CURRENT ISSUES IN
LINGUISTIC THEORY | 136

Recent Advances in Natural Language Processing

Edited by Ruslan Mitkov
and Nicolas Nicolov

JOHN BENJAMINS PUBLISHING COMPANY

Hahn, Udo. 1989. "Making Understanders out of Parsers: Semantically Driven Parsing as a Key Concept for Realistic Text Understanding Applications". *International Journal of Intelligent Systems* 4:3.345-393.

Hahn, Udo, Susanne Schacht & Norbert Bröker. 1994. "Concurrent, Object-oriented Natural Language Parsing: The *ParseTalk* Model". *International Journal of Human-Computer Studies* 41:1/2.179-222.

Hayes, Patrick J. 1985. "The Second Naive Physics Manifesto". *Formal Theories of the Commonsense World* ed. by J. Hobbs & R. Moore, 1-36. Norwood, N.J.: Ablex.

MacGregor, Robert. 1991. "The Evolving Technology of Classification-based Knowledge Representation Systems." *Principles of Semantic Networks* ed. by J. Sowa, 385-400. San Mateo, Calif.: Morgan Kaufmann.

MacGregor, Robert & Raymond Bates. 1987. *The LOOM Knowledge Representation Language.* Information Sciences Institute, University of Southern California (ISI/RS-87-188).

Mars, Nicolaas J. I. 1994. "The Role of Ontologies in Structuring Large Knowledge Bases". *Knowledge Building and Knowledge Sharing* ed. by K. Fuchi & T. Yokoi, 240-248. Tokyo, Ohmsha and Amsterdam: IOS Press.

Palmer, Martha S. et al. 1986. "Recovering Implicit Information". *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL'86)*, 10-19. New York, N.Y.

Rada, Roy, Hafedh Mili, Ellen Bicknell & Maria Blettner. 1989. "Development and Application of a Metric on Semantic Nets". *IEEE Transactions on Systems, Man, and Cybernetics* 19:1.17-30.

Resnik, Philip. 1995. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, vol.I, 448-453. Montreal, Canada.

Rips, L. J., E. J. Shoben & E. E. Smith. 1973. "Semantic Distance and the Verification of Semantic Relations". *Journal of Verbal Learning and Verbal Behavior* 12:1.1-20.

Simmons, Geoff. 1992. "Empirical Methods for 'Ontological Engineering'. Case Study: Objects". *Ontologie und Axiomatik der Wissensbasis von LILOG* ed. by G. Klose, E. Lang & Th. Pirlein, 125-154. Berlin: Springer.

Strube, Michael & Udo Hahn. 1995. "*ParseTalk* about Sentence- and Text-level Anaphora". *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, 237-244.

Wada, Hajime. 1994. "A Treatment of Functional Definite Descriptions." *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, vol.II, 789-795. Kyoto, Japan.

Improving a Robust Morphological Analyser Using Lexical Transducers

Iñaki Alegria, Xabier Artola & Kepa Sarasola

University of the Basque Country

Abstract

This paper describes the components of a robust and wide-coverage morphological analyser for Basque and their transformation into lexical transducers. The analyser is based on the two-level formalism and has been designed in an incremental way with three main modules: the standard analyser, the analyser of linguistic variants, and the analyser without lexicon which can recognise word-forms without having their lemmas in the lexicon. This analyser is a basic tool for current and future work on automatic processing of Basque and its first three applications are a commercial spelling corrector and a general purpose lemmatiser/tagger. The lexical transducers are generated as a result of compiling the lexicon and a cascade of two-level rules (Karttunen et al. 1994). Their main advantages are speed and expressive power. Using lexical transducers for our analyser we have improved both the speed and the description of the different components of the morphological system. Some slight limitations have been found too.

1 Introduction

The two-level model of morphology (Koskenniemi 1983) has become the most popular formalism for highly inflected and agglutinative languages. The two-level system is based on two main components: (i) a lexicon where the morphemes (lemmas and affixes) and the possible links among them (morphotactics) are defined; (ii) a set of rules which controls the mapping between the lexical level and the surface level due to the morphophonological transformations.

The rules are compiled into transducers, so it is possible to apply the system for both analysis and generation. There is a free available software, PC-Kimmo (Antworth 1990) which is a useful tool to experiment with this formalism.

Different flavours of two-level morphology have been developed, most of them changing the continuation-class based morphotactics by unification based mechanisms (Ritchie et al. 1992; Sproat 1992).

We did our own implementation of the two-level model with slights variations, and applied it to Basque (Agirre et al. 1992), a highly inflected and agglutinative language.

In order to deal with a wide variety of linguistic data we built a Lexical Database (LDBB). This database is both source and support for the lexicons needed in several applications, and was designed with the objectives of being neutral in relation to linguistic formalisms, flexible, open and easy to use (Agirre et al. 1995). At present it contains over 60,000 entries, each with its associated linguistic features (category, sub-category, case, number, etc.).

In order to increase the coverage and the robustness, the analyser has been designed in a incremental way. It is composed of three main modules (see Figure 1): the standard analyser, the analyser of linguistic variants produced due to dialectal uses and competence errors, and the analyser without lexicon which can recognise 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, far from ad-hoc solutions.

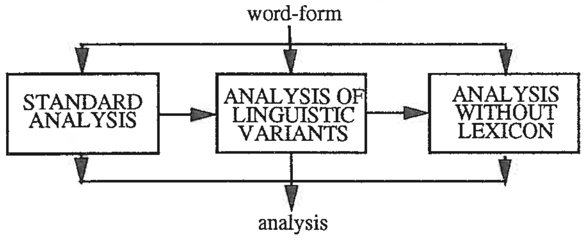

Fig. 1: *Modules of the analyser*

This analyser is a basic tool for current and future work on automatic processing of Basque and its first two applications are a commercial spelling corrector (Aduriz et al. 1994) and a general purpose lemmatiser/tagger (Aduriz et al. 1995).

Following an overview of the lexical transducers and the description of the application of the two-level model and lexical transducers to the different steps of morphological analysis of Basque are given.

2 Lexical transducers

A lexical transducer (Karttunen et al. 1992; Karttunen 1994) is a finite-state automaton that maps inflected surface forms into lexical forms, and can be seen as an evolution of two-level morphology where:

- Morphological categories are represented as part of the lexical form. Thus it is possible to avoid the use of diacritics.
- Inflected forms of the same word are mapped to the same canonical dictionary form. This increases the distance between the lexical and surface forms. For instance *better* is expressed through its canonical form good (*good*+COMP:*better*).
- Intersection and composition of transducers is possible (see Kaplan & Kay 1994). In this way the integration of the lexicon (the lexicon will be another transducer) in the automaton can be resolved and the changes between lexical and surface level can be expressed as a cascade of two-level rule systems (Figure 2).

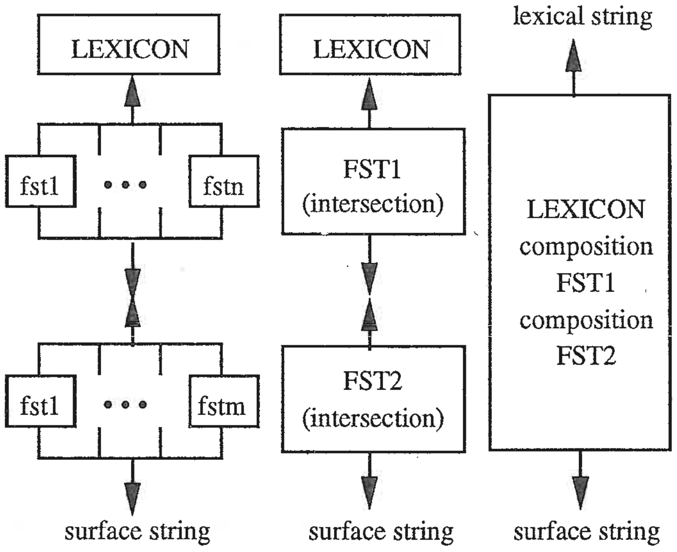

Fig. 2: *Lexical transducers (from Karttunen et al. 1992)*

In addition, the morphological process using lexical transducers is very fast (thousands of words per second) and the transducer for a whole morphological description can be compacted in less than 1 MB.

Different tools to build lexical transducers (Karttunen & Beesley 1992; Karttunen 1993) have been developed in Xerox and we are using them. Uses of lexical transducers are documented by Chanod (1994) and Kwon & Karttunen (1994).

3 The standard analyser

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 spoken in the surrounding countries.

We have applied the two-level model defining the following elements (Agirre et al. 1992; Alegria 1995):

- Lexicon: over 60,000 entries have been defined corresponding to lemmas and affixes, grouped into 154 sublexicons. The representation of the entries is not canonical because 18 diacritics are used to control the application of morphophonological rules.
- Continuation classes: they are groups of sublexicons to control the morphotactics. Each entry of the lexicon has its continuation class and all together define the morphotactics graph. The long distance dependencies among morphemes can not be properly expressed by continuation classes, therefore in our implementation we extended their semantics defining the so-called extended continuation classes.
- Morphophonological rules: 24 two-level rules have been defined to express the morphological, phonological and orthographic changes between the lexical and the surface levels that appear when the morphemes are combined.

The morphological analyser attaches to each input word-form all possible interpretations and its associated information that is given in pairs of morphosyntactic features.

The conversion of our description to a lexical transducer was done in the following steps:

1. Canonical forms and morphological categories were integrated in the lexicon from the lexical data-base.

2. Due to long distance dependencies among morphemes, which could not be resolved in the lexicon, two additional rules were written to ban some combinations of morphemes. These rules can be put in a different rule system near to the lexicon without mixing morphotactics and morphophonology (see Figure 3).
3. The standard rules could be left without changes (mapping in the lexicon canonical forms and arbitrary forms) but were changed in order to change diacritics by morphological features, doing a clearer description of the morphology of the language.

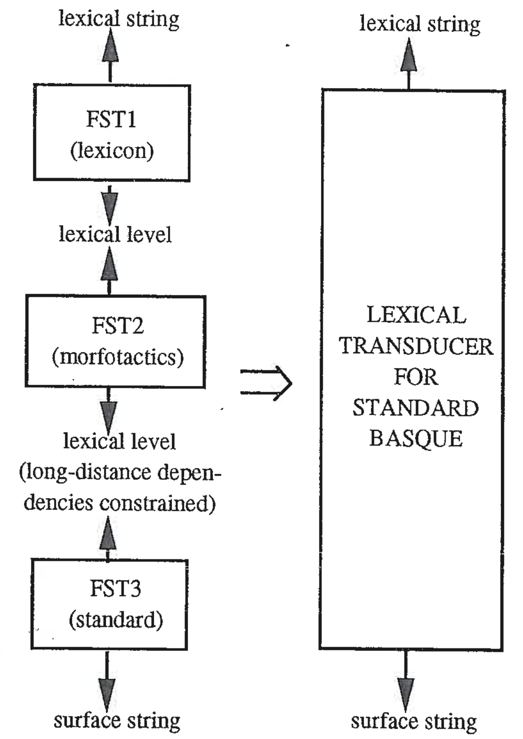

Fig. 3: *Lexical transducer for the standard analysis of Basque*

The resultant lexical transducer is about 500 times faster than the original system.

4 The analysis and correction of linguistic variants

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. To increase the coverage of the morphological processor an additional two-level subsystem was added (Aduriz et al. 1993). 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. 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 morpheme, linked to its corresponding right entry tik will be able to analyse and correct word-forms such `etxetikan`, `kaletikan`, ... (variants of `etxetik` *from the house*, `kaletik` *from the street*, ...). Changing the continuation class of morphemes morphotactic errors can be analysed.
2. New two-level rules describing the most likely regular changes that are produced in the variants. These rules have the same structure and management than the standard ones. Twenty five new rules have been defined to cover the most common competence errors. For instance, the rule `h:0 => V:V_V:V` describes that between vowels the h of the lexical level may disappear in the surface level. In this way the word-form `bear`, misspelling of `behar`, *to need*, can be analysed. All these rules are optional and have to be compiled with the standard rules but some inconsistencies have to be solved because some new changes were forbidden in the original rules.

To correct the word-form the result of the analysis has to be entered into the morphological generation using correct morphemes linked to variants and original rules. To correct `beartzetikan`, variant of `behartzetik`, two steps, analysis and generation, are followed as it is shown in Figure 4.

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

1. The additional morphemes linked to the standard ones are solved using the possibility of expressing two levels in the lexicon. In one level the non-standard morpheme will be specified and in the other (the correspondent to the result of the analysis) the standard morpheme.
2. The additional rules do not need to be integrated with the standard ones (Figure 5), and so, it is not necessary to solve the inconsistencies.

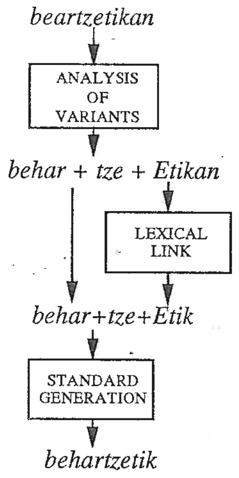

Fig. 4: *Steps for correction*

As Figure 5 (B) shows, it is possible and clearer to put these rules in other plane near to the surface, because most of the additional rules are due to phonetic changes and do not require morphological information. Only the surface characters, the morpheme boundary and additional information about one change (the final a of lemmas) complete the intermediate level between the two rule systems.

3. In our original implementation it was possible to distinguish between standard and non-standard analysis (the additional rules are marked and this information can be obtained as result of the analysis), and so the non- standard information can be additional; but with lexical transducers, it is necessary to store two transducers one for standard analysis and other for standard and non-standard analysis.

Although in the original system the speed of analysis using additional information was two or three times slower than the standard analysis, using lexical transducers the difference between both analysis is very slight.

5 The analysis of unknown words

Based on the idea used in speech synthesis (Black et al. 1991), a two-level mechanism for analysis without lexicon was added to increase the robustness of the analyser.

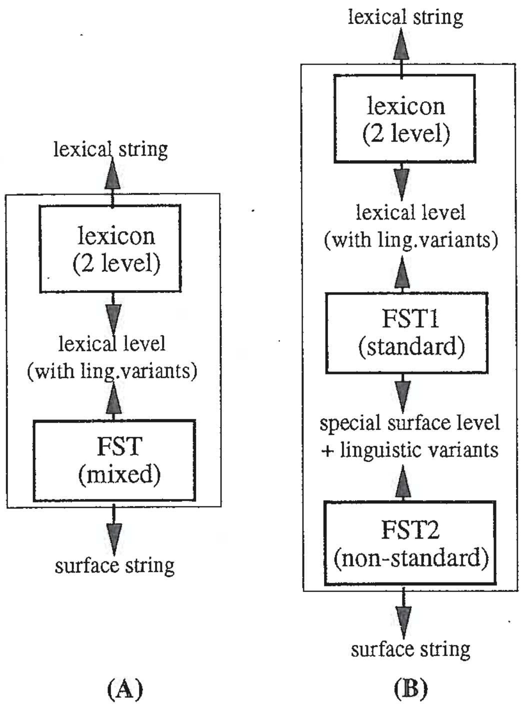

Fig. 5: *Lexical transducer for the analysis of linguistic variants*

This mechanism has the following two 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 organised with the affixes in a small two-level lexicon
2. two additional rules in order to express the relationship between the generic lemmas at lexical level and any acceptable lemma of Basque, which are combined with the standard ones

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 two-level mechanism is also used to analyse the unknown forms, and the obtention of at least one analysis is guaranteed. In order to eliminate the great number of ambiguities in the analysis, a local disambiguation process is carried out.

By using lexical transducers the two additional rules can be placed independently (see Figure 6), and so, the original rules can remain unchanged. In this case the additional subsystem is arranged close to the lexicon because it maps the transformation between generic and hypothetical lemmas at lexical level. The resultant lexical transducer is very compact and fast.

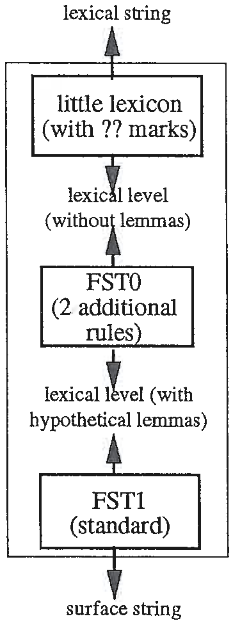
lexical string
little lexicon (with ?? marks)
lexical level (without lemmas)
FST0 (2 additional rules)
lexical level (with hypothetical lemmas)
FST1 (standard)
surface string

Fig. 6: *Lexical transducer for the analysis of unknown words*

Our system has a user lexicon and an interface to the update process too. Some information about the new entries (mainly part of speech) is necessary to add them to the user lexicon. The user lexicon is combined with the general one increasing the coverage of the morphological analyser. This mechanism is very useful in the process of spelling correction but an on-line updating of the user lexicon is necessary. This treatment is carried out in our original implementation but, when we use lexical transducers the updating operation is slow (it is necessary to compile everything together) and therefore, there are problems for on-line updating.

Carter (1995) proposes compiling affixes and rules, but no lemmas, in order to have flexibility when dealing with open lexicons, but it presents problems managing compounds at run-time.

6 Conclusions

A two-level formalism based morphological processor has been designed in a incremental way in three main modules: the standard analyser, the analyser of linguistic variants produced due to dialectal uses and competence errors, and the analyser without lexicon which can recognise word-forms without having their lemmas in the lexicon. This analyser is a basic tool for current and future work on automatic processing of Basque.

Concept	A	B	A+B
Number of words	4.846	2.343	7.207
Different words	2.607	1.429	4.036
Unknown words	307	85	392
Linguistic variants	101	28	129
Analysed	85	22	107
	(84%)	(79%)	(83%)
Full wrong analysis	21	4	25
Precision	99,2%	99,7%	**99,4%**

Table 1: *Figures about the different kinds of analysis*

Figures about the precision of the analyser are given in Table 6. Two different corpora were used: (A) a text of a magazine where foreign names appear and (B) a text about philosophy. The percents of unknown words and precision are calculated on different words, so, the results with all the corpus would be better.

Using lexical transducers for our analyser we have improved both the speed and the description of the different components of the tool. Some slight limitations have been found too.

Acknowledgements. This work had partial support from the local Government of Gipuzkoa and from the Government of the Basque Country. We would like to thank to Xerox for letting us using their tools, and also to Ken Beesley and Lauri Karttunen for their help in using these tools and designing the lexical transducers. We also want to acknowledge to Eneko Agirre for his help with the English version of this manuscript.

REFERENCES

Aduriz, Itziar, E. Agirre, I. Alegria, X. Arregi, J.M. Arriola, X. Artola, A. Diaz de Illarraza, N. Ezeiza, M. Maritxalar, K. Sarasola & M. Urkia. 1993. "A Morphological Analysis Based Method for Spelling Correction". *Proceedings of the 6th Conference of the European Association for Computational Linguistics (EACL'93)*, 463-463. Utrecht, The Netherlands.

——, E. Agirre, I. Alegria, X. Arregi, J.M. Arriola, X. Artola, Da Costa A., A. Diaz de Illarraza, N. Ezeiza, M. Maritxalar, K. Sarasola & M. Urkia. 1994. "Xuxen-Mac: un corrector ortografico para textos en euskara". *Proceedings of the 1st Conference Universidad y Macintosh, UNIMAC*, vol.II, 305-310. Madrid, Spain.

——, I. Alegria, J.M. Arriola, X. Artola, Diaz de Ilarraza A., N. Ezeiza, K. Gojenola, M. Maritxalar. 1995. "Different issues in the design of a lemmatiser/tagger for Basque". *From Text to Tag Workshop, SIGDAT (EACL'95)*, 18-23. Dublin, Ireland.

Agirre, Eneko, I. Alegria, X. Arregi, X. Artola, A. Diaz de Illarraza, M. Maritxalar, K. Sarasola & M. Urkia. 1992. "XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology". *Proceedings of the 3rd Conference Applied Natural Language Processing (ANLP'92)*, 119-125. Trento, Italy.

——, X. Arregi, J.M. Arriola, X. Artola, A. Diaz de Illarraza, J.M. Insausti & K. Sarasola. 1995. "Different issues in the design of a general-purpose Lexical Database for Basque". *Proceedings of the 1st Workshop on Applications of Natural Language to Data Bases (NLDB'95), Versailles, France*, 299-313.

Alegria, Iñaki. 1995. *Euskal morfologiaren tratamendu automatikorako tresnak.* Ph.D. dissertation, University of the Basque Country. Donostia, Basque Country.

Antworth, Evan L. 1990. *PC-KIMMO: A two-level processor for morphological analysis.* Dallas, Texas: Summer Institute of Linguistics.

Black, Alan W., Joke van de Plassche & Briony Williams. 1991. "Analysis of Unknown Words through Morphological Descomposition". *Proceedings of the 5th Conference of the European Association for Computational Linguistics (EACL'91)*, vol.I, 101-106.

Carter, David. 1995. "Rapid development of morphological descriptions for full language processing system". *Proceedings of the 5th Conference of the European Association for Computational Linguistics (EACL'95)*, 202-209. Dublin, Ireland.

Chanod, Jean-Pierre. 1994. "Finite-state Composition of French Verb Morphology". Technical Report (Xerox MLTT-005). Meylan, France: Rank Xerox Research Center, Grenoble Laboratory.

Kaplan, Ronald M. & Martin Kay. 1994. "Regular models of phonological rule systems". *Computational Linguistics* 20:3.331-380.

Karttunen, Lauri & Kenneth R. Beesley. 1992. "Two-Level Rule Compiler". Technical Report (Xerox ISTL-NLTT-1992-2). Palo Alto, Calif.: Xerox. Palo Alto Research Center.

——, Ronald M. Kaplan & Annie Zaenen. 1992. "Two-level morphology with composition". *Proceedings of the 14th Conference on Computational Linguistics (COLING'92)*, vol.I, 141-148. Nantes, France.

——. 1993. "Finite-State Lexicon Compiler". Technical Report (Xerox ISTL-NLTT-1993-04-02). Xerox. Palo Alto Research Center. 3333 Coyote Hill Road. Palo Alto, CA 94304

——. 1994. "Constructing Lexical Transducers". *Proceedings of the 15th Conference on Computational Linguistics (COLING'94)*, vol.I, 406-411. Kyoto, Japan.

Koskenniemi, Kimmo. 1983. *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production.* Publications 11. University of Helsinki.

Kwon, Hyuk-Chul & Lauri Karttunen. 1994. "Incremental construction of a lexical transducer for Korean". *Proceedings of the 15th Conference on Computational Linguistics (COLING'94)*, vol.II, 1262-1266. Kyoto, Japan.

Ritchie, Graeme D., Alan W. Black, Graham J. Russell & Stephen G. Pulman. 1992. *Computational Morphology*. Cambridge, Mass.: MIT Press.

Sproat, Richard. 1992. *Morphology and Computation*. Cambridge, Mass.: MIT Press.

II

SEMANTICS AND DISAMBIGUATION