Discovering Regularities in Non-Native Speech

Julie Carson-Berndsen [1], Ulrike Gut [2] & Robert Kelly [1]

[1] Department of Computer Science,
University College Dublin, Ireland,
{Julie.Berndsen, Robert.Kelly}@ucd.ie
Fax: +353 1 2697262

[2] Faculty of Linguistics and Literary Studies,
University of Bielefeld, Germany
gut@spectrum.uni-bielefeld.de
Fax: +49 521 106 6008

This paper presents ongoing collaborative research which focuses on the application of computational linguistic techniques to the analysis of a corpus of native and non-native speech. The aim of this research is to use computational tools for phonological acquisition and representation to identify regularities and sub-regularities between different speaker groups. The corpus is being collected and annotated at different levels as part of ongoing research into the acquisition of prosody by non-native speakers at the University of Bielefeld (see Milde & Gut, 2001). The computational tools have been designed and implemented at University College Dublin as part of a suite of tools aimed at providing a development environment for modeling, testing and evaluating phonotactic descriptions of lesser studies languages (Carson-Berndsen, 2002). The two hitherto separate research directions have now come together to apply computational linguistic tools to a corpus-based investigation of non-native speaker phonotactics. The term *phonotactics* refers to the permissible combination of sounds in a language. There are various ways of acquiring representations of phonotactic constraints. One approach is to manually construct a set of rules based on the linguistic intuitions of a native speaker. Another approach is to learn such constraints from a data set. The latter is the approach taken in this paper.

Currently the corpus consists of 253 annotated recordings of between 2 and 30 minutes length by 88 different speakers with 21 different native languages. The corpus is annotated at a number of linguistic levels such as the level of the intonational phrase, the word, the syllable, and the skeletal (CV) structure, and comprises annotations of the prosodic structures of intonation and pitch range. Each level of annotation is viewed as a *tie*r, analogous to the representations of autosegmental phonology (Goldsmith 1990). Analysis can take place either with respect to individual tiers or with respect to an associated set of tiers. In the latter case, one tier is chosen as the primary tier and the others are associated with it in terms of overlap and precedence relations between the units as suggested in Carson-Berndsen (1998: 60). Using the computational linguistic tools, finite state automaton and finite state transducer representations of the tiers are extracted automatically from the annotated corpus. Regularities in the data are then identified either with respect to a single tier or with respect to an associated set of tiers.

The majority of previous studies on the acquisition of phonotactic constraints are based on small numbers of participants, which reflects the time-consuming nature of a manual analysis of this kind of data. Results on an initial subset of the corpus for Italian and Polish speakers of German demonstrate that the task of identifying phonotactic regularities and sub-regularities in the corpus data can be performed elegantly and efficiently using the computational linguistic tools. Distinct differences between the phonotactic violations produced by the Italian and the Polish speakers have been found and will be documented in detail in the full paper. Further experiments are now underway on a larger corpus.