A TRADE-OFF BETWEEN ROBUSTNESS AND OVERGENERATION IN MORPHOLOGY

Iñaki Alegria, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, Kepa Sarasola

Informatika Fakultatea, Euskal Herriko Unibertsitatea
P.K. 649. 20080 DONOSTIA Basque Country
acpalloi@si.ehu.es

Abstract

This paper describes a multilevel method in lexical-morphological analysis which offers robustness and avoids overgeneration. The proposed levels are three: the analysis of standard forms, the analysis of linguistic variants, and the analysis without lexicon. This method can be seen as a variant of constraint relaxation used in syntax. This solution is being used in two different applications: a spelling corrector and a general purpose lemmatizer/tagger.

Keywords: Morphology, corpus analysis, POS tagging, lemmatization, robustness.

1 Introduction

Although some years ago the simplicity of English inflection reduced the interest in research on morphological analysis by computer, in the last years the importance of the morphological treatment of languages has had a great increase due to two important facts: robustness and multilingualism. It is clear that morphological analysis is a first task when doing NLP, specially in the case of highly inflected languages.

Moreover the appearance of new formalisms more flexible and powerful have made easier the morphological description of different languages. Among these formalisms the two-level morphology (Koskenniemi 83) has became very popular because of their good attributes: bidirection (useful for analysis and generation), speed and clear separation between linguistic information and program, and among the three main elements (lexicon, morphotactics and phonological changes). Different flavours of two-level morphology have been developed (Ritchie *et al.* 92) (Karttunen 94) (Carter 95). PC-Kimmo (Antworth 90) is a freely available software tool which is useful to experiment with this formalism (e-mail: *pc-parse-owner@sil.org*).

In addition to its role as preprocess for syntax and semantics, the applications of automatic morphological treatment are multiple:

- spelling checking/correction and OCR verification
- lemmatization: very useful in lexicography and information retrieval
- interfaces in natural language
- computer aided learning

2 The problem

In the design of a morphological processor it is necessary to keep in mind two criteria that often tend to be quite incompatible: to avoid overgeneration and to get robustness. While robustness is basic in corpus analysis and natural language understanding, avoiding overgeneration is very important in spelling checking and language generation. In addition to this, in robust systems overgeneration increases the ambiguity and many times this ambiguity is not real. So, for these applications a trade-off between robustness and overgeneration is necessary.

An example of the difficult compatibility between these features is derivation. Derivation is not regular and is very productive generating new lexical terms. Although a fine-grained description can give us a good approach, in general it is necessary to decide in favour of generalization that contributes to robustness or in favour of precision to avoid overgeneration. Another example, as we will examine below, is the treatment of non-standard uses of the language.

The compromise between overgeneration and robustness must be handled in the three main elements of the morphological description: lexicon, morphotactics and phonological changes (Sproat 92). The choice of two-level morphology in the proposed method is due to the attributes referenced in the introduction, but the proposed method can be extended to other kinds of morphological description.

3 Our proposal

We propose a multilevel method in lexical-morphological analysis which combines both features —robustness and avoiding overgeneration— in order to build a general purpose morphological analyzer/generator. The analyzer for Basque built applying this method has been used in the development of a commercial spelling corrector (Agirre et al. 92) and also to design a lemmatizer/tagger (Aduriz et al. 95). In this way the method guarantees reusability and modularity, because, as we said above, they are two applications with different requirements of robustness and overgeneration.

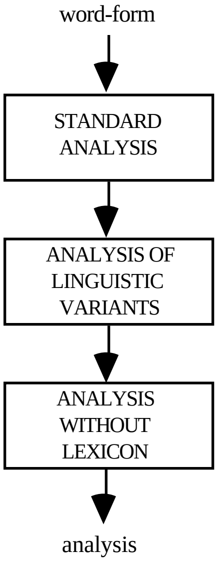

Figure 1: Modules of the analyzer

The morphological analyzer that we propose is designed in an incremental way. It is composed of three main modules (see Figure 1): the analysis of standard forms, the analysis of linguistic variants —due to dialectal uses and competence errors—, and the analysis without lexicon, which is able to recognize word-forms without having their lemmas in the lexicon. An important feature of this method is its homogeneity —the three different steps can be based on two-level morphology and use the same program modules— far from ad-hoc solutions.

3.1 The standard analysis

This module is able to analize standard language word-forms. Different models or morphology can be used to describe the main components —lexicon, morphotactics and morphonology— but those based in two-level morphology are becoming the most successful for real applications (Antworth 90, Ritchie et al. 92, Oflazer 94). In our applications for Basque we defined —using a database— about 60,000 entries in the lexicon, more than 130 patterns of morphotactics and 24 morphonological rules.

Concept	total
Unknown words.	392 (%100)
A.-Non-standard use	129 (%32,9)
B1.-Loan-words	33 (%8,4)
B2.-Out of lexicon	84 (%21,4)
B3.-New derivatives	46 (%11,7)
B4.-Foreign words	53 (%13,5)
C.-Errors	40 (%10,2)
D.-Others	7 (%1,8)

Table 1: Causes of the faults

When standard morphology and a closed lexicon are used although overgeneration is almost avoided the coverage is not satisfactory enough. After testing a corpus of Basque the coverage was about the 95%. We tried to find the causes of the faults sorting them into different sets (see Table 1)

Keeping in mind these figures, it seems necessary to manage non-standard uses and forms whose lemmas are not in the lexicon if we wanted to develop a comprehensive analyser, so three different modules are proposed: management of user lexicon, analysis of linguistic variants and analysis without lexicon.

In addition to the standard description a user lexicon and an interface for updating is added in order to be able to analyze text-dependent vocabulary. The user lexicon is combined with the general one increasing the coverage of the morphological analyzer, but does not produce overgeneration. As the standard description avoids overgeneration, this whole description will be used for generation and spelling checking.

3.2 The analysis of linguistic variants

Because of the non-standard or dialectal uses of the language and the competence errors, the standard morphology is not enough to offer good results when analyzing real text corpora. This problem becomes critical in languages like Basque where standardisation is still in process and dialectal forms are still of widespread use.

An additional morphological subsystem which analyzes —and generates— linguistic variants is added in order to increase the robustness of the morphological processor. This subsystem has three main components:

1) New morphemes linked to their corresponding standard ones. They are added to the lexical system and describe particular variations, mainly dialectal forms. Thus, in our application for Basque, the new entry tikan, dialectal form of the ablative singular morpheme, linked to its corresponding standard entry tik, will be able to analyze and correct word-forms such etxetikan, kaletikan,... —variants of etxetik (*from the house*), kaletik (*from the street*), ...
2) Changing the morphotactical information —continuation class in the original two-level morphology— corresponding to some morphemes, morphotactical errors can be analyzed. For example, in Basque the base-form batzu (*some*) must be declined in plural but it is quite usual to use the non-determinate declension paradigm; so, if this kind of inflection is assigned to the base-form in the new subsystem, non-standard inflections can also be analized.
3) New rules describing the most likely regular changes that are produced in the linguistic variants. These rules have the same structure and management as the standard ones. All these rules are optional and have to be combined with the standard rules. Some inconsis-tencies have to be solved because some new changes were forbidden in the original rules. For instance, the rule **h:0 => V:V_V:V** describes that between vowels the h may disappear. In this way the word-form bear, misspelling of behar (*to need*), can be analyzed.

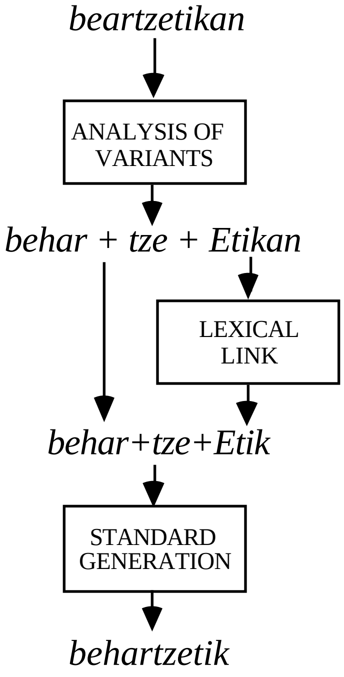

Figure 2: Correction process of variations

Forms containing a combination of different kinds of errors —i.e. beartzetikan, variant of behartzetik (*from the need*)— can be analyzed and corrected too (see Figure 2). This is very useful to apply in correction and also adds robustness to the system without mixing standard and non-standard languages.

When more than one analysis is obtained in this step, a local disambiguation process is carried out in order to select the “most standard” possibilities. To be able to disambiguate, our analyzer records the non-standard entries and rules applied in the analysis. The disambiguator computes the number of them, selecting the analysis with less non-standard features, giving priority —when the number is the same— to the lexical changes (particular ones) over the others (general ones).

3.3 The analysis of unknown words (guesser)

The problem of unknown words does not disappear with the previous modules. In order to deal with it a two-level mechanism for analysis without lexicon is added, thus increasing the robustness of the analyzer. This module is based on the idea used in speech synthesis (Black et al. 91). There are other proposals but most of them —for instance (Chanod & Tapanainen 95)— are oriented to tagging and do not obtain full analysis but only tags.

This mechanism has the following three main components in order to be capable of treating unknown words:

1) generic lemmas represented by "*??*" —one for each possible open category or subcategory— which are organized with their affixes in a small two-level lexicon.
2) new morphotactical information in order to generalize non-standard morphotactics —i.e. derivation.
3) two additional rules in order to express the relationship between the generic lemmas and any acceptable lemma of Basque, which are combined with the standard rules. Some standard rules have to be modified because surface and lexical level are specified, and in this kind of analysis the lexical level of the lemmas changes.

The obtaining of at least one analysis is guaranteed but the ambiguity rate is very high. In order to decrease the great number of ambiguous analysis, a local disambiguation process based on statistical data is carried out based on the next criteria:

- it will remain at least one analysis by each category part of speech (POS). We suppose that a later treatment which manages the context —a tagger for example— will improve the final disambiguation.
- the core of this step is to decide among different hypothetical lemmas. A decision is taken in function of the length of the lemma —shorter lemmas use to be more adequate because more affixes have been found in the analysis— and statistics linking POS and endings of lemmas.

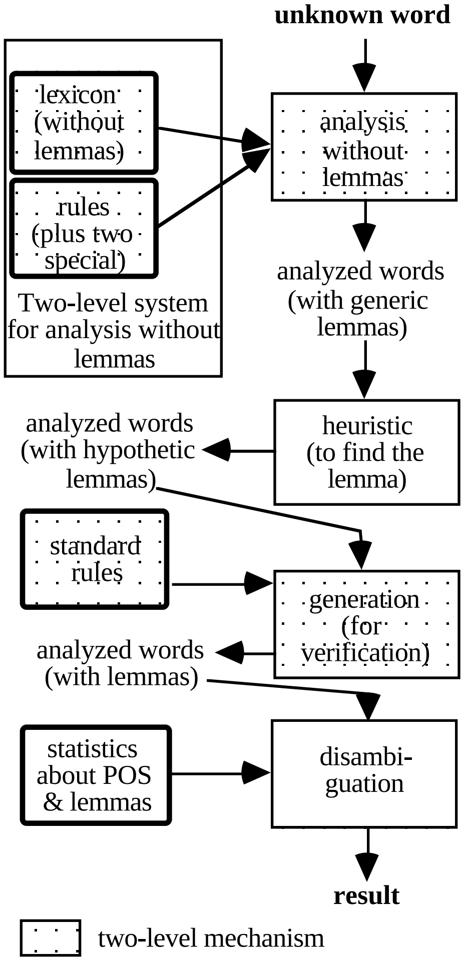
unknown word
lexicon (without lemmas)
rules (plus two special)
Two-level system for analysis without lemmas
analysis without lemmas
analyzed words (with generic lemmas)
heuristic (to find the lemma)
analyzed words (with hypothetic lemmas)
standard rules
generation (for verification)
analyzed words (with lemmas)
statistics about POS & lemmas
disambi-guation
result
two-level mechanism

Figure 3: Analysis of unknown words

The result of this process for Basque is that over the initial ambiguity of 11.4 analysis per word the remaining ambiguity is 5.8 —2.9 in standard forms— with a precision —the right analysis of the unknown word remains— higher than 92%. The whole process of analysis of unknown words can be seen in Figure 3.

4 Results

The analysis process is performed in an incremental way, so that, if standard analysis are obtained subsequent steps are suspended, and if analysis as linguistic variant is reached the treatment of unknown words is not done. Thus, overgeneration is avoided in a great number of cases but robustness is guaranteed.

A commercial spelling corrector and also a lemmatizer/tagger for Basque have been built using this method. In our application a precision rate of 99.4% has been obtained (see Table 2). The main reason of the errors are misspellings in close categories —categories where unknown words are not expected—.

Concept	Total
Different words (list)	4.036
Unknown words in standard analysis	392 %10
Linguistic variants Recognized variants	129 107 (%83)
Errors after all analyses **Precision**	25 **%99,4**

Table 2: Precision of the analyzer

The speed of the analysis depends on the implementation of the formalism. Recently new proposals made the two level formalism a very fast mechanism, able to analyze thousands of words per second (Karttunen 94). We evaluated this proposal (Alegria et al. 95) obtaining very good results.

The explained method can be seen as a variant of constraint relaxation techniques used in syntax (Stede, 92), where the first constraint demands standard language, the second one standard plus studied variants, and the third one allows free language.

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

References:

(Aduriz et al. 95) Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. *Different issues in the design of a lemmatizer/tagger for Basque*. “From text to tag” Workshop (SIGDAT,EACL), 18-23. 1995.

(Agirre et al. 92) Agirre E., Alegria I., Arregi X., Artola X., Diaz de Illarraza A,. Maritxalar M., Sarasola K., Urkia M. XUXEN: *A spelling checker/corrector for Basque based on Two-Level morphology*, Proc. of the 3. ANLP, 119-125. 1992.

(Alegria 95) Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak*. Ph.D. Thesis. Univ. of the Basque Country. 1995.

(Alegria et al. 95) Alegria I., Artola X., Sarasola K. (1995). *Improving a robust morphological analyser using lexical transducers*. Recent advances in Natural Language Processing. Bulgary.

(Antworth 90) Antworth E.L. *PC-KIMMO: A two-level processor for morphological analysis*. Occasional Publications in Academic Computing, No. 16, Dallas, Texas. 1990.

(Black et al. 91) Black A., van de Plassche,J., Williams B. *Analysis of Unkown Words through Morphological Descomposition*. Proc. of 5th Conference of the EACL, vol. 1, 101-106.1991.

(Carter 95) Carter D. Rapid development of morphological descriptions for full language processing system. *Proc. of EACL´95*.

(Chanod & Tapanainen 95) Chanod J.P., Tapanainen P. *Creating a tagset, lexicon and guesser for a French tagger*. “From text to tag” Workshop (SIGDAT,EACL), 18-23. 1995.

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

(Koskenniemi 83) 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.

(Oflazer 94) Oflazer K. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, vol.9, No. 2, 137-148.

(Ritchie et al. 92) Ritchie G., A.W.Black, G.J. Russell and S.G. Pulman. *Computational Morphology*. The MIT Press. 1992.

(Sproat 92) Sproat R. *Morphology and Computation*. The MIT Press. 1992.

(Stede 92) Stede M. *The Search for Robustness in Natural Language Understanding*. Artificial Intelligence Review 6, 383-414. 1992.