An Environment for Learner Corpus Research and Error Analysis: The Study of Determiner Errors in Basque

Larraitz Uria, Department of Computer Languages and Systems, University of the Basque Country, Donostia-San Sebastián, Spain

Montse Maritzalar, Department of Computer Languages and Systems, University of the Basque Country, Donostia-San Sebastián, Spain

Igone Zabala, Department of Computer Languages and Systems, University of the Basque Country, Donostia-San Sebastián, Spain

ABSTRACT

This article presents an environment developed for Learner Corpus Research and Error Analysis which makes it possible to deal with language errors from different points of view and with several aims. In the field of Intelligent Computer Assisted Language Learning (ICALL), our objective is to gain a better understanding of the language learning process. In the field of Natural Language Processing (NLP), we work on the development of applications that will help both language learners and teachers in their learning/teaching processes. Using this environment, several studies and experiments on error analysis have been carried out, and thanks to an in-depth study on determiner-related errors in Basque, some contributions in the above mentioned fields of research have been made.

Keywords: Automatic Exercise Generation, Determiner errors analysis, Intelligent Computer Assisted Language Learning (ICALL), Natural Language Processing (NLP)

INTRODUCTION

It is currently assumed that languages are learnt by making errors. In fact, errors are considered good examples of the language learning process since they indicate the steps that learners are taking in that process (Corder, 1967). Therefore, error analysis is a significant starting point in the research field of Intelligent Computer Assisted Language Learning (ICALL). Analyzing language learners’ errors demonstrates, to a certain extent, language competence and the way learners are learning the language. It is also possible to identify many of the difficulties
they may find when learning a new language. Additionally, taking into account learners’ errors and their needs, computer tools and resources can be developed within the field of Natural Language Processing (NLP).

Learner corpora provide an essential basis for the identification of frequently occurring mistakes in learner language. But they can reveal not only what is in a learner corpus, but also what is not in it. That is why learners’ writings offer invaluable information for Second Language Acquisition (SLA) research. In the last years, interesting research has been conducted and contributions made in the area of learner corpus research and language learners’ error analysis in many languages and with different aims (Garnier et al., 2003; Granger, 2007, 2004; Izumi & Isahara, 2004; Tono, 2003; among many others).

The IXA research group1 at the University of the Basque Country is working since 1988 on NLP with the aim of developing tools and resources for the analysis and processing of the Basque language. For this purpose, it is indispensable to analyze the language at all linguistic levels, including lexicon, morphology, syntax, semantics and pragmatics. Language errors must be also analyzed since error analysis can bring contributions in different areas such as ICALL and NLP.

Hence, this article presents an environment which is part of a research methodology (Aldabe et al., 2005; Aldabe et al., 2006a) that allows for language errors to be treated from different perspectives and with several goals. In the field of NLP, our goal is mainly to develop tools and resources for automatic error treatment (Ansa et al., 2004; Uria et al., 2009). In the field of ICALL, our aims are to analyse language learners’ errors and texts to understand their language learning process (Uria et al., 2011) as well as to develop assistant tools for teachers and learners, tools such as an automatic exercise generator (Aldabe et al., 2006b) and an essay scoring system (Castro-Castro et al., 2008).

The article focuses on a study of determiner-related errors in Basque.

**METHODOLOGY**

As regards error analysis methodology, language errors are usually collected from learner corpora, tagged and classified. Then, analysing those texts and the tagged errors, it is possible to make some conclusions on the learners’ interlanguage (IL), the developing language of second language (2L) learners. Interlanguage often differs from the target language (L2), and the annotation of such corpora is an important means of accessing its unique characteristics (Granger, 2003a).

However, manual annotation is a difficult and costly job. That is why many computer tools for language learner error analysis have been developed (Dagneaux, Denness, & Granger, 1998; Meunier, 1998) which enable the error tagging to be performed more accurately, quickly and easily. Apart from annotating errors for language learning and teaching purposes, our idea when developing the environment here presented was to take the advantage of those annotated corpora to carry out the automatic error detection, diagnosis and correction of the tagged error examples.

This environment (Figure 1) consists of several resources and tools including learner corpora, an error classification, the ETIKERRO editor, the Learners and Errors databases and the IRAKAZI and ERREUS applications (Figure 1). The information that is collected, annotated, stored and managed by means of those resources makes it possible the development of new tools in the ICALL and NLP fields.

The main tools and resources that constitute the environment are presented next:

**Language Learner Corpora**

Language learner corpora are collections of authentic texts produced by foreign/second language learners and stored in electronic format. Naturalistic learner productions are an important empirical resource for SLA research. It is a common misconception that corpus building means collecting lots of texts from the Internet and pasting them all together (Cobb, 2003).
A Review of CALL and L2 Reading: Glossing for Comprehension and Acquisition
www.igi-global.com/article/review-call-reading/61346?camid=4v1a