iCorpora: Compiling, Managing and Exploring Multilingual Data

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

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

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 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>compilation tools are scarce or proprietary</td>
<td>simplistic with limited features [6]</td>
<td></td>
</tr>
<tr>
<td>built to compile one monolingual corpus at a time</td>
<td>or do not cover the entire compilation process (i.e. they do not allow managing and exploring both parallel and multilingual comparable corpora)</td>
<td></td>
</tr>
</tbody>
</table>

iCorpora

<table>
<thead>
<tr>
<th>iCompileCorpora</th>
<th>iManageCorpora</th>
<th>iExploreCorpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offers the functionality of compiling monolingual and multilingual corpora</td>
<td>Management and reusability:</td>
<td>It will allow to:</td>
</tr>
<tr>
<td>Allows for the manual upload of documents from a local or remote directory</td>
<td>• edit, copy and paste sentences and documents from and to documents and corpora, respectively</td>
<td>• search for words in context</td>
</tr>
</tbody>
</table>

References