A spelling corrector for Basque based on morphology

Itziar Aduriz, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, Kepa Sarasola and Miriam Urkia
Basque Country University and UZEI

1 Introduction

This paper describes the components used in the elaboration of the commercial *Xuxen* spelling checker/corrector for Basque. Because Basque is a highly inflected and agglutinative language, the spelling checker/corrector has been conceived as a by-product of a general purpose morphological analyser/generator (Alegria et al., 96). The two-level model of morphology (Koskenniemi, 83) that we use is based on two main components —see Sproat (1992):

- A lexicon where the morphemes (lemmas and affixes) and the possible links among them (morphotactics) are defined.
- A set of rules which controls the mapping between the lexical level and the surface level due to the morphonological transformations (morphophonemics). There are four kind of rules: context restriction rules “=>“ (lexical character may be realized as the lexical one in the given context), surface coercion rules “<=“ (lexical character must be realized as the lexical one in the given context), composite rules “<=>“ (lexical character must be realized as the lexical one in the given context and this change is licit only in this context) and exclusion rules (lexical character may not be realized as the lexical one in the given context). The rules are independent from the morphotactics. The rules are compiled into transducers, so it is possible to apply the system for both analysis and generation.

In order to increase the coverage and the robustness, the analyser has been designed in an incremental way and it consists of three main modules: the standard analyser, the analyser of linguistic variants —due to dialectal uses and competence errors—, and the analyser without lexicon which can recognize word-forms without having their lemmas in the lexicon. An important feature of the analyser is its homogeneity as the three different steps are based on two-level morphology, very different from ad-hoc solutions.

This analyser is a basic tool for current and future work on automatic processing of Basque and its first applications is the commercial spelling corrector named *Xuxen* that is presented here. First we describe the subsystem added to the analyser in order to increase relevantly the coverage in competence errors

2 The Analysis of Linguistic Variants

As we said in (Alegria et al., 96) because of the recent standardisation and the widespread dialectal use of Basque, the standard morphology is not enough to offer good results when analysing corpora.

Three types of linguistic variants are distinguished: morpheme variants —i.e. *haundi* is used instead of standard *handi* (big)—, morphotactical variants —i.e. the standard declension of *batzu* (someone) is plural but it is often declined as indeterminate— and morphonological variants or regular non-standard changes —i.e. the use of the *h* was controversial and it is not yet well known.

The treatment of these variants has been carried out by means of an additional two-level subsystem (Aduriz et al., 93), thus increasing the coverage of the morphological processor. This tool is the main component in the correction of competence errors in Xuxen.

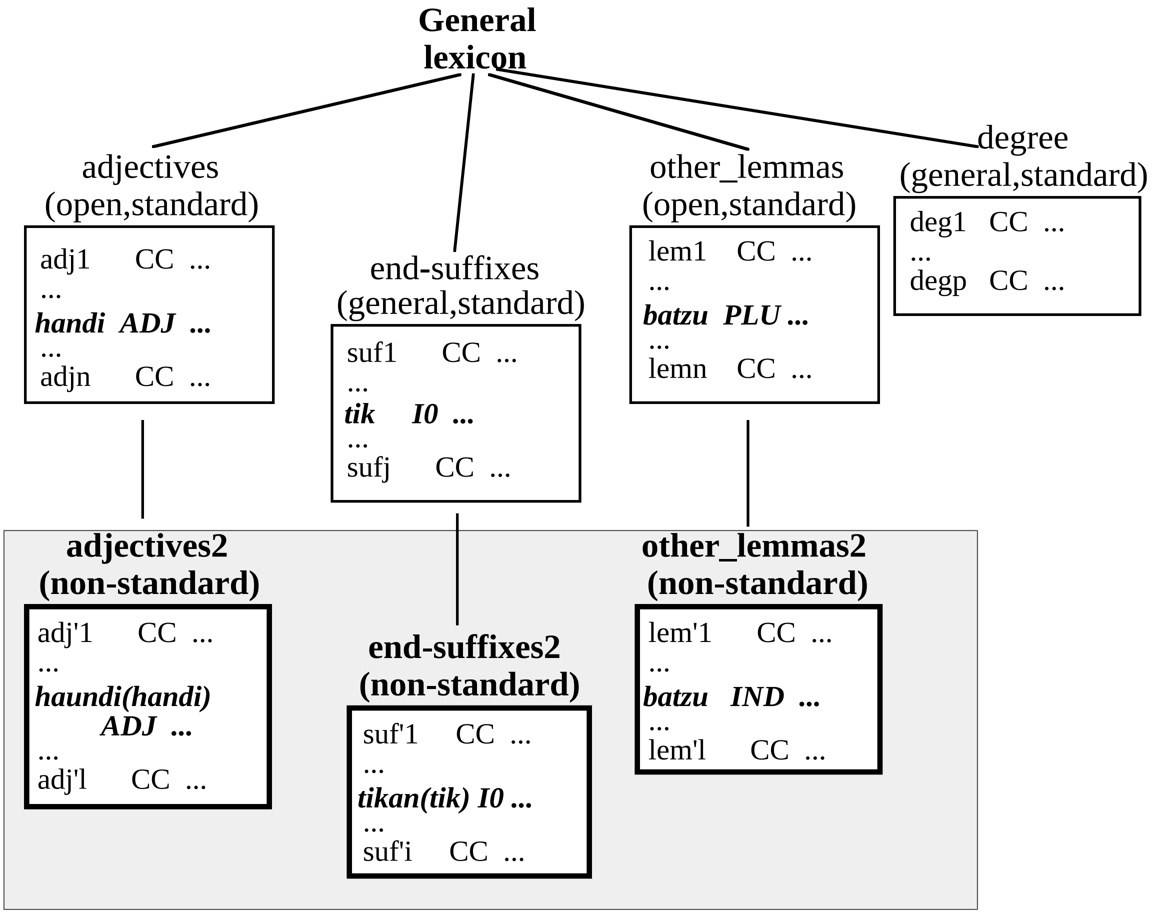

Fig. 1 Standard and non-standard morphemes in the lexicon

This subsystem is also used in the spelling corrector to manage competence errors and has two main components:

1) New morphemes linked to the corresponding correct ones (Fig. 1). They are added to the lexical system and they describe particular variations, mainly dialectal forms. Thus, the new entry tikan, dialectal form of the ablative singular, linked to its corresponding right entry tik will make the system be able to analyse and correct word-forms such as etxetikan, kaletikan,... (variations of etxetik (from the house), kaletik (from the street), ...). Morphotactical variations can be analysed changing the continuation class (*CC*) of morphemes (see *batzu* in Fig. 6). More than 1000 non-standard morphemes —mainly dictionary entries— have been included in this subsystem.

2) New two-level rules describing the most likely regular morphonological changes that are produced in the variations. These rules have the same structure and management than the original ones. Eighteen new rules have been defined (see appendix 1) to cover the most common competence errors.

   For instance, the next rule describes that between vowels or at the beginning of a word before a vowel the h of the lexical level may disappear in the surface level and vice versa. In this way the word-form bear, misspelling of behar (to need), can be analysed.

```
"description: losing and generating h"
h:0 => [ Beg | Vowel ] _ Vowel ;
                   ! behar:bear
                   ! hau:au
0:h => [ Beg | Vowel ] _ Vowel ;
                   ! ziur:zihur
                   ! esparru:hesparru
```

All these rules are optional (context restriction rules: =>) and have to be compiled with the standard rules but some inconsistencies have to be solved because some of the changes described in the new subsystem were forbidden in the original rule-set.

It is possible to correct the morpheme and morphonological variations using standard morphemes linked to variants and entering them into the morphological generation with standard rules. This has proved very interesting when applied to spelling correction.

In our system it is also possible to identify the kind of variant that has been analysed. As we can see below the result of the analysis tells us whether the analysis is standard or not —in this case the analysis is marked as *VAR*— and gives us the standard morphemes as well as the variant —*Etik* (standard) and *Etikan* (variant)— and the rules applied when non-standard morphonological rules are used —the change from *zuhaitz* (standard) to *suaitx* (non-standard) is analysed using the 2th rule (changes among sibilants) two times and the 6th rule (losing of h) once. This is being used in ICALL —Intelligent Computer Aided Language Learning— applications for Basque (Maritxalar & Diaz, 93).

```
((form "kaletikan")
  ((anal VAR1)
   ((lemma "kale")((POS NOUN))))
   ((morph "0")((POS DEC)(NUM S)(DET DEF))))
   ((morph "Etik") (var3 "Etikan")((POS DEC)(CAS ABL)))))
)

((form "suaitxetikan")
  ((anal VAR1)
         ((lemma "zuhaitz")(var "suaitx")((POS NOUN))(R2,R6,R2))
   ((morph "0")((POS DEC)(NUM S)(DET DEF))))
   ((morph "Etik") (var3 "Etikan")((POS DEC)(CAS ABL)))))
)
```

The non-standard analyses are rejected if there are standard ones. When different non-standard analyses are obtained there is a disambiguation process that prefers concrete analysis (morpheme or morphotactical variants) to general ones (morphonological variants) and, among these analyses, those with less non-standard morphonological rules are applied.

3 The Spelling Checker/Corrector

Xuxen is a spelling checker/corrector for Basque based on two-level morphology (Agirre et al. 92) which was comercialized in 1994. Languages with a high level of inflection such as Basque make it impossible to store every word-form in a dictionary even in a very compressed way; so, 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 —if the lemma is in the lexicon all the declension is known— and security.

The spelling checker accepts as good any word which allows a correct standard morphological breakdown, while the objective of the morphological analyser is to obtain all of the possible breakdowns and the corresponding information. In order to speed the process buffers with the most frequent words, the most frequent misspellings and the previous word that appeared in the text are used (Peterson, 80). The user-lexicon explained in section 4.1 is offered to the users in order to increase the coverage and to manage specific terminology.

When a word is not known by the checker, it is assumed to be a misspelling and a warning is given to the user who has different options, two of most interesting being entering its entry in the user lexicon, and asking for possible corrections.

Although there is a wide bibliography about the problem of correction —the compilation of Kukich (1992) is very interesting — 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. Only Oflazer and Guzey (1994) face the problem of correcting words in agglutinative languages, but their proposal, although interesting, is computationally too complex if very fast analysers (lexical transducers for example) are not used.

When we faced the problem of correcting misspelled words, the main problems found in designing the correction strategy were:

- As has been said due to the high level of inflection of Basque, it is impossible to store every word-form in a dictionary, even in a compressed way (Agirre et al. 92).
- 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.
- The word-forms which are generated without linguistic knowledge must be fed into the spelling checker to check whether they are valid or not.

Having in mind the points above we have designed a strategy based on two steps, which are complementary and that can be carried out in parallel: treatment of competence errors and treatment of typographical errors.

3.1 Correcting Competence Errors

The need of managing competence errors —also named orthographic errors— has been mentioned and reasoned by different authors.

> “... Most of the correction methods currently in use in spelling checkers are biased toward the correction of typographical errors. We argue that this is not the right thing to do. Even if orthographical errors are not as frequent as typographical errors, they are not to be neglected for a number of good reasons. First, orthographical errors are *cognitive* errors, so they are more persistent than typographical errors: proof-reading by the author himself will often fail to lead to correction. Second, orthographical errors leave a worse impression on the reader than typographical errors. Third, the use of orthographical correction for standardization purposes (e.g. consistent use of either British or American spelling) is an important application appreciated by editors. ...” (van Berkel & de Smedt, 88:77).

Our treatment of competence errors is based on the parallel use of a two-level subsystem designed to analyse non-standard uses and competence errors previously typified, which is added to the two-level system used by the checker.

As we have shown in section 3.3, this subsystem has two main components:

- New two-level rules describing the most likely changes that are produced in the orthographic errors.
- New morphemes linked to the corresponding correct ones. They are added to the lexical system and they describe particular errors, mainly dialectal forms.

When a word-form is not accepted by the checker the competence error subsystem is added and the system retries the morphological checking. If the incorrect form can be recognized now —i.e. it contains a competence error— the correct lexical level form is directly obtained and, as the two-level system is bi-directional, the corrected surface form will be generated from the lexical form using only standard two-level rules.

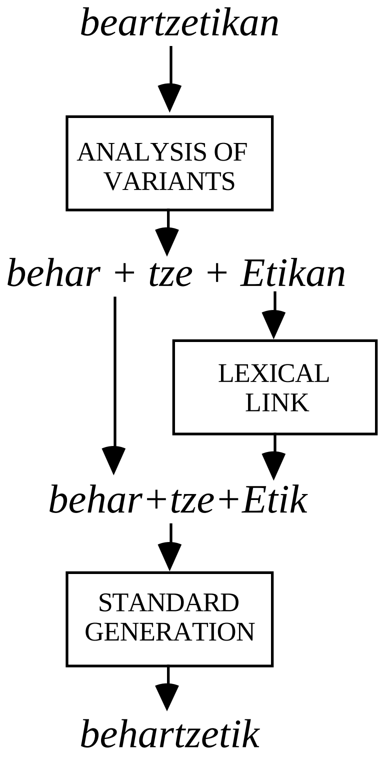

Fig. 2 Correction process of misspellings

For example, as is shown in Fig. 2, the word-form `beartzetikan,` misspelling of `behartzetik` (from the need) can be corrected although the edit-distance (Damerau, 64) is three. The complete process of the correction process would be the following:

- The word is analysed and decomposed into three morphemes: *behar* (to need) using a non-standard rule to guess the "h", *tze* (nominalization) and *Etikan* (non-standard ablative singular).
- *Etikan* is a non-standard use of *Etik* and they are linked in the lexicon, so the last one is chosen.
- The standard generation is performed to obtain the correct word.

Examining the results reported in section 3.3 more than the 80% of the competence errors can be corrected with the proposed subsystem.

3.2 Handling Typographical Errors

The treatment of typographical errors is quite conventional and performs the following steps (Fig. 2):

- Generating proposals to typographical errors using Damerau's classification.
- Trigram analysis. It is performed during the generation of the proposals: proposals with trigrams below a certain probability threshold are discarded, while the rest are classified in order of trigramic probability.
- Spelling checking of proposals. On the basis of the previous criteria only, incorrect word-forms could be offered to the user. Therefore, these word-forms must be fed into the spelling checker to check whether they are valid or not.

The whole process would be especially slow, mostly due to the checking of alternatives. To speed it up the following techniques have been used:

- All the proposals are looked up before in the buffer of frequent words, and only the proposals of word-forms that have not been detected as an orthographic error, will be morphologically verified.

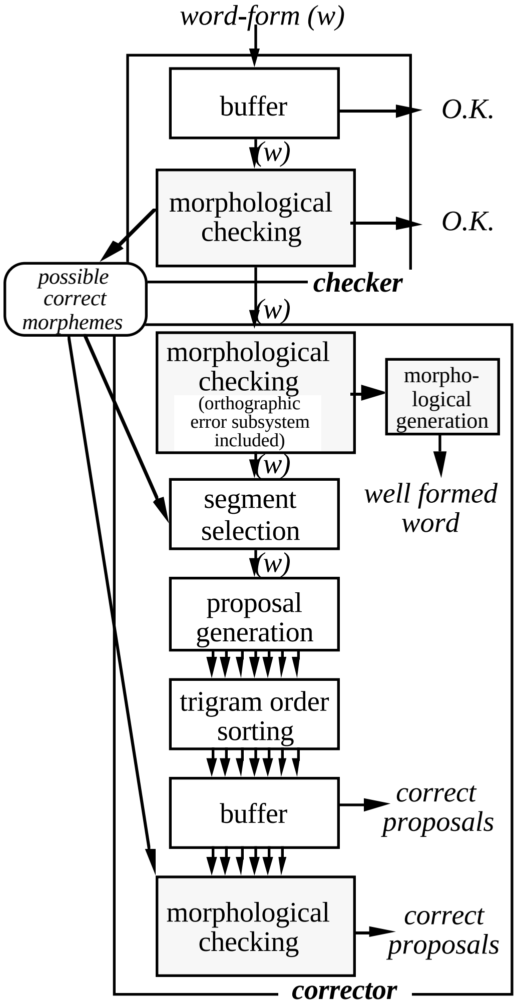

Fig. 2 Architecture of the checker/corrector

- 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, decreasing the number of alternatives. This criterion is applied on the basis that far fewer "typos" are committed at the beginning of a word (Yannakoudakis, 83). 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 to be checked is also limited by filtering the words containing very low frequency trigrams, and never exceeds a maximum number of forms. At any rate, after having obtained three correct proposals, the process will end.

3.3 Results

The results are very good in the case of competence errors —they could be even better by improving the non-standard lexicon— and not so good for typographical errors. In the last case only errors with an edit-distance of one can be corrected due to the techniques used to speed the system.

The results are explained in Table 1. Three different sets of 100 misspellings —coming from different kinds of text— are studied, showing the percentage of right correction obtained at the first proposal *(1)*, among the first three ones *(3)* and among all the proposals *(n)*; and, finally, the time to obtain all the proposals. The first column correspond to the explained method and the second to the same method without limiting the number of morphological checks. So, it seems that the chosen speed method is a good trade-off between speed and precision.

Texts		**RESULT**	**UNLIM.**
Text A (students) 100 misspellings	(n)	**%82**	%89
	(3)	**%81**	%86
	(1)	**%74**	%75
	time(s/w)	**0,3**	15
Text B (technical report) 100 misspellings	(n)	**%63**	%88
	(3)	**%62**	%86
	(1)	**%49**	%68
	time(s/w)	**0,4**	12,5
Text C (newspaper) 100 misspellings	(n)	**%70**	%89
	(3)	**%68**	%85
	(1)	**%59**	%71
	time(s/w)	**0,35**	16,7
TOTAL 300 misspellings	(n)	**%72**	%89
	(3)	**%70**	%86
	(1)	**%61**	%71
	time(s/w)	**0,35**	14,7

Table 1 Precision of the corrector

Without changing the main idea of the correction method, the precision can be improved slowing it (assuming the speed of morphological checking is constant). For example it would be possible, but very slow with our analyser, to generate and test all the possible words with an edit-distance higher than one from the original misspelling. Another way could be investigating in the line proposed by Oflazer and Guzey (1994); based on flexible morphological decomposition, although by the moment we have found the same problems of response time.

Conclusions

The spelling checker/corrector named Xuxen is based on the two-level morphological processor. The correction strategy for misspelled words in the spelling checker/corrector has been described. It deals with both competence and typographical errors and, in the first case, a new correction strategy has been used. An additional two-level subsystem enables recognizing dialectal variants and regular non-standard changes. The results have been described in detail to explain the quality, scale and precision of this tool.

Acknowledgements

This work has had partial support from the Economy Department of the Local Government of Gipuzkoa and from the Education Department of the Government of the Basque Country. We are in debt with Hitesh Patel for his help in the English version of this paper.

References

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M. (1993). A Morphological Analysis Based Method for Spelling Correction. *Proc. of the 6th Conference of the EACL*, p.463.

Agirre E., Alegria I., Arregi X., Artola X., Diaz de Illarraza A,. Maritxalar M., Sarasola K., Urkia M. (1992). XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology, *Proc.of the Third ANLP*, 119-125.

Alegria I. (1995). *Euskal morfologiaren tratamendu automatikorako tresnak*. Ph.D. Thesis. In Basque.

Alegria I., Artola X., Sarasola K., Urkia M. (1996). Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, vol.XX, No. X, XXX.

Damerau F. (1964). A technique for computer detection and correction of spelling errors. *Comm. of ACM* vol. 7 pp. 171-176.

Euskaltzaindia (1985). *Euskal Gramatika: Lehen urratsak (I, II, III eta IV)*. Euskaltzaindia, Bilbo.

Kaplan R. M. and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguistics*, vol.20(3), 331-380.

Karttunen L. and Beesley K.R. (1992). *Two-Level Rule Compiler*. Xerox ISTL-NLTT-1992-2.

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.

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

Maritxalar M., Diaz de Ilarraza A. (1993). *Integration of Natural Language Techniques in the ICALL System Field: The treatment of incorrect knowlegment*. Barne-txostena. EHU/LSI/TR 993.

Mitton R. (1987). Spelling checker, spelling correctors and the misspellings of poor spellers. *Information Processing and Management,* Vol 23, N.5, pp.495-505.

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

Peterson J.L. (1980). Computer Programs for detecting and correcting spelling errors, *Comm. of ACM*, vol.23, No.12.

Solack A, Oflazer K. (1993). Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing*, vol.8, No. 3, 113-130.

Sproat R. (1992). *Morphology and Computation*. The MIT Press.

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.

Yannakoudakis E.J. (1983). The rules of spelling errors. *Information Processing & Management* vol.19 no.2.