Designing spelling correctors for inflected languages using lexical transducers

I. Aduriz**, I. Aldezabal, I. Alegria, O. Ansa, J. M. Arriola, N. Ezeiza, Alexander Da Costa*
University of the Basque Country
**Hizkia*
***UZEI*

1 Introduction

This abstract describes the components used in the design of the commercial *XuxenII* spelling checker/corrector for Basque. It is a new version of the Xuxen spelling corrector (Aduriz et al., 97) which uses lexical transducers to improve the process. A very important new feature is the use of user dictionaries whose entries can recognise both the original and inflected forms.

In languages with a high level of inflection such as Basque spelling checking cannot be resolved without adequate treatment of words from a morphological standpoint. In addition to this, the morphological treatment has other important features: coverage, reusability of tools, orthogonality and security.

The tool is based in lexical transducers and is built using the *fst* library of *Inxight*[i]. A lexical transducer (Karttunen 94) is a finite-state automaton that maps inflected surface forms to lexical forms, and can be seen as an evolution of two-level morphology (Koskenniemi, 83) where the use of diacritics and homographs can be avoided and the intersection and composition of transducers is possible. In addition, the process is very fast and the transducer for the whole morphological description can be compacted in less than 1Mbyte.

The design of the spelling corrector consists of four main modules:

- the standard checker,
- the recogniser using user-lexicons,
- the corrector of linguistic variants —proposals for dialectal uses and competence errors—
- the corrector of typographical errors

An important feature is its homogeneity. The different steps are based on lexical transducers, far from ad-hoc solutions.

2 The Spelling Checker

The spelling checker accepts as correct any word which allows a correct standard morphological breakdown. When a word is not recognised by the checker, it is assumed to be a misspelling and a warning is given to the user who has different options, being one of most interesting including its lemma in the user-lexicon.

2.1 The user lexicons

The user-lexicon is offered in order to increase the coverage and to manage specific terminology. Our tool recognises all the possible inflections of a root. The use of a lexical transducer for this purpose is difficult because it is necessary to compile the new entries with the affixes and the rules to update it but this process is slow. The mechanism we have implemented has the following two main components in order to be able to treatment declensions:

1) a general transducer which use standard rules but totally opened lexicon. The result of the analysis is not only if the word is known or not, but also all the possible lemmas

[i] Inxight Software, Inc., a Xerox New Enterprise Company (`www.inxight.com`)

corresponding to this word-form and the grammatical category of each one. The resulting lexical transducer is very compact and fast.

2) a searcher of these hypothetical lemmas in the user-lexicons. If one of them is found, the checker will accept the word, otherwise it will suppose that it has to be corrected.

For this process the system has an interface to update the user lexicon because the part of speech of the lemmas is necessary when they are added to the user lexicon.

3 The Spelling Corrector

Although there is a wide bibliography about the problem of correction (Kukich, 92), it is significative that almost all of them do not mention the relation with morphology and assume that there is a whole dictionary of words or that the system works without lexical information. Oflazer and Guzey (1994) face the problem of correcting words in agglutinative languages.

3.1 Correcting Competence Errors

The need of managing competence errors —also named orthographic errors— has been mentioned and reasoned by different authors (van Berkel & de Smedt, 88). When we faced the problem of correcting misspelled words the main problem found was that because of the recent standardisation and the widespread dialectal use of Basque, competence errors or linguistic variants are more likely and therefore their treatment becomes critical.

When we decided to use lexical transducers for the treatment of linguistic variants, the following procedure was applied to build the transducer:

1) Additional morphemes are linked to the standard ones using the possibility of expressing two levels in the lexicon.
2) Definition of additional rules for competence errors that do not need to be integrated with the standard ones. It is possible and clearer to put these rules in other plane near to the surface and compose them with the standard rules, because most of the additional rules are due to phonetic changes.

When a word-form is not accepted the word is checked against this second transducer. If the incorrect form is recognised now —i.e. it contains a competence error— the correct lexical level form is directly obtained and, as the transducers are bi-directional, the corrected surface form will be generated from the lexical form using only standard transducer.

For example, the word-form `beartzetikan`, misspelling of `behartzetik` (from the need) can be corrected although the edit-distance is three. The process of correction is the following:

- Decomposition into three morphemes: *behar* (using a rule to guess the h), *tze* and *tikan*.
- *tikan* is a non-standard use of *tik* and as they are linked in the lexicon is chosen.
- The standard generation of *behar+tze+tik* obtains the correct word *behartzetik*.

3.2 Handling Typographical Errors

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

- Generating proposals to typographical errors using Damerau's classification <u>(edit distance of one). These proposals are ranked in order of trigramic probability.</u>
- Spelling checking of proposals.

3.3 Results

The results are very good in the case of competence errors and not so good for typographical errors because in the last case only errors with an edit-distance of one have been planned. <u>In 89% of the cases the right proposal is generated and in 71% is the first</u>. It would be possible to generate and test all the possible words with an edit-distance higher, but the number of proposal would be very high. The corrector has been integrated in several tools. A demonstration can be seen in http://ixa.si.ehu.es.

Acknowledgements

This work has had partial support from the Culture Department of the Government of the Basque Country. We would like to thank to Xerox for letting us using their tools, and also to Lauri Karttunen for his help.

References

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. (1997) A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford. 1997.

Alegria I., Artola X., Sarasola K (1997). Improving a Robust Morphological Analyser using Lexical Transducers. Recent Advances in Natural Language Processing. *Current Issues in Linguistic Theory (CILT) series*. John Benjamins publisher company. Vol. 136. pp 97-110. 1997.

Karttunen L. (1992). *Constructing Lexical Transducers*, Proc. of COLING´94, 406-411. 1994.

Koskenniemi, K. (1983). *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics. Publications nº 11, 1983.

Kukich K. (1992). Techniques for automatically correcting word in text. *ACM Computing Surveys*, vol.24, No. 4, 377-439

Oflazer K, Guzey C. (1994). Spelling Correction in Aglutinative Languages, *Proc. of ANLP-94*, Sttutgart.

Van Barkel B, De Smedt K. (1988). Triphone analysis: a combined method for the correction of orthographic and typographical errors. *Procedings of the Second Conference ANLP* (ACL), pp.77-83.