MASTER THESIS PROPOSAL

Supervision by Prof. John Nerbonne and Dr. Marjolijn Verspoor

M.A student: *Victor Dias de Oliveira Santos (EM LCT Masters)*

BACKGROUND

The automatic assessment of students' level in a second or foreign language (referred to from now on as "L2") has for quite some time now been a sort of "holy grail" in Applied Linguistics. Being able to automatically (i.e, computationally) classify someone's piece of writing in a foreign language into one of a number of possible levels has many desired implications. Some of them are:

a) *a welcome decrease in the need to depend solely on humans who are able to manually analyze the essays and decide on the corresponding level*. By using an automatic system, we can not only improve efficiency (being able to classify many more essays than would be possible with a manual process), but also reduce the amount of financial resources required, while increasing the reliability of the classification (avoiding thus possible biases or differences in classification due to differences in opinions and beliefs of those carrying out the manual analysis).

b) *Such a system can be easily tested for its accuracy by comparing its classifications to human judgements (gold standard)*. Once the system has been proved to be sufficiently accurate, it can be applied to an unlimited number of essays and also used by different companies, institutions or governments.

c) *Such a system, in which a great number of indicative linguistic features (variables) are qualitatively and quantitatively analyzed, can be used to deepen the understanding of the influence of one's first language (L1) on the L2 in case*. Many of the patterns found in the L2 essays can be accounted for by taking into consideration what the writer's L1 is. By comparing large sets of essays from different L1 backgrounds, we arrive at significant conclusions regarding the manner of influence of specific L1s on a specific L2. This knowledge can, in its turn, be used, among other things, to guide practices in the language learning classroom.

To arrive at such a system, one needs therefore to have the following: *a large body of essays in the L2 classified into different levels* (called the "training set" and used for training the algorithm/program), *a fixed number of features/variables and their respective measurements/values*,(which will be taken into consideration in order to classify future essays) *and lastly a test set*, containing essays which will then be classified into different levels by the newly-built classifier and against which the accuracy of the classifier can be checked.

Many recent attempts to develop (automatic) language level classifiers only take into account a small number of features considered to be relevant and indicative of one's proficiency level. Common features (which are many times, however, not quite linguistically-motivated) are *Mean Word Length, Type-Token Ratio, Mean Period Unit Length, Unique Bigram Ratio* (Schulze et all, 2008), *Amount of Subordination* (Michel et all, 2007) and others. Some researchers also make use of some more linguistically motivated features, such as raw tallies of *Verbal Morphology* (verb classes, tense, etc) , *Syntactic structures* (infinitival sentences, wh-clauses, etc) and others (Ellis and Yuan, 2005).

RESEARH QUESTIONS

RQ 1: Can a linguistically-sound computational model/approach assign proficiency levels to:

a) groups of students?

b) individual students?

RQ2: Can such a system be nearly or just as accurate in its assessment as humans with knowledge in the field would?

RQ3: By use of statistical analysis and methods, can we identify those features which correlate highly with each other and therefore use this knowledge in order to optimize our system? Can we also automatically detect those features which are the most distinctive as well as which weight each feature should have in our classifier?

THE DATA SET

In our research for this Masters' thesis, we will take advantage of the great amount of available hand-coded material collected by Dr. Marjolijn Verspoor. The aforementioned professor has collected thousands of essays in English by Dutch speakers from different schools levels, all of which have been judged holistically and classified into one out of six possible proficiency levels by a team of experts (gold standard). A sub-set of these essays (about 500) have been hand-annotated for approximately 60 features. This body of essays, i.e, the *corpus*, has been collected within the framework of the OTTO project and financed by the OCW (Dutch Ministry of Education), the European Platform and the Network of TTO (Tweetalig Onderwijs = Bilingual Education).

INTENDED APPROACH

We intend to answer the research questions mentioned previously and to build our automatic classifier by taking the following steps:

1) *Pre-processing of the data set* (already in digital version), so that it is in a format which is optimal for us to extract the measurements of the variables we are interested in (around 60) and to be used in a classification system.

2) *Use of statistical methods and data-mining/ machine-learning software* (such as WEKA) in order to come to a conclusion as to which of our features correlate with each other and to guide us with regard to the weights to be used for each feature in our classification system. We intend to use a Bayesian classifier in our project.

3) *Check the performance of our classifier against a development set* (manually classified essays which were not used during the training phase and which can give us a reliable idea about the performance and accuracy of our system as it stands).

4) *Check the real accuracy of our classifier against a test/final set* (manually classified essays which were not used during either the training or development phase).

WORKPLAN (2011)

February / March – *Background reading and data-processing*

April / May – *Statistical Analysis and Machine-Learning*

June – *Development of classifier*

July – *Testing and Improvement of Classifier*

August / September – *Writing and defense of thesis*

REFERENCES:

Bird, S., Klein, E., Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Lu, X., Thorne, S. L., & Gamson, D. (submitted). Toward a Framework for Computational Assessment of Linguistic Complexity of Grade-level Reading Materials. *Journal of Applied Linguistics.*

Manning, C.D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Norris, J.M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics, 30(4), 555-578*.

Ortega, L.(2003). Syntactic Complexity Measures and Their Relationship to L2 Proficiency: A Research Synthesis of College-Level L2 writing. *Applied Linguistics, 24(4), 492-518*.