

# CNGL Undergraduate Students as Researchers Programme 2011 **PROJECT DESCRIPTION**

<b>Institution/Track:</b>	Dublin city University	
<b>Project Title:</b>	Arabic Multiword Expressions Annotation	
<b>Suitable for students who are studying in the following areas:</b>	Computational linguistics, linguistic annotation of natural language corpora.	
<b>Skills needed:</b>	Computer Science and a good knowledge of Arabic grammar would be ideal.	
<b>Project Description:</b>	<p><b>The research project aims to manually validate and annotate an automatically built lexicon of Arabic named entities (NEs) and multiword expressions (MWEs).</b></p> <p>Our research group have investigated the automatic acquisition of NEs and MWEs combining different computational approaches and using available language resources such as: Wikipedia, WordNet and the Arabic Gigaword corpus.</p> <p><b>To the best of our knowledge, we have built the largest, most mature and well-structured Arabic NEs-MWEs lexical resource to date.</b> In fact, we have automatically created a <b>lexicon of 70 000 entries</b> and we have published 2 papers describing the research.</p> <p>The evaluation of this dataset against gold standards has revealed the good overall quality of the collected information, yet there is a wide discrepancy in the accuracy scores for different sections of the lexicon ranging from 98% to 63%. Therefore, in order to release this lexicon, it is essential to carry out a complete manual verification to give a higher value to the product. If the proposal is accepted we will be able to release the lexicon with 99% accuracy for the entire collection.</p>	
<b>he Role of the student &amp; benefits gained from participation in this project:<sup>1</sup></b>	<ul style="list-style-type: none"> <li>• Learn about research team environment and work</li> <li>• Practice communication skills</li> <li>• Develop research and programming skills</li> <li>• learn about: annotation procedures, annotation evaluation, annotation software and frameworks, etc</li> <li>• Publication might be expected</li> </ul>	
<b>Who will be working with you?</b>	Dr. Lamia Tounsi	
<b>Short description of the group:</b>	<p>The National Centre for Language Technology is based in the School of Computing in Dublin city University. The Centre carries out basic and applied research in the areas of machine translation, natural language parsing, grammar induction, question answering, sentiment analysis, computer-aided language learning, software localisation, speech recognition and speech synthesis. Its researchers are drawn from the School of Computing, the School of Applied Languages and Intercultural Studies and the school od Electroninis Engineeering. The Centre is affiliated with the Center for Next Generation Localisation.</p> <p><a href="http://www.nclt.dcu.ie/">http://www.nclt.dcu.ie/</a></p>	
<b>Recommended Reading Material:</b>	<p>1) Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith. 2010. Automatic Extraction of Arabic Multiword Expressions. COLING 2010 Workshop on Multiword Expressions: from Theory to Applications, China.</p> <p>2) Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini and Josef van Genabith. 2010. 'An automatically built Named Entity lexicon for Arabic'. LREC 2010, Malta.</p>	
<b>Other information:</b>		
<b>For further details on this project please contact:</b>	<b>Name:</b> <b>Phone:</b> <b>E-Mail:</b> <b>Website:</b>	<b>Dr. Lamia Tounsi</b> <b>(0)1 700 6905</b> <a href="mailto:lamia.tounsi@computing.dcu.ie">lamia.tounsi@computing.dcu.ie</a>

<sup>1</sup> *This is an initial description of the role of the student and it is liable to change following discussions between the investigators and the student.*