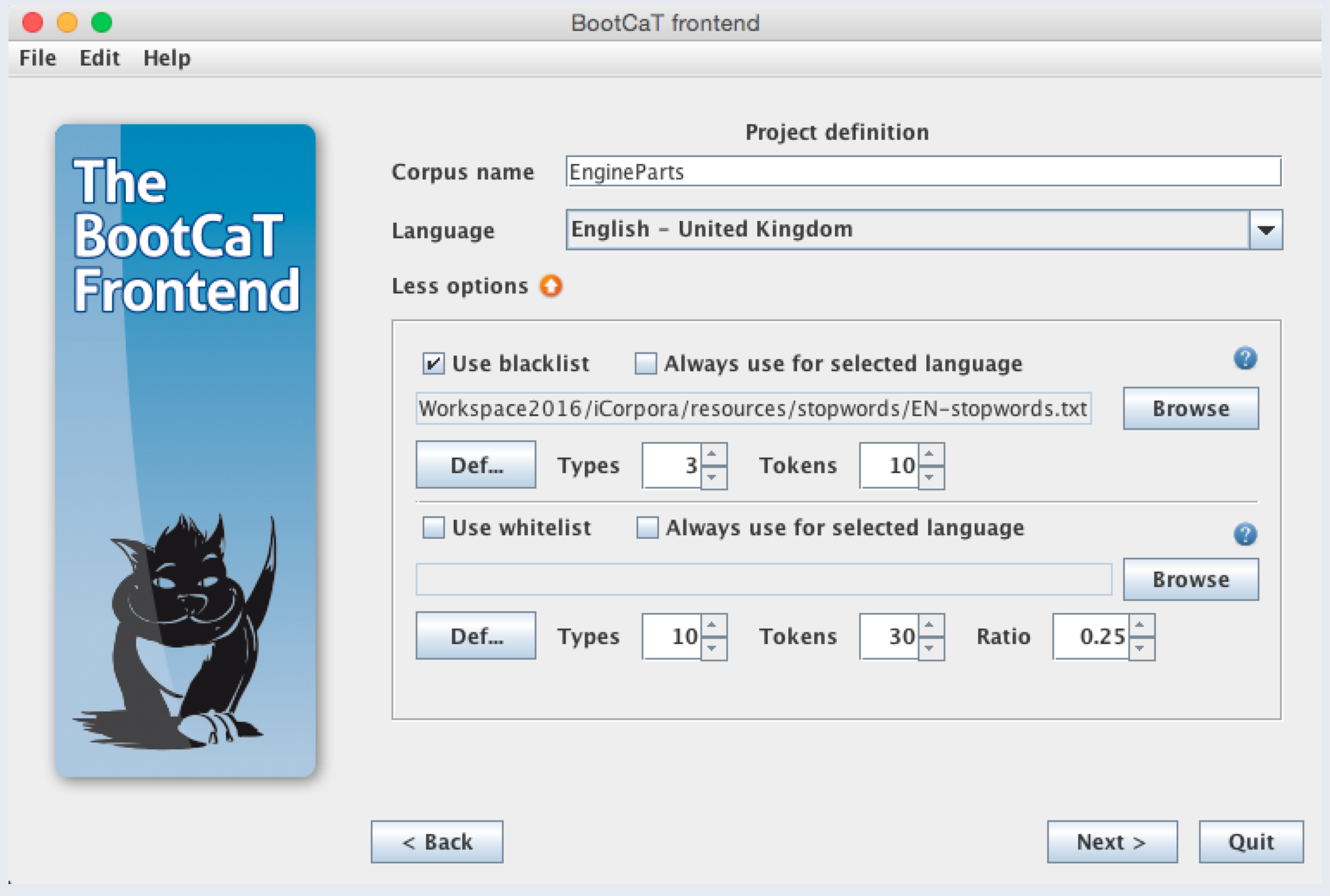
- The interest in mono-, bi- and multilingual corpora is vital in many research areas, such as:
  - terminology and specialised language
  - automatic and assisted translation
  - language teaching
  - natural language processing
- Particularly in translation, their benefits have been demonstrated by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

Objectives

- Against the background of increasing importance of comparable corpora given the scarcity of suitable parallel corpora, iCorpora's objectives are:
  - to develop of a novel flexible and robust web-based application for compilation, management and exploitation of comparable corpora
  - to address the needs of translators and interpreters as well as other professional and casual users

Existing Corpora Compilation Solutions and their limitations

BootCaT [5]



BootCaT frontend

Project definition

Corpus name: EngineParts

Language: English – United Kingdom

Less options

☑ Use blacklist ☐ Always use for selected language

Workspace2016/iCorpora/resources/stopwords/EN-stopwords.txt Browse

Def... Types 3 Tokens 10

☐ Use whitelist ☐ Always use for selected language

Browse

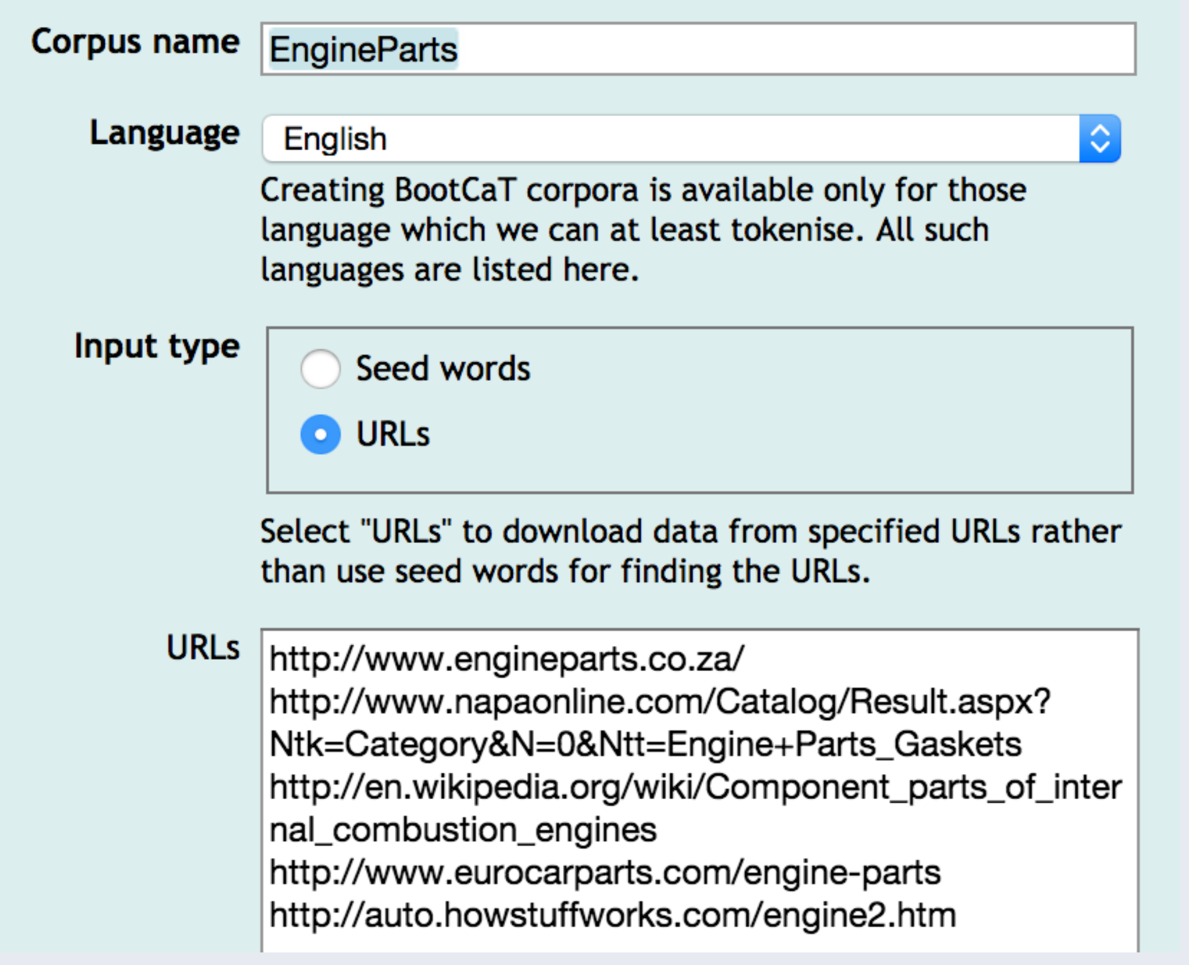
Def... Types 10 Tokens 30 Ratio 0.25

< Back Next > Quit

Current limitations

- compilation tools are scarce or proprietary
- simplistic with limited features [6]
- 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]



Corpus name: EngineParts

Language: English

Creating BootCaT corpora is available only for those language which we can at least tokenise. All such languages are listed here.

Input type: ○ Seed words ● URLs

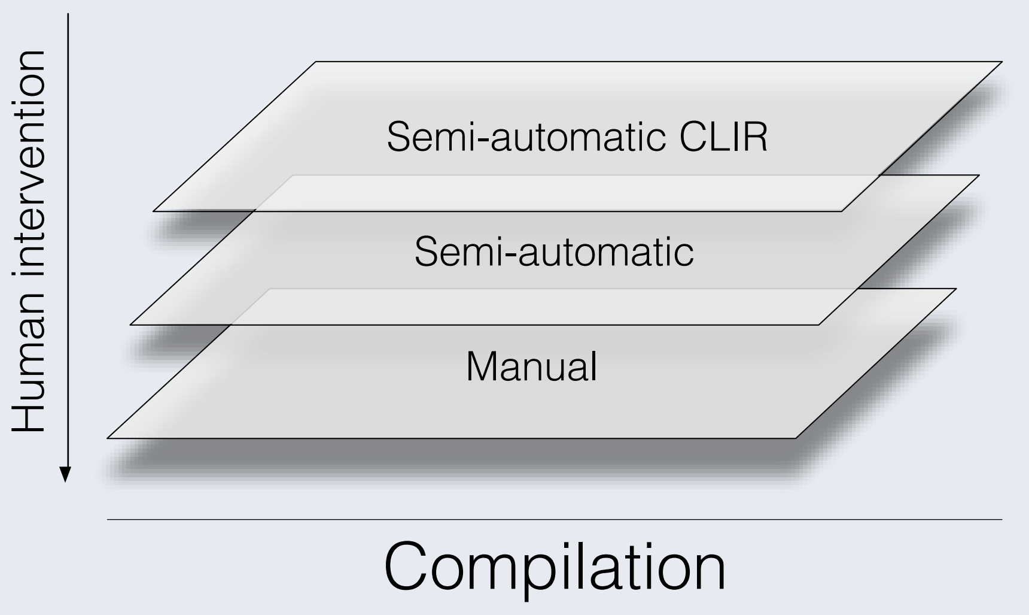
Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

URLs:
http://www.engineparts.co.za/
http://www.napaonline.com/Catalog/Result.aspx?Ntk=Category&N=0&Ntt=Engine+Parts_Gaskets
http://en.wikipedia.org/wiki/Component_parts_of_internal_combustion_engines
http://www.eurocarparts.com/engine-parts
http://auto.howstuffworks.com/engine2.htm

iCorpora

iCompileCorpora

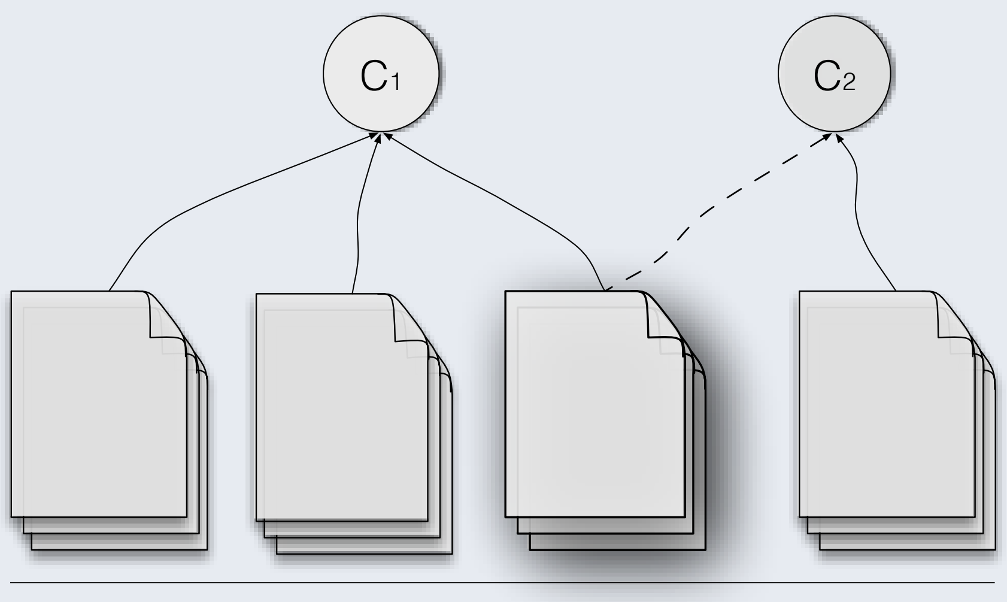
- Offers the functionality of compiling monolingual and multilingual corpora
- Allows for the manual upload of documents from a local or remote directory



Compilation

iManageCorpora

- Management and reusability:
  - edit, copy and paste sentences and documents from and to documents and corpora, respectively
  - measure the similarity between documents
  - explore the representativeness of the corpora



Management

iExploreCorpora

- It will allow to:
  - search for words in context
  - extract the most frequent words and multi-words

Exploration

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] 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.

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

[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.

AIETI7 | January, 2015 | Malaga, Spain
Hernani Costa (hercos@uma.es)