# Analysis of Derivational Morphology by Finite State Transducers *

Cristina MOTA

**Abstract:** This paper presents a new morphological parser for INTEX aiming to recognise the words resulting from productive derivational processes that are not included in the dictionaries. The parsing of these derivatives is based on the description of morphological derivational rules, represented using finite state transducers (FST), and on the lexical information already contained in the dictionaries. In particular, it will be shown how to use FST to represent those rules; a new replace operator for INTEX transducers, which is fundamental in the description of the morphological derivational rules, will be presented as well. Furthermore, we will describe how to establish lexical constraints when locating morphological patterns.

**Keywords:** morphological parsing, finite state transducers, derivational analysis, large-coverage lexicons.

**Mots clés :** analyseur morphologique, transducteurs à état fini, analyse dérivationnelle, lexiques à large couverture.

## 1. Introduction

INTEX is a powerful environment for the processing of natural languages, based on Finite State Technology. As more complete the linguistic resources of a language are in the system, the most accurate its processing is.

The dictionaries integrated in INTEX already present a large-coverage completion (for example, the Portuguese simple word dic-

* This research was partly supported by the FCT (Programmes PRAXIS XXI, Proj. 2/2.1/CSH/775/95 and POCTI, Proj. PLP/34729/99-00).

✉ Cristina MOTA, CAUTL-IST
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
Tel: +351 21 8417170. Fax: +351 21 8417167 e-mail : cristina@label2.ist.utl.pt

tionary contains more than 1 million words). Nevertheless, it is easy to notice that new words are constantly being formed by adding affixes to words already present in the lexicon, which implies a frequent update of the resources. The inclusion of these words is fundamental, because the non-recognition of a word may compromise the whole syntactic analysis of the sentence where it is included.

Therefore, our goal is to develop a new morphological parser able to analyse those derived forms, minimising the number of unknown words of the text being analysed. With the information produced by the analyser, the linguist may decide to add, or not, the word to the lexicon.

This paper describes such analyser, which will be based on the description of morphological derivation rules using finite state transducers; to account for morphological modifications that occur during the derivational process, a new replace operator will be presented as well.

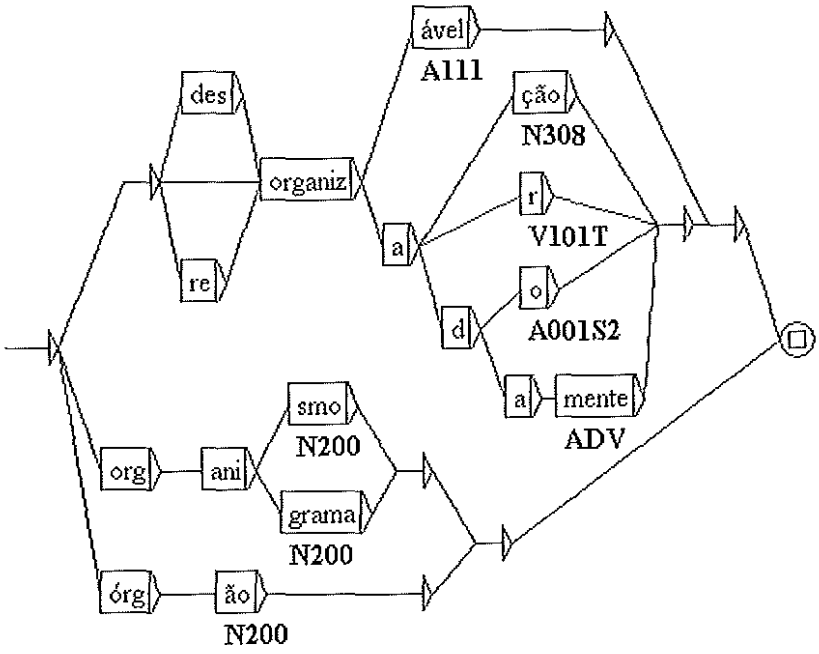
Furthermore, it will be described how to locate patterns that satisfy lexical constraints.

The use of the parser will be illustrated using examples of the Portuguese language.

## 2. Morphological Derivation in INTEX

INTEX allows the description of words created by derivational processes. The sets of words morphologically related by derivation, can be represented using a graph such as Graph 1, where the input represents the lemma of the derived form and the output its inflection code. So, this graph generates the equivalent following DELAS entries:

desorganização,N308
desorganizadamente,ADV
desorganizado,A001S2
desorganizar,V101T
desorganizável,A111
organigrama,N200
organismo,N200
organização,N308
organizadamente,ADV
organizado,A001S2
organizar,V101T
organizável,A111
órgão,N200
reorganização,N308
reorganizadamente,ADV
reorganizado,A001S2
reorganizar,V101T
reorganizável,A111



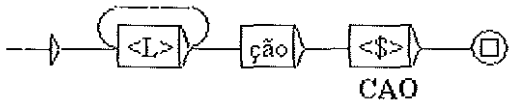

*Graph 1*

In order to locate words beginning by a certain prefix, for instance *re-*, one can use a graph such as:


*Graph 2*

which identifies words that can be either a derived form: *reabrir* (re-open), *reactivado* (reactivated), *reorganização* (reorganization), ..., or not: *rebelde* (rebeld), *rebentar* (burst), *rebuliço* (fuss), etc.

Likewise, words ending by a certain suffix, for example *-ção*, can be identified using the following graph:




*Graph 3*

which, in the same way, recognizes either derived forms: *abstracção* (abstraction), *aceleração* (acceleration), *celebração* (celebration), ..., or base forms: *coração* (*heart*), função (*function*), estação (*station*), etc.

Both graphs are applied as lexical resources, producing the corresponding lexical information:

| | |
|---|---|
| **reabre,reabre.RE** | **abstracção,abstracção.CAO** |
| **reactivadas,reactivadas.RE** | **aceleração,aceleração.CAO** |
| **reorganização,reorganização.RE** | **celebração,celebração.CAO** |
| rebeldes,rebeldes.RE | coração,coração.CAO |
| rebentam,rebentam.RE | estação,estação.CAO |
| rebuliço,rebuliço.RE | função,função.CAO |

The above forms can be located in the texts by using the regular expressions <RE> and <CAO>.

## 3. The new parser for derivatives

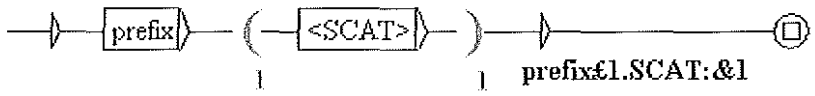
The representation of families of derivatives as described in the previous section, is an accurate method, since it only produces valid forms. Nevertheless, (i) it is time consuming (all the families of words have to be described) and (ii) it is useless to identify well-formed derived neologisms, not included in the graph but appearing in texts. So, it does not present any overall advantages in comparison with the description by lists of words.

However, it is essential to identify those neologisms resulting of productive derivational processes, as well as to assign them the corresponding lexical information.

Aiming at the identification of these derivatives, the new morphological parser uses finite state transducers, not to describe families of words having a common stem, but to describe rules of morphological derivation.

### 3.1. Prefixation Rules

In general, the morphological process of adding a prefix to a base form creates a derived word with the same lexical information of the form that generated it. We propose to describe this type of rule with a graph such as:

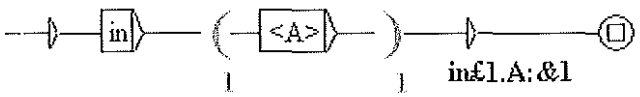



*Graph 4*

where the symbols *£1* and *&1* are local variables that store, respectively, the lemma and the inflectional information of the sequence of letters recognised has being a SCAT.

Assuming that the base form: *base,lemma.SCAT:inflectional_information* is in the dictionary, this FST identifies forms that will be associated with the following lexical information: *prefixbase, prefixlemma.SCAT:inflectional_information.*

As a simple example, the graph




*Graph 5*

identifies words formed by adding the prefix *in-* to adjectives. So, if the base adjectives:

| | |
|---|---|
| felicíssimo,feliz.A:Sms | (very happy) |
| satisfeita,satisfeito.A:fs | (satisfied) |
| capazes,capaz.A:mp | (capable) |

were in the dictionary, the derived forms, *infelicíssimo* (very unhappy), *insatisfeita* (unsatisfied) or *incapazes* (incapable) can be recognised by the above graph and associated with the following lexical information :

| | |
|---|---|
| infelicíssimo,infeliz.A:Sms | (very unhappy) |

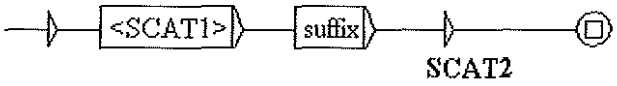
| | |
|---|---|
| insatisfeita,insatisfeito.A:fs | (unsatisfied) |
| incapazes,incapaz.A:mp | (incapable) |

### 3.2. Suffixation Rules

When a suffix is added to a base form, the resulting derivative, usually, does not have the same syntactic category of the base form. As a matter of fact, the syntactic category and inflectional paradigm of the derived form depends mostly in the suffix being added. Since some derivatives inflect and some others do not, we will make a distinction between them.

#### 3.2.1. *Non-Inflective Derivatives*

A suffixation process that produces derivatives that do not inflect, will be described using a FST of the following type:




*Graph 6*

which recognises forms constructed by adding *-suffix* to a SCAT1 form, and associates to them the lexical information: *derivative, derivative.SCAT2*.

For instance, the graph:


*Graph 7*

identifies adverbs formed by adding the suffix *-mente* to the feminine form of an adjective, recognizing forms such as: *felizmente* (happily) or *deliciosamente* (deliciously), if the adjectives:

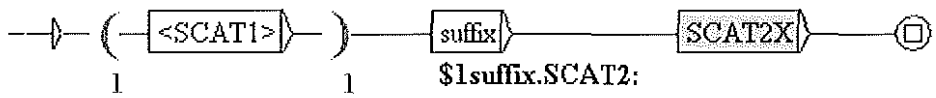
| | |
|---|---|
| feliz,feliz.A:ms:fs | (happy) |
| deliciosa,delicioso.A:fs | (delicious) |

are present in the dictionary, thus producing the lexical information:

felizmente,felizmente.ADV (happily)
deliciosamente,deliciosamente.ADV (deliciously)

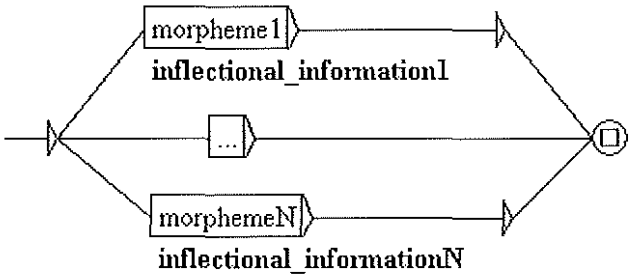
3.2.2. *Inflective Derivatives*

For the representation of suffixation rules that generate inflective derivatives, we propose to use a graph such as:




*Graph 8*

where SCAT2X is the graph that describes the inflectional paradigm associated with the suffix being added to the base form:



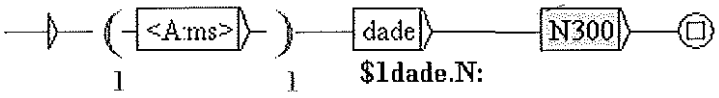

*Graph 9*

This FST generates the lexical information:

*derivative,lemma.SCAT2:inflectional_information*

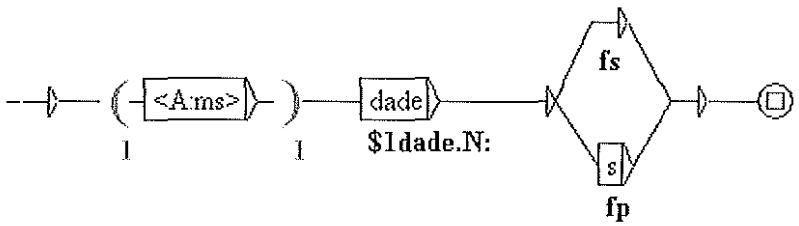
where *derivative* is a word formed by adding *suffix*, inflected by paradigm SCAT2X, to a SCTA1, *lemma* is a word formed by adding *suffix* to the SCAT1 word and *inflectional_information* is the information produced by the inflectional graph SCAT2X.

For example, the following FST:




*Graph 10*

which is equivalent to:




*Graph 11*

recognises inflected nouns derived from adding the suffix *-dade* to adjectives. So, if the adjectives:

| | |
|---|---|
| cruel,.A:ms:fs | (cruel) |
| igual,.A:ms:fs | (equal) |
| leal,.A:ms:fs | (loyal) |

were in the dictionary, the derived forms:

| | |
|---|---|
| crueldade,crueldade.N:fs | (cruelty) |
| crueldades,crueldade.N:fp | (cruelties) |
| igualdade,igualdade.N:fs | (equality) |
| igualdades,igualdade.N:fp | (equalities) |
| lealdade,lealdade.N:fs | (loyalty) |
| lealdades,lealdade.N:fp | (loyalties) |

would be identified.

### 3.3. Replace Operator

In the previous examples, the words to which the affixes were added did not undergo morphological modifications. However, this is not always the case. Actually, most cases of suffixation result in several formal modifications either in the suffix or in the form to which it is added.

For instance, the following examples show that different types of modifications may occur[1]:

[1] The noun audição could be regarded as being derived from the verb *ouvir* taking into account the morphological modification ***ouvir*** → ***audição***. However,

| | |
|---|---|
| aceit**ável** + dade → aceit**abili**dade | (acceptable + ity → acceptabi**lity**) |
| amig**ável** + mente → amig**avel**mente | (friend + ly → friend**ly**) |
| organizar + ção → organiz**ação** | (organize + ation → organiz**ation**) |
| des + **humano** → desumano | (in + human → inhuman) |
| **in** + possível → **im**possível | (**in** + possible → **im**possible) |

When morphological modifications occur, a graph such as the one to identify nouns formed by adding the suffix *-dade* to an adjective will not recognise the noun *aceitabilidade*, because *aceitabili* is not an adjective; the adjective is *aceitável*.

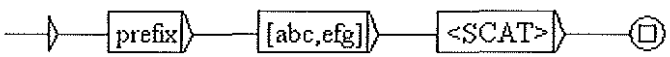
Consequently, before trying to match *aceitabili* with an adjective, it should be modified to *aceitável* and then this last form will be the one matched with the adjective.

In order to account for these and other morphological modifications in the description of derivational processes, we propose a new replace operator that will have two arguments: the source and the destination sequences, and that will be represented with the expression:

[*source,destination*]

The replace operator will be applied to an input sequence being parsed by a certain graph, and will not produce an output. Instead, it will modify the input sequence itself before further analysis.

Therefore, in a prefixation rule such as:




*Graph 12*

after *prefix* has been matched and before matching the remaining sequence *r* with a SCAT form, the replace operator will test if the sequence *r* begins by subsequence *abc*, substituting it by *efg*. So, the form that will be matched with SCAT will be a sequence beginning by *efg* and not by *abc*.

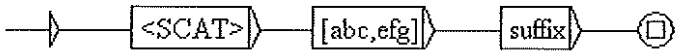
---

audição truly derives from the Latin base word *audire* (that naturally does not make part of the lexicon), and so, this and other cases of derivation involving Latin and Greek base forms will not be treated for now.

For instance, if the sequence being analysed is *prefixabcxxxx*, it will be parsed by the above prefixation rule, assuming that *efgxxxx* is a SCAT form, in the following way:

1. *prefix* matches *prefix*; the remaining input sequence is *abcxxxx*;
2. [abc,efg] matches *abc* and substitutes *abc* by *efg*; the remaining sequence is now *efgxxxx*;
3. *efgxxxx* matches <SCAT>

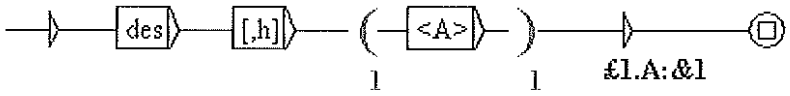
Likewise, the replace operator can be used in a suffixation rule. For example, if *xxxxefg* is a SCAT form, the graph:




*Graph 13*

parses the sequence *xxxxabcsuffix*, beginning by matching *suffix*; then, the replace operator will match *abc* in the remaining sequence, *xxxxabc*, substituting *abc* by *efg*; finally, sequence *xxxxefg* matches the SCAT form.

As a first simple example, the FST:



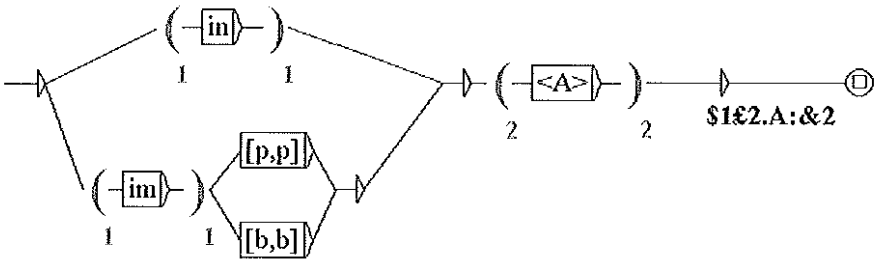

*Graph 14*

recognises adjectives formed by adding the prefix *des-* (in- or un-) to adjectives beginning by the letter *h*, such as:

| | |
|---|---|
| desumano = des + humano | (inhuman = in + human) |
| desabitado = des + habitado | (uninhabited = un + inhabited) |

In this case, the replace operator is used to add the letter *h* to the subsequence following the prefix *des-*, before verifying if it is an adjective, since the original adjective lost the *h*, during the derivational process.

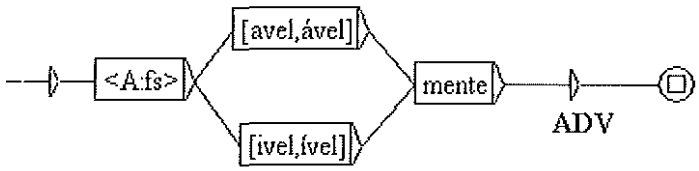
Another example of the use of the replace operator in a prefixation rule, is shown in the graph:




*Graph 15*

that identifies adjectives formed by adding the prefix *in-* to a base adjective. The replace operator is used to guarantee that if the adjective begins by a *p* or a *b*, the suffix added to the adjective form is the variation *im-* of the prefix *in-*. This graph will then recognise forms such as *infeliz* (unhappy) or *incapaz* (incapable), as well as, *impossível* (impossible) or *imbebível* (undrinkable).

In the suffixation graph:

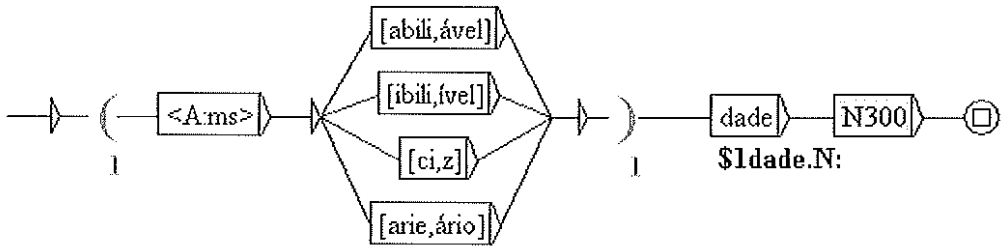



*Graph 16*

the replace operator represents the type of modifications that occur when forming adverbs from adjectives ending in *-ável* or *-ível*, such as:

| | |
|---|---|
| admira**avel**mente = admir**ável** + mente | (admirably) |
| pun**ivel**mente = pun**ível** + mente | (punishably) |

whereas, in the graph:



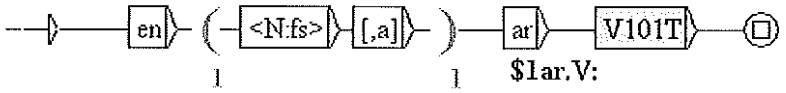

*Graph 17*

it accounts for some modifications occurring in the derivational process of nouns from adjectives by adding the suffix *-dade*. This graph will recognise nouns such as:

aplic**abili**dade = aplic**ável** +dade (applicability = applicable + ity)
poss**ibili**dade = poss**ível** + dade (possibility = possible + ity)
fel**icid**ade = fel**iz** + dade (happiness = happy + ness )
prec**arie**dade = prec**ário** + dade (precariousness = precarious + ness)

### 3.4. Combining prefixation and suffixation

There are words formed by adding to the base a prefix and a suffix simultaneously. For instance, the verb *enlatar* (to put in cans; to can) is created from the noun *lata* (can) by adding the affixes *en-* and *-ar* at the same time, since the form *enlata* is not a noun nor *latar* is a verb. In order to identify the inflected forms of verbs formed in the same way, it is necessary to use prefixation and suffixation in the same graph, as it can be seen in the FST:




*Graph 18*

So, this transducer recognises inflected forms such as *enlatámos* (we canned) and *enlatarei* (I will can), associating to them the following lexical information:

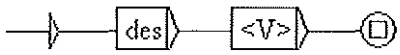
enlatarei,enlatar.V:F1s
enlatámos,enlatar.V:J1p

### 3.5. Locating patterns that satisfy lexical constraints

The derivational transducers presented in the previous sections specify lexical constraints: they match a sequence of letters that is composed by one or more subsequences that satisfy a certain lexical information, eventually after having suffered some morphological modifications.

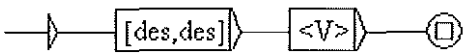
Therefore, used as automata, these transducers can locate patterns that also satisfy lexical constraints.

For instance, the automaton:




*Graph 19*

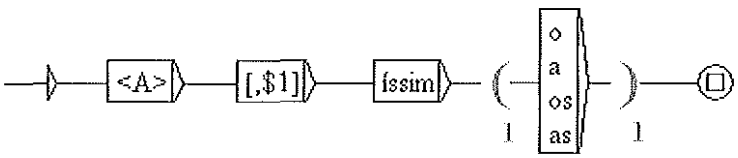
locates words, usually verbs, that are formed by adding the suffix *des-* to a verbal form, e.g., *desencaixotar* (unpack) or *desfazer* (undo), and the corresponding inflected forms, whereas the graph:




*Graph 20*

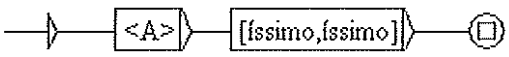
only identifies forms that are verbs beginning by sequence *des*, independently of this sequence being a prefix added to a verb, or not. As a result, besides *desencaixotar* and *desfazer* (undo), it will locate verbs such as *desenhar* (draw) or *destruir* (destroy) that are not formed by derivation from other verbs. This type of graph can be viewed as a simplification of the process of locating patterns that begin (or end) by certain subsequences (as illustrated in section 2.) and, in addition, constraints the following (or preceding) sequence to belong to a specific syntactic category.

Similarly, the FSA:




*Graph 21*

recognises the superlative form of adjectives, even if they are not included in the dictionaries as adjectives, because the graph only tests if the base form is an adjective. On the contrary, the automaton:




*Graph 22*

only identifies superlative forms of adjectives ending with the suffix -*íssimo* that are included in the dictionaries.

## 4. Conclusion and Future Work

This paper described a new morphological parser for INTEX able to analyse well-formed derived words that appear in texts but are not yet included in the dictionaries. The proposed analyser is based on the description of rules of morphological derivation, using finite state transducers. In order to represent morphological modifications that occur during the process of derivation, a new replace operator was also presented. Moreover, it was illustrated how to establish lexical constraints when locating morphological patterns.

The next step of our research work will be the implementation of the analyser and the construction of a full description of the derivational processes in Portuguese.

## Bibliography

CLEMENCEAU (David) : 1993, *Structuration du lexique et reconnaissance de mots dérivés* (Ph.D. thesis, Université ParisVII).

CLEMENCEAU (David) : 1997, "Finite-State Morphology: Inflections and Derivations in a Single Framework Using Dictionaries and Rules", *Finite-State Language Processing*, ROCHE (Emmanuel) and SCHABES (Yves) eds. (Cambridge, Mass./London : MIT Press), p. 67–98.

ELEUTÉRIO (Samuel), RANCHHOD (Elisabete), FREIRE (Helena), BAPTISTA (Jorge) : 1995, "A System of Electronic Dictionaries

of Portuguese", *Lingvisticae Investigationes*, XIX:1, (Amsterdam/Philadelphia: John Benjamins), p. 57–82.

GROSS (Maurice) : 1995, "Representation of Finite Utterances and the Automatic Parsing of Texts", *Language Research*, 31, n° 2 (Language Research Institute : Seoul), p. 291–307.

GROSS (Maurice) : 1997, "The Construction of Local Grammars", in *Finite-State Language Processing*, ROCHE (Emmanuel) and SCHABES (Yves) eds. (Cambridge, Mass./London : MIT Press), p. 329–354.

MOTA (Cristina) : 1999, "Enhancing the INTEX Morphological Parser with Lexical Constraints", *Lingvisticae Investigationes*, XXII (Amsterdam/Philadelphia: John Benjamins), p. 413–423.

RANCHHOD (Elisabete M.), MOTA (Cristina), BAPTISTA (Jorge) : 1999, "A Computational Lexicon of Portuguese for Automatic Text Parsing", in *SIGLEX99: Standardizing Lexical Resources*, 37th Annual Meeting of the ACL (College Park, Maryland, USA, June 20-26, 1999), p. 74-80, http://label2.ist.utl.pt/LabEL/siglex99.ps.

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX* (Paris : Masson).

SILBERZTEIN (Max) : 1996, *INTEX 3.4 reference manual* (Paris : LADL).

SILBERZTEIN (Max) : 1997, "The Lexical Analysis of Natural Language", in *Finite-State Language Processing*, eds. Emmanuel Roche and Yves Schabes (Amsterdam/Philadelphia: John Benjamins) p. 175-203.

SILBERZTEIN (Max) : 1998, "Finite State Transducers and the Processing of Natural Languages", distributed with INTEX version 4.12.