XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K.
Informatika Fakultatea
P.K. 649
20080 DONOSTIA (Basque Country - Spain)
xabier@si.ehu.es

Urkia M.

U.Z.E.I.
Aldapeta, 20
20009 DONOSTIA (Basque Country)

Abstract

The application of the formalism of two-level morphology to Basque and its use in the elaboration of the XUXEN spelling checker/corrector are described. This application is intended to cover a large part of the language.

Because Basque is a highly inflected language, the approach of spelling checking and correction has been conceived as a by-product of a general purpose morphological analyzer/generator. This analyzer is taken as a basic tool for current and future work on automatic processing of Basque.

An extension for continuation class specifications in order to deal with long-distance dependencies is proposed. This extension consists basically of two features added to the standard formalism which allow the lexicon builder to make explicit the interdependencies of morphemes.

User-lexicons can be interactively enriched with new entries enabling the checker from then on to recognize all the possible flexions derived from them.

Due to a late process of standardization of the language, writers don't always know the standard form to be used and commit errors. The treatment of these "typical errors" is made in a specific way by means of describing them using the two-level lexicon system. In this sense, XUXEN is intended as a useful tool for standardization purposes of present day written Basque.

1 Introduction

This paper describes the application of two-level morphology to Basque, along with its use in the elaboration of the XUXEN spelling checker/corrector. The morphological analyzer included in XUXEN has been designed with the aim of laying the foundations for further development of automatic processing of Basque. The fact that Basque is a highly inflected language makes the correction of spelling errors extremely difficult because collecting all the possible word-forms in a lexicon is an endless task.

The simplicity of English inflections made for reduced interest in research on morphological analysis by computer. In English, the most common practice is to use a lexicon of all of the inflected forms or a minimum set of morphological rules (Winograd, 83). That means that while a great many language independent tools have been developed for syntactic and semantic analysis, the same cannot be said for morphological tools. In 1981, Kaplan and Kay (Kaplan *et al.*, 81) made a valuable contribution in designing a formalism for phonological generation by means of rules compiled in an automaton. This idea would later be followed up by Koskenniemi (Koskenniemi, 83-85; Karttunen *et al.*, 87) in the two-level formalism. The computational model for two-level morphology has found widespread acceptance in the following years due mostly to its general applicability, declarativeness of rules and clear separation of linguistic knowledge from the program . The essential difference from generative phonology is that there are no intermediate states between lexical and surface representations. Word recognition is reduced to finding valid lexical representations which correspond to a given surface form. Inversely, generation proceeds from a known lexical representation and searches for surface representations corresponding to it. The complexity of the model is studied in depth in (Barton, 85), who with few exceptions agrees with Karttunen (Karttunen, 83) in feeling that the complexity of a language has no significant effects on the speed of analysis or synthesis.

There have been many implementations of the two-level model for very different languages, some of them taking a full coverage of the language: Finnish, English and Arabic among others. Our implementation is intended to cope extensively with present day Basque.

XUXEN manages user-lexicons which can be interactively enriched during correction by means of a specially designed human-machine dialogue which allows the system to acquire the internal features of each new entry (sublexicon, continuation class, and selection marks).

Moreover, XUXEN deals with errors often due to recent standardization of Basque. An additional lexicon includes alternative variants to the standard entries and additional rules

model erroneous morphophonological changes; this allows a specialized treatment of "typical errors".

Following are given an overview of Basque morphology and the application of the two-level model to Basque, then the lexical database built as a support for this and other applications is described, and finally, the strategies followed in the design and implementation of the spelling checker-corrector.

2 Brief Description of Basque Morphology

Basque is an agglutinative language; that is, for the formation of words the dictionary entry independently takes each of the elements necessary for the different functions (syntactic case included). More specifically, the affixes corresponding to the determinant, number and declension case are taken in this order and independently of each other (deep morphological structure).

One of the principal characteristics of Basque is its declension system with numerous cases, which differentiates it from the languages from surrounding countries. The inflections of determination, number and case appear only after the last element in the noun phrase. This last element may be the noun, but also typically an adjective or a determiner. For example:

etxe zaharreAN	(*etxe zaharrean*: in the old house)
etxe: noun	(house)
zahar: adjective	(old)
r and *e*: epenthetical elements	
A: determinate, singular	
N: inessive case	

So, these inflectional elements are not repeated in each individual word of a noun phrase as in the Romance languages.

Basque declension is unique; that is, there exists a single declension table for all flexionable entries, compared to Latin for instance,which has 5 declension paradigms.

As prepositional functions are realized by case suffixes inside word-forms, Basque presents a relatively high power to generate inflected word-forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. Moreover, while 77 of them are simple combinations of number, determination, and case marks, not capable of further inflection, the other 58 are word-forms ended with one of the two possible genitives or with a sequence composed of a case mark and a genitive mark. If the latter is the case, then by adding again the same set of morpheme combinations (135) to each one of those 58 forms a new, complete set of forms could be recursively generated. This kind of construction reveals a noun ellipsis inside a complex noun phrase and could be theoretically extended *ad infinitum*; in practice, it is not usual to find more than two levels of this kind of recursion in a word-form but, in turn, some quite frequent forms contain even three or more levels. This means that a morphological analyzer for Basque should be able to recognize the amount of 77 + 58 (77 + 58 (77 + 58)) = 458683 inflected forms for each noun taking into account only these two levels of recursion.

e.g.	*semeA*	(the son)
	semeArI	(to the son)
	semeArEN	(of the son)
	semeArEN etxeA	(the house of the son)
	semeArENA	(the one (house) of the son)
	semeArENArI	(to the one (house) of the son)

This generation capability is similar for all parts of speech. In the case of adjectives, due to the possibility of graduation, this capability is 4 times greater.

The grammatical gender does not exist in Basque; there are not masculine and feminine. However, the verb system uses the difference sometimes, depending on the receiver and the grade of familiarity: this is the case of the allocutive verb forms.

Verb forms are composed of a main verb and an auxiliary finite form. The verb system in Basque is a rich one: it is often found in a single finite verb form morphemes corresponding to ergative, nominative and dative cases.

Derivation and composition are quite productive and they are widely used in neologism formation.

3 Application of Two-Level Morphology to Basque

3.1 The Rules

The correlations existing between the lexical level and the surface level due to morphophonological transformations are expressed by means of the rules. In the case of Basque 21 two-level rules have been defined. These rules are due to the four following reasons: eminently phonological (7 rules), morphological (3 rules), orthographical (5 rules), and both phonological and morphological (6 rules). The effects of the rules are always phonological. Given that suppletion cases are rare in Basque, phonemically unrelated allomorphs of the same morpheme are included in the lexicon system as separated entries. No rules deal with these phenomena. The rules are applied to express three types of realizations: adding or removing a character, or alternation of a character from the lexical to the surface level. These basic transformations can be combined.

In order to control the application of the rules 17 selection marks are used. Since two-level rules are sensitive only to the form of the word, these marks inform on part of speech, special endings and other features needed for handling exceptions in rules.

Examples of rules:

2nd rule: ADDITION OF EPENTHETICAL e.

2:e ⇔ [C:C / 8: / 6:r / 4:] _

8th rule: VOICING OF t.

t:d ⇔ [l&2:l / n&2:n / n2:n] _

where C represents any consonant, 2 is the selection mark stated at the beginning of affixes requiring epenthetical e, 8 is the selection mark stated at the end of those lemmas with final au diphthong, 6 is the selection mark stated at the end of those lemmas with final hard r, 4 is the selection mark

stated at the end of verb infinitives with final n, and & is the selection mark stated at the end of place names with final l or n which forces voicing of following t.

The first rule states that the selection mark 2 is realized as surface e, always and only when it is preceded either by a consonant or a selection mark 8, or a selection mark 6 realized as surface r, or a selection mark 4.

The second rule specifies the voicing of lexical t, always and only when it is preceded either by a n or l followed by the selection marks & and 2, or a n followed by the selection mark 2.

At the moment, the translation of rules into automata required by the two-level formalism is made by hand.

3.2 The Lexicon System

Among the morphological phenomena handled by our system so far, we would like to emphasize the following: whole declension system —including place and person names, special declension of pronouns, adverbs, etc.—, graduation of adjectives, relational endings and prefixes for verb forms —finite and non-finite— and some frequent and productive cases of derivation and compounding.

The lexicon system is divided into sublexicons. Lexical representation is defined by associating each entry to its sublexicon and giving it the corresponding continuation class.

a) Sublexicons: Lemmas, auxiliaries of verbs and finite verb forms, and different affixes corresponding to declension, determination, number, verb endings, and so on are distinguished.
All of the entries in the sublexicons are coded with their continuation class and morphological information. At present near 15,000 items are completely coded in the lexicon system: 8,697 lemmas, 5,439 verb forms and 120 affixes. They are grouped into 94 different sublexicons. Within short time, this number will be increased in order to code all the 50,000 entries present at the moment in the database supporting the lexicon. The entry code gives, when appropriate, information on part of speech, determination, number, declension case, gender (exceptional cases), relation (of subordination), part of speech transformation that a relational affix produces, type of verb, root of finite verb forms, tense-mood, grammatical person, etc. along with the specific information each entry requires.

b) Continuation class: Generalizations are not always possible. For example, while with nouns and adjectives the assignment of a single continuation class to all of the elements of each category has been possible, adverbs, pronouns and verbs have required more particularized solutions. A number of 79 continuation classes have been defined.

The system permits the unlimited accumulation and treatment of information as it extracts data from the dictionary according to the segmentation found. This feature is essential to Basque given that: a) a large amount of morpho-syntactic knowledge can be derived from a single word-form, and b) there is no set theoretical limit to the potential recursion of genitives.

Separated representation for homographs and homonyms —in the main sublexicon, with the same or different continuation classes— has been made possible. Although this distinction is not necessarily relevant to morphological analysis, future work on syntax and semantics has been taken into consideration.

3.3 Some Problems and Possible Solutions

Although until now, the notation and concept of continuation class have been used, in authors' opinion it is the weakest point of the formalism. Specially in dealing with the Basque auxiliary verb, cases of long-distance dependencies that are not possible to express adequately have been found. Different solutions have been proposed to solve similar problems for other languages (Trost, 90; Schiller, 90). The solution suggested below is not as elegant and concise as a word-grammar but it seems expressive enough and even more efficient when dealing with this kind of problems. To this end, an improved continuation class mechanism is being implemented. This mechanism supports the following two extra features:

- **bans** that can be stated altogether with a continuation class; they are used to express the set of continuation classes forbidden further along the word-form (from the lexical entry defined with this restricted continuation class).
Examples:

  bait (PERTSONA - LA - N)

  this states that among the morphemes in the word-form following to the verb prefix *bait* are to be allowed those belonging to the continuation class PERTSONA but also that further on in the word no morphemes belonging to the continuation classes LA or N will be accepted.
  continuation class-tree: the lexicon builder has the possibility of restricting the set of allowed continuation morphemes for a given one, by means of making explicit these morphemes through different segments in the word-form; this explicitation is done by giving a parenthesized expression representing a tree. This mechanism improves the expressiveness of the formalism providing it with the additional power of specifying constraints to the set of morphemes allowed after the lexicon entry, stating in fact a continuation "path" —not restricted to the immediate morpheme— which makes explicit that set in a conditioned way.
  Examples:

joan	*NA TZAI O*	(I went to him)
joan	*NA TZAI T **	(I went to me*)
etorri	*HA TZAI T*	(You came to me)
etorri	*HA TZAI N **	(You came to you* (fem.))

  Long-distance dependency cases are found in the verb finite form instances above: the initial morpheme *na*

(nominative, first person) allows dative morphemes corresponding to the third person after the morpheme *tzai* (root) but not those corresponding to the first person. Analogously the theoretically possible *hatzain** is not grammatical in Basque because it combines two second person morphemes in nominative and dative cases. The continuation corresponding to *na* can be stated as follows:

na (KI (DAT23 (N_KE)), TZAI (DAT23 (LAT)))

which specifies two alternative continuation "paths" allowed after this morpheme: the one including the morphemes in the continuation class KI and that which includes those in the continuation class TZAI. In both cases DAT23 restricts the set of morphemes potentially permitted as continuation of those in KI or TZAI, allowing only the 2nd and 3rd person dative morphemes. Without this extension of the formalism, it would be possible to do it by storing repeatedly the morpheme *tzai* in two or more different lexicons, but this is not very useful when the distance between dependent morphemes is longer. Similarly:

ha (KI (DAT13 (N_KE)), TZAI (DAT13 (LAT)))

is the way to express that *ha* (nominative, 2nd person) is to be combined with dative morphemes of 1st and 3rd person but not with those of 2nd.
Continuation classes N_KE and LAT further restrict the morphemes allowed conditioning them in this case to the classes KI and TZAI respectively. Note that in this example two different cases of long-distance dependency are present.

4 The Lexical Database

The lexical database is supported permanently in a relational system. This database is intended as an independent linguistic tool. Within this framework, information about the two-level lexicon system is stored in three different relations.

Each lexicon is mainly characterized by the susceptibility of its components to be the initial morpheme in a word-form and by whether or not they are of semantic significance.

In another relation, continuation classes are defined in terms of lexicons or other continuation classes. It is possible to store examples as well.

Finally, the main component of the database is the set of lexicons with their associate entries: the two-level form of the entry is stored along with its original form, the source from which it has been obtained, examples, and in some cases (lemmas) the usage frequency. Obviously, the linguistic knowledge related to the entry is also stored in this relation.

A user friendly interface allows the lexicon builder to do the operations of addition and updating of entries, consistency checking, etc. in a comfortable way. Selection marks depending on knowledge contained in the database such as part of speech, subcategorization of nouns, special endings for certain categories, etc. may be automatically derived from the information in the base.

The production of the up-to-date run-time lexicon and continuation class definitions in the format required by the two-level system is obtained automatically from this database by means of specially designed procedures.

5 The Spelling Checker/Corrector

The morphological analyzer-generator is an indispensable basic tool for future work in the field of automatic processing of Basque, but in addition, it is the underlying basis of the spelling checker/corrector. The spelling checker accepts as good any word which permits a correct morphological breakdown, while the mission of the morphological analyzer is to obtain all of the possible breakdowns and the corresponding information. Languages with a high level of inflection such as Basque make impossible the storage of 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.

From the user's point of view XUXEN is a valid system to analyze documents elaborated by any word processor. It operates at a usual speed and takes up reasonable amount of space, thus allowing it to work with any microcomputer.

5.1 The Spelling Checker

The basic idea of accepting words which have a correct morphological analysis is fulfilled with classic techniques and tools for detecting spelling errors (Peterson, 80). A filter program appropriate for the punctuation problems, capital letters, numbers, control characters and so on has been implemented. In addition to the mentioned problems, difficulties intrinsic to Basque, like word-composition, abbreviations, declension of foreign words, etc. have been also taken into account. Besides this filter, interactive dialogue with the user, buffers for the most frequent words (in order to improve the performance of the system), and maintenance of the user's own dictionary (following the structure of the two-level lexicon) are the essential elements to be added to the morphological analyzer for the creation of a flexible and efficient spelling checker.

It is very important to notice the necessity of a suitable interface for lexical knowledge acquisition when it comes to managing with precision the inclusion of new lemmas in the user´s own dictionary. Without this interface, morphological and morphotactical information essential to the checker would be left unknown and, so, no flexions could be accepted. Currently, the system acquires information from the user about part of speech, subcategorization for nouns —person or place names, mainly— and some morphonological features like final hard-or-soft r distinction. So, the user, giving to the system several answers, makes possible the correct assignment of continuation class and selection marks to the new lemma. In this way, open class entries may be accepted and adequately treated. Entries belonging to other classes may also be entered but no flexions of them will be recognized. This ability of the checker to deal correctly with new lemmas

requires, in turn, certain grammatical knowledge from the user.

Our prototype, running on a SUN 3/280 and using a buffer containing 4,096 of the most frequent word-forms, checks an average of 17.1 words per second in a text with a rate of misspellings and unknown words (not present in the current lexicon) of 12.7%. Considering the word-forms the system deems as erroneous, statistical tests have shown that 60% are actual misspellings, 16% would have been recognized had the general lexicon been more comprehensive, and the rest correspond to specific words (technical terms, proper nouns, etc.) which the user should include in his own dictionary.

Within a short time minor changes will provide greater performance. A PC version is also in use.

5.2 The Spelling Corrector

When a word is not recognized by the spelling checker, the user can choose, among other options, to ask the system for suggestions for replacing the erroneous word. These suggestions, logically, must be correct words which will be similar to the word-form given by the user.

To find similar words to propose, there exists two working lines:

1) Using as a guide the "sources of error" described by Peterson (Peterson, 80), errors are basically of two types:

- Errors due to lack of knowledge of the language: these errors are often not dealt with on the assertion that they are infrequent, but Pollock and Zamora (Pollock, 84) evaluate their frequency at between 10% and 15%. Moreover, because Basque is a language whose standardization for written use has begun only in recent years, a higher degree of error would be expected for it.
- Typographical errors. According to the classic typification by Damerau (Damerau, 64) 80% of "typos" are one of the following four types: one exceeding character, one missing character, a mistaken character, or the transposition of two consecutive characters.
  Following that, $n+26(n-1)+26n+(n-1)$ possible combinations (n being the length of a word) can be generated; they must be examined to verify their validity and the most probable must be selected. For this examination it is normal to use statistical methods which, though not very reliable, are highly efficient (Pollock, 84).

2) Definition of a measurement of distance between words and calculation of which words of the dictionary give a lesser distance with respect to the erroneous word (Angell, 83; Tanaka, 87). The most frequently used measure is the "distance of Levenshtein".

This second method, measurement of distance, is slower but much more reliable than the first one, though it is not suitable for a lexicon system where the words are incomplete, as is the case. Due chiefly to this, the chosen option has been the adaptation of the first method, taking into account the following criteria:

- **Handling of typical errors.** A linguistic study has been carried out on typical errors, that is, errors most frequently committed due to lack of knowledge of the language itself or its latest standardization rules, or due to the use of dialectal forms. To store typical errors a parallel two-level lexicon subsystem is used. In this subsystem, each unit is an erroneous morpheme which is directly linked to the corresponding correct one. When searching for words the two-level mechanism is used together with this additional lexicon subsystem. When a word-form is not accepted by the checker the typical errors subsystem is added and the system retries the orthographical checking. If the incorrect form is now correctly analyzed —so, it contains a typical error— the correct morpheme corresponding to the erroneous one is directly obtained from the typical errors subsystem. There will also be additional two-level rules, which will reflect the erroneous, but typical morphonological alternations in dialectal utilizations or training periods.
- **Generating alternatives.** Generating alternatives to typographical errors using Damerau's classification.
- **Trigram analysis.** In generating the alternatives, trigram analysis is used both for discarding some of them as well as for classifying them in order of probability.
- **Spelling checking of proposals.** On the basis of the three previous criteria, incorrect word-forms would be offered to the user. Therefore, the word-forms must be fed into the spelling checker to check whether they are valid or not.

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

- If during the analysis of the word considered wrong a correct morpheme has been found, the criteria of Damerau are applied only in the part unrecognized morphologically, so that the number of possibilities will be considerably lower. 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 continues from the state it was in at the end of that last recognized morpheme.
- On doing trigrammatical analysis a trigram table mechanism is used, by means of which generated proposals will be composed only of correct trigrams and classified by their order of probability; thus, correction analysis (the slowest element of the process) is not carried out with erroneous trigrams and the remaining analyses will be in the order of trigrammatical probability. Besides that, the number of proposals to be checked is also limited by filtering the words containing very low frequency trigrams, and never exceeds 20 forms. At any rate, after having obtained three correct proposals, the generation process will end.

- If a word is detected as a typical error, it will not be verified as a possible "typo". This requires the analysis of typical errors to take place previous to that of "typos", in spite of being less probable. The justification is that we are particularly interested in giving preferential treatment to typical errors and, what's more, these can be handled more speedily.

The average time for the generation of proposals for a misspelt word-form, on the SUN machine cited above, is 1.5 s. The best case is when three or more alternatives are found in the buffer of most frequent words, and takes less than 0.1 s. The worst case, when no correct proposals are found for a long word-form and when no correct initial morphemes were recognized during its analysis, takes around 6 s.

6 Conclusions

The XUXEN analyzer/checker/corrector has been described as based on the two-level morphological formalism. It deals with Basque, a highly inflected language recently standardized. At the moment a prototype of the system has been implemented in C language. This implementation is a general tool for Basque useful for texts written by any word processing programme.

As is well known, in the two-level model morphemes are stored in the sublexicons without alterations, unlike in other systems. From a linguistic standpoint, the clarity and respect for the lexical unit promoted by this way of focusing morphological analysis is of great importance. However, long-distance dependencies between morphemes can not be adequately expressed by means of the continuation class mechanism. An improved continuation-class mechanism to solve this problem is suggested.

At present, the lexicon system contains nearly 15,000 items, now the coding of new lemmas in order to reach 50,000 entries is being completed. At this moment finite verb forms (approximately 2,000) are in the lexicon, although they could be seen as analyzable forms. These verb forms have been described by means of their component morphemes taking into account the long-distance dependency problems they present. This have been done using the extension of the continuation-class formalism described in 3.3 which is currently being implemented.

With the lemmas and morphemes coded so far, XUXEN is able to recognize approximately three millions different word-forms without at all counting forms produced by genitive recursion. Considering that most of lemmas in the lexicon can take genitive suffixes, our present implementation of the spelling checker would recognize thousands of millions of word-forms.

User-lexicons can be interactively enriched with new entries enabling XUXEN to recognize from then on all the possible flexions derived from them.

An additional two-level lexicon subsystem is used in our system to store the so-called typical errors. Typical errors are due often to the recent standardization of the language and dialectal uses. This lexicon subsystem is used preferably when suggesting alternatives to the user.

Acknowledgements

Prof. Koskenniemi for his fruitful comments on an earlier version of this paper.

References

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Urkia M. Aplicación de la morfología de dos niveles al euskara. S.E.P.L.N, vol. 8, 87-102. 1989.

Angell R., Freund G., Willety P. Automatic Spelling Correcting using a trigram similarity measure. Information Processing & Management, vol 19, nº4, 1983.

Barton, E. *Computational Complexity in two-level Morphology*, 1985.

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

Euskaltzaindia. *Aditz laguntzaile batua*. Euskaltzaindia, Bilbo 1973.

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

Kaplan, R. M., and M. Kay. *Phonological rules and finite-state transducers*. Paper read at the annual meeting of the Linguistic Society of America in New York City, 1981.

Karttunen, L. *KIMMO : A two-level Morphological Analyzer*. Texas Linguistic Forum, Vol 22, Pp.165-186, 1983.

Karttunen L., Koskenniemi K., Kaplan R. *A Compiler for Two-Level Phonological Rules* in "Tools for Morphological Analysis", Center for the Study of Language and Information, Report No. CLSI-87-108.

Kay, M. *Morphological Analysis*.. A.Zampolli & N. Calzolari eds. (1980). Proc. of the Int. Conference on Computational Linguistics (Pisa), 1973.

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

Koskenniemi, K. *Compilation of Automata from Morphological Two-level Rules*. Pp. 143-149. Publication nº 15. University of Helsinki, 1985.

Peterson J.L. Computer Programs for detecting and correcting spelling errors. Comm. of ACM vol.23 nº12 1980.

Pollock J., Zamora A. Automatic spelling correction in scientific and scholarly text. Comm. of ACM vol.27 358-368, 1984.

Ritchie, G.D., S.G. Pulman, A.W.Black and G.J. Russell. A Computational Framework for Lexical Description. Computational Linguistics, vol. 13, numbers 3-4, 1987.

Sarasola, I. *Gaurko euskara idatziaren maiztasun-hiztegia.* (3gn. liburukia), GAK, Donostia, 1982.

Schiller A. Steffens P. A lexicon for a German two-level morphology. Paper read at Euralex 1990 (Benalmádena).

Tanaka E., Kojima Y. A High Speed String Correction method using a hierarchical file. IEEE transactions on pattern analysis and Machine Intelligence vol.9 nº6, 1987.

Trost, H. The application of two-level morphology to non-concatenative German morphology. COLING-90, Helsinki, vol.2 371-376.

Winograd, T. *Language as a cognitive process.* Vol.1: Syntax, pp 544-549. Addison-Wesley, 1983.

Yannakoudakis E.J. The rules of spelling errors. Information Processing & Management vol.19 nº2, 1983.