# Towards the Automatic Lemmatization of 16th Century Mexican Spanish: a Stemming Scheme for the CHEM

Alfonso MEDINA-Urrea

GIL, Instituto de Ingeniería, UNAM
Ciudad Universitaria, 04510 Coyoacán, DF, MEXICO
amedinau@ii.unam.mx

**Abstract.** Two of the problems that should arise when developing a stemming scheme for diachronic corpora are: (1) morphological systems of natural languages may vary throughout time, and these changes are normally not documented sufficiently; and (2) they exhibit very diverse orthographic characteristics. In this short paper, a stemming strategy for a diachronic corpus of Mexican Spanish is briefly described, which partially faces up to these problems. Success rates of the method are contrasted to those of a Porter stemmer.

## 1 Introduction

Diachronic corpora for the Spanish language have become available for various kinds of research. Two widely known corpora are the RAE's *Corpus diacrónico del español*, CORDE (`http://www.rae.es/`), and Mark Davies' *Corpus del español* (`http://www.corpusdelespanol.org/`). Recently, a first version of the *Corpus histórico del español de México*, CHEM (`http://www.iling.unam.mx/chem/`), became available to the public for the study of the Spanish used in Mexico from the arrival of Europeans to the 19th century.

Many tools for the exploitation and analysis of corpora require a lemmatization process, which is often reduced to simple stemming or graphical word truncation to eliminate inflections. Simple techniques such as the Porter algorithm [1] are regularly applied to corpora of many languages, but they require knowledge of their morphology. Fortunately, in comparison with other languages, Spanish morphology has changed relatively little during the last five centuries. So, a Porter stemmer for today's Spanish could presumably be applied to those centuries in order to accomplish inflection removal. However, given that techniques exist which can be used for stemming without having to code morphological knowledge into the algorithm, it is worthwhile to compare them to the Porter method in order to appreciate what scheme would be better for the CHEM.

In this short paper, the stemming strategy devised for this corpus is described and contrasted with an implementation of the Porter stemmer.[1] The strategy

[1] Various implementations of the Porter algorithm for Spanish are available (based on `http://snowball.tartarus.org/`). In this experiment a version for contemporary Spanish developed at GIL-IINGEN-UNAM, was used.