



Institute of Technology

Ciência sem Fronteiras / Science Without Borders

Postgraduate Project Template

Institution:	Institute of Technology Blanchardstown Dublin, IRELAND
Title of Postgraduate Opportunity: (include level of study)	PhD – level 10
PI Name & Contact Details:	Dr. Brian Nolan Head of Department of Informatics Institute of Technology Blanchardstown Blanchardstown Road North Blanchardstown Dublin 15 IRELAND Email: brian.nolan@itb.ie
Department/School:	Informatics
Research Centre /Group:	Computational and Functional Linguistics Research Group
Research Centre/Group website:	
<p>Brief Summary of PI research / research group /centre activity</p> <p>The key research themes/interests of this group incorporates aspects of computer science, linguistics, computational linguistics, rich media plus spoken and Sign languages</p> <p>This domain is concerned with the treatment and processing of human language within a functional linguistic paradigm and how this can be modelled within software. It is also concerned with the development of robust language aware Internet enabled software applications that provide motivated links between the interface between the semantics and syntax of human languages. We welcome research proposals in the areas of linguistically motivated machine translation of human languages using interlingua-bridge approaches, computational processing of human sign languages, the structural description of language and the use of digital corpora, computational frameworks for multilingual agents, and the development and deployment of conversational agents and avatars. We are also interested in matter of linguistic complexity and the computational adequacy of linguistic models of human languages.</p>	

Brief Description of PhD /Post-Doc Project

This research project proposes to create a proof-of-concept rule-based machine translation application in software that will accept as input a source text of one of a choice of native Brazilian languages (from a restricted domain) and generate English text. To achieve this we assemble a corpus native Brazilian data validated by native Brazilian speakers. For Brazilian languages we mean any of Brazilian Portuguese, indigenous Brazilian languages or Brazilian Sign language. The major aim is to build in software a system can accept input of text in the source language and translate it in real time to English text. The secondary aim in this to present an extensible interlingua architecture which is not only successful in translating (simple) source text sentences to corresponding English sentences but also does it in most optimal way that adheres to the principles of functional linguistics, as articulated by Role and reference Grammar (RRG), and incorporating a strong model of the lexicon and a bi-directional linking system across the syntax-semantics interface. This research project will build a lexically driven machine translation engine that by design is extensible to other languages. The research strategy will be built on an interlingua bridge architecture to facilitates translation from the Brazilian source language to English into one domain across the three general stages for machine translation: parsing, transfer and generation.

According to [Wikipedia], the official language of Brazil is Portuguese and almost all of the population speaks this. An exception to this is Brazilian Sign Language, more commonly known by its Portuguese acronym LIBRAS, in education and government services. The language must be taught as a part of the education and speech and language pathology curricula. Brazilian Portuguese has had its own development, mostly similar to 16th century Central and Southern dialects of European Portuguese, with some influences from the Amerindian and African languages, especially West African and Bantu. As a result, the language is somewhat different, mostly in phonology, from the language of Portugal and other Portuguese-speaking countries (the dialects of the other countries, partly due to the more recent end of Portuguese colonialism in these regions, have a closer connection to contemporary European Portuguese). These differences are comparable to those between American and British English. Brazil is the only Portuguese-speaking nation in the Americas, making the language an important part of Brazilian national identity and giving it a national culture distinct from those of its Spanish-speaking neighbours. Minority languages are spoken throughout the nation. One hundred and eighty Amerindian languages are spoken in remote areas and a significant number of other languages are spoken by immigrants and their descendants. [Wikipedia] <http://en.wikipedia.org/wiki/Brazil#Language>

The **research hypothesis** is that it is possible, using current software models, to design a (proof-of-concept) rule-based MT system, with a knowledge of the morphological and syntactic facts of the language, that is sufficiently powerful and robust to translate one of the native Brazilian source language source text to English within a restricted domain of language usage.

To satisfy the hypothesis, we need to answer the following research questions:

1. How can an MT system read, understand and parse native Brazilian source language source text that is presented, for example, in the traditional orthography, and with full rich agreement features, as appropriate.
2. How can an MT system manage the person, gender and number agreement that is morphologically rich in the particular native Brazilian source language text and successfully and accurately generate text in a morphologically impoverished language like English.
3. What level of morphological analysis is required to be undertaken on the particular native Brazilian source language and how can this be achieved using computational linguistic techniques within software in real-time.

4. How can these computational linguistic techniques inform a sentence level syntactic analysis understood within a functional Role and reference grammar characterisation of the particular native Brazilian language (from the languages defined earlier above).
5. How can the information retrieved from the morphological analysis, including, person, number, gender and consonant templates for vowel insertion be represented for the parsed source language sentence or utterance.
6. How can this be optimally used to motivate the generation of the equivalent English sentence representation and the corresponding English sentence as text.

The project consists of the following general **work-packages**:

1. Following a read-in period, we will **collect a corpus of native source language data** for a restricted domain.
2. On this corpus, we will perform a **morphological analysis of the native source language data** and to identify the relevant word level inflectional and derivation rules along with the application of the constructional templates in the lexicon and constructional repository within the RRG linguistic model.
3. We will perform a **syntactic analysis of the native source language data** and determine how grammatical function and agreement is encoded. We will to identify the rules that operate at the level of the declarative source language sentence.
4. We will **implement a suitable meta-representation scheme** for the information coded by the morphological analyser and the parser that can be instantiated in software.
5. We will **implement a model for a grammar of the** source language with lexicon, lexical items and grammatical rules according to the linguistic principles of Role and reference Grammar (RRG). We will design and construct the morphological analyser in software. We will determine a method of interacting 'on-screen' with native orthographic script in addition to the processing of collected corpus. We will design and construct the parser to process the output of the morphological analyser.
6. We will design, build and **implement in software the transfer strategy**, from source language to English, of the meta-information collected at the morphological analyser and the parser stages.
7. Using the transfer strategy, we will **implement an English sentence generator** that will be capable of producing a grammatically correct and meaningful equivalent English sentence. This will be rigorously tested using bi-lingual native source language speakers.
8. We will **evaluate and critique the results, strategies and techniques** designed and employed in this research and contrast with practice in the literature. This will be followed by results dissemination.

Key Attributes of Project for Brazilian Postgraduate Students

The PhD research will have the opportunity to develop a grammar of one of a number of the Brazilian languages and to incorporate this into software with leading linguistic and computational researchers in a functional linguistic context. The research student will have access to software and labs to support the research in an intellectually supporting environment. This research is of value in in the application of software techniques in computational functional linguistics to support Brazilian languages ad culture.

Name and contact details for project queries, if different from PI named above:

As above

Please indicate graduate disciplines which are eligible for application:

- Computer Science with strong programming skills.

- Strong linguistic background, ideally to masters level.

Alignment with Science Without Borders Priority Areas:

Please indicate the specific programme priority area under which the proposed postgraduate project fits – choose only one (tick box)

Information and Communication Technologies (ICTs)	Y