1 Introduction

Xuxen is a spelling checker/corrector for Basque which is going to be commercialized next year. The checker recognizes a word-form if a correct morphological breakdown is allowed. The morphological analysis is based on two-level morphology.

The correction method distinguishes between orthographic errors and typographical errors.

- Typographical errors (or misstypings) are uncognitive errors which do not follow linguistic criteria.
- Orthographic errors are cognitive errors which occur when the writer does not know or has forgotten the correct spelling for a word. They are more persistent because of their cognitive nature, they leave worse impression and, finally, its treatment is an interesting application for language standardization purposes.

2 Correction Method in Xuxen

The main problems found in designing the checking/correction strategy were:

- Due to the high level of inflection of Basque, it is impossible to store every word-form in a dictionary; therefore, the mainstream checking/correction methods were not suitable.
- Because of the recent standardization and widespread dialectal use of Basque, orthographic errors are more likely and therefore their treatment becomes critical.
- The word-forms which are generated without linguistic knowledge must be fed into the spelling checker to check whether they are correct or not.

In order to face these issues the strategy used is basically the following (see also Figure 1).

Handling orthographic errors

The treatment of orthographic errors is based on the parallel use of a two-level subsystem designed to detect misspellings previously typified. This subsystem has two main components:

- Additional two-level rules describing the most likely changes that are produced in the orthographic errors. Twenty five new rules have been defined to cover the most common orthographic errors. For instance, the rule $h:0 \Rightarrow v:v \rightarrow v$ describes that between vowels the $h$ of the lexical level may disappear in the surface. In this way bæhr, typical misspelling of bæhar (to need), will be detected and corrected.

- Additional morphemes linked to the corresponding correct ones. They describe particular errors, mainly dialectal forms. Thus, using the new entry tikan, dialectal form of the ablative singular, the system is able to detect and correct word-forms as etxetik, kaletikan,... (variants of etxetik (from the house), kaletik (from the street),...)

Handling typographical errors

The treatment of typographical errors is quite conventional and performs the following steps:

- Generating proposals to typographical errors using Damerau’s classification.
- Trigram analysis. Proposals with trigrams below a certain probability threshold are discarded, while the rest are classified in order of trigramic probability.
- Spelling checking of proposals.

To speed up this treatment the following techniques have been used:

- If during the original morphological checking of the misspelled word a correct morpheme has been found, the criteria of Damerau are applied only to the unrecognized part. Moreover, on entering the proposals into the checker, the analysis starts from the state it was at the end of the last recognized morpheme.
- The number of proposals is also limited by filtering the words containing very low frequency trigrams.