Multi-Point Inflection System

Samuel Eleutério[1] and Elisabete Ranchhod[2]

[1]Instituto Superior Técnico,
Av. Rovisco Pais, 1000 Lisboa
[2]Faculdade de Letras da Universidade de Lisboa,
Alam. da Universidade, 1700 Lisboa
http://label.ist.utl.pt/

Abstract. The Multipoint Inflection System is a free software developed at LabEL (Laboratório de Engenharia da Linguagem), IST, (download from: http://label.ist.utl.pt/pt/downloads_pt.php).

The algorithm was conceived for inflecting automatically multiword lexical units (MWUs). By MWU we mean a lexical unit constituted of a string of more than one simple word (for formal and linguistic details see [?], [?], [?]). Although the initial goal was to generate the inflection of Portuguese MWUs, in its present stage the algorithm is language independent, and it can be used efficiently to generate the inflection of MWUs of different languages (Romance, Germanic and Slavic languages) .

The present implementation of the algorithm was developed to be integrated into UNITEX [?], a system based on finite-state technology. It generalizes other algorithms oriented towards the inflection of single words.

Key words: Multi-words, Inflection

1 Multi-Point Inflection System

1.1 Introduction

The main goal of this program is to allow the inflection graphs to be applied to several positions of the lexical entries and not only to the last one (at the end of the string).

Such situations arise when the inflection occurs inside a word (resulting from the agglutination of two words, for instance) or when a canonical form inflects at more than one inflection point. Another example concerns multiword lexical units, where all or only some of their constituent words can inflect.

These situations are simply and easily resolved with the implementation of the *multi-point inflection system*. This new system does not change the present rules of inflection but it allows the implementation of new rules. It is a natural extension of the existing system.

1.2 Implementation

The implementation of the *multi-point inflection system* is applied by the introduction of a *multi-point inflection mark* (character '|'). This mark gives to the system the point or points at witch the inflection process takes place. If no marks exist (present situation) the inflection process occurs at the end of the lemma.

From a formal point of view, to the first mark in the lexical entry corresponds the first mark of the inflection graph (the mechanism gives at that point the same result as those we have presently at the end of the lexical entry and corresponding graph). After this, the process is reapplied at the next mark. If there is no more marks in the lexical entry (or graph), the system will be applied at the end. In the cases in which an internal point does not change in a given inflectional paradigm, this point will be marked with the *multi-point inflection mark*, and the empty mark is not necessary.

1.3 Changes in 'Inflect.cpp' program

The changes of this program in 'Inflect.cpp' program are applied by the macro:
"#define SME_LABEL_FLEX_CP 1"
where '1' is apply and '0' is not apply.

Note: the introduction of the *multi-point inflection mark* ('|') does not change the sort of the lemmas files (Delas or Delac), and the *multi-point inflection mark* is removed in the inflection process. So, the module 'sort' dictionaries will ignore this mark when sorting them.

1.4 Examples

A small dictionary will be a good example to illustrate the functioning of this program

```
abertura| contratual,N301_A311
acção| de boicote,N308
director| regional| adjunto,N005_A111_A001+NAA+Func
mesa|-de-cabeceira,N301_dht301
percurso| pedonal|,N201_A211
qual|quer,N111
sorriso| amarelo,N201_A201
```

the corresponding graphs are defined in Anexo I.

The *multi-point inflection mark* is optional when the change occurs at the end of the lemma (”abertura| contratual” and ”percurso| pedonal|”). The results are then identical even though the implementation is different. The sorted result of the inflection rules is:

```
abertura contratual,abertura contratual.N:fs
aberturas contratuais,abertura contratual.N:fp
abertura contratual,abertura contratual.N:fs
aberturas contratuais,abertura contratual.N:fp
acção de boicote,acção de boicote.N:fs
acções de boicote,acção de boicote.N:fp
director regional adjunto,director regional adjunto.N+NAA+Func:ms
directora regional adjunta,director regional adjunto.N+NAA+Func:fs
directoras regionais adjuntas,director regional adjunto.N+NAA+Func:fp
directores regionais adjuntos,director regional adjunto.N+NAA+Func:mp
mesa-de-cabeceira,mesa-de-cabeceira.N:fs
mesas-de-cabeceira,mesa-de-cabeceira.N:fp
mesinha-de-cabeceira,mesa-de-cabeceira.N:Dfs
mesinhas-de-cabeceira,mesa-de-cabeceira.N:Dfp
mesita-de-cabeceira,mesa-de-cabeceira.N:Dfs
mesitas-de-cabeceira,mesa-de-cabeceira.N:Dfp
percurso pedonal,percurso pedonal.N:ms
percursos pedonais,percurso pedonal.N:mp
quaisquer,qualquer.N:mp:fp
qualquer,qualquer.N:ms:fs
sorriso amarelo,sorriso amarelo.N:ms
sorrisos amarelos,sorriso amarelo.N:mp
```

A different interesting example of application of this new program concerns multiwords that have more than one graphical representation. This is the case of ”director regional adjunto”, which can be written with spaces (like above), with hyphens or with both spaces and hyphens (below). To obtain all these possibilities the inflection class ”N005_A111_A001” has to be changed (and renamed to ”N005_A111_A001_sh”):

```
director— regional— adjunto,N005_A111_A001_sh+NAA+Func
```

The inflection class is defined in the Anexe II, and its application gives:

```
director regional adjunto,director regional adjunto.N+NAA+Func:ms
director regional-adjunto,director regional adjunto.N+NAA+Func:ms
director-regional adjunto,director regional adjunto.N+NAA+Func:ms
director-regional-adjunto,director regional adjunto.N+NAA+Func:ms
directora regional adjunta,director regional adjunto.N+NAA+Func:fs
directora regional-adjunta,director regional adjunto.N+NAA+Func:fs
directora-regional adjunta,director regional adjunto.N+NAA+Func:fs
directora-regional-adjunta,director regional adjunto.N+NAA+Func:fs
directoras regionais adjuntas,director regional adjunto.N+NAA+Func:mp
directoras regionais-adjuntas,director regional adjunto.N+NAA+Func:mp
directoras-regionais adjuntas,director regional adjunto.N+NAA+Func:mp
directoras-regionais-adjuntas,director regional adjunto.N+NAA+Func:mp
directores regionais adjuntos,director regional adjunto.N+NAA+Func:fp
directores regionais-adjuntos,director regional adjunto.N+NAA+Func:fp
directores-regionais adjuntos,director regional adjunto.N+NAA+Func:fp
directores-regionais-adjuntos,director regional adjunto.N+NAA+Func:fp
```

Notice that the inflection graph used in this example can be simplified by introducing a sub-graph that represents the variation hyphen - space (named ”SHifen”). A second version of the graph ”N005_A111_A001_sh”

”N005_A111_A001_sh_v2”

can be used to describe this process (see Anexe II).

2 Pre-defined replacements

2.1 Definition

The *pre-defined replacements* are optional functionalities that allow to make a set of *replace* operations when such options are expressed in a graph.

The search starts from the activated position to the beginning of the word, and only the first occurrence of the list *pre-defined replacements* is applied. The search stops at the begin of the word or at the operator *multiple inflections mark* (’|’, see below).

Notice that, whenever this rule is applied to a multiword entry (string of simple words), it only operates on the specific determined word(s).

If multiple replacements are to be made, that has to be explicitly specified by using the replace ’operator’ every time a replacement is required.

The replace rule only operates if there are conditions for its application. In the case that no conditions are found, the word is not changed and no error information is returned.

The advantage of this functionality is to reduce significantly the number of the inflection classes. The changes that it introduces turn in a reduction of the number of inflection classes, which is proportional to the number of entries of the substitution list.

The code representing this option is a ’T’ in the inflection graph.

2.2 Implementation

The set of replacements to be applied by this functionality is defined in a file [1] *Presently this file is "Inflect_Replace_Char.txt", which must be in the inflection directory.*. Each line defines a replace operation, and it has two elements separated by a space: the first is the search string and the second is the replace string. The non-existence of this file corresponds to the non-existence of replace operations, so to the disregarding of that functionality.

For technical reasons the file 'Alphabet' has to be in the inflection directory till a more suitable solution is found.

2.3 Changes in 'Inflect.cpp' program

In the code of the program 'Inflect.cpp' that option is defined by the macro:

"#define SME_LABEL_CHANGE 1"

where '1' is apply and '0' is not apply.

The implementation of this functionality on Unitex should allow its activation and the introduction of the file from the implementation window. Since each language has a specific alphabet file, the address of the 'Alphabet' file has also to be made explicit.

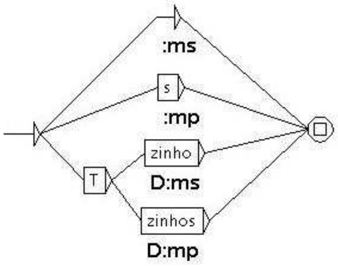
The option 'change_all', which is defined in the program, will allow the implementation of all possible replacements from de inflection window.

2.4 Exemples

In Portuguese, a very frequent situation is that the adjunction of inflectional suffixes (e.g. diminutive, superlative) often erases or changes the position of graphical accents.

1. The three situations below can be resolved by one rule (graph 'T-zinho'):

- café, cafés, cafezinho, cafezinhos
- cálice, cálices, calicezinho, calicezinhos
- estímulo, estímulos, estimulozinho, estimulozinhos

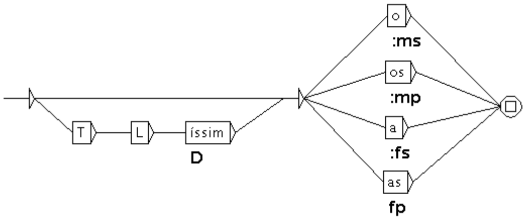


2. An identical situation is found with feminine words (graph of the same type of the previous one):

- árvore, árvores, arvorezinha, arvorezinhas
- chávena, chávenas, chavenazinha, chavenazinhas
- dúvida, dúvidas, duvidazinha, duvidazinhas

[1] (1)

3. Superlative of adjectives (T-L-íssimo):



- diminuído, diminuída, diminuídos, diminuídas,
  diminuidíssimo, diminuidíssima, diminuidíssimos, diminuidíssimas
- pálido, pálida, pálidos, pálidas,
  palidíssimo, palidíssima, palidíssimos, palidíssimas
- sério, séria, sérios, sérias,
  seriíssimo, seriíssima, seriíssimos, seriíssimas

3 Assembling homograph inflected forms (with the same lemma)

3.1 Introduction

When inflection graphs are being constructed to inflect a given lemma, it is more easy (and it avoids errors) to organize the inflection forms according to the structure of the grammatical information than according to the homography of words (for instance, in verb inflection it is more easy to organize the inflection forms by tense).

As a result of these option there is a lack of well defined criteria to assemble the homograph information of the same lexical entry. To avoid that, we conceived a rule that assembles (in the DELAF and DELACF dictionaries) all the homograph forms of a given lemma (i.e. all the information of homograph forms of a given lemma appears in the same line).

3.2 Implementation

The implementation of this rule can be made as an option in the inflection window.

3.3 Changes in 'Inflect.cpp' program

The changes to be made to the 'Inflect.cpp' program correspond to the macro:
"#define SME_LABEL_INFLECT_CONCAT 1"
where '1' is activate and '0' is do not activate.

In the program, the implementation of this rule is made by a vector of structures that contains the inflection forms of a lemma. Once the generation of all inflection forms of a lemma is concluded, the homograph forms are assembled together under the same entry. Then they will be written to the output file.

At the end, this option writes the number of inflections obtained and total number of inflections.

To activate it, a new option can be appended ('-c').

References

1. Gross, M. (1986). Lexicon-grammar. The representation of compound words. Proceedings of COLING '86.
2. Eleutério, Samuel; E. Ranchhod; H. Freire; J. Baptista (1995). A System of Electronic Dictionaries of Portuguese. Lingvisticae Investigationes, XIX:1, pp. 57-82, Amsterdam/ Philadelphia: John Benjamins.
3. Ranchhod, Elisabete (2005). Using Corpora to Increase Portuguese MWE Dictionaries. Tagging MWE in a Portuguese Corpus. In: Proceedings from The Corpus Linguistics Conference Series, Vol. 1, no. 1.
4. UNITEX, http://www-igm.univ-mlv.fr/˜unitex.

Anexe I - Multi-Point Inflection Graphs

Mini - dictionary

Inflection classes used to inflect the mini-dictionary (examples above):

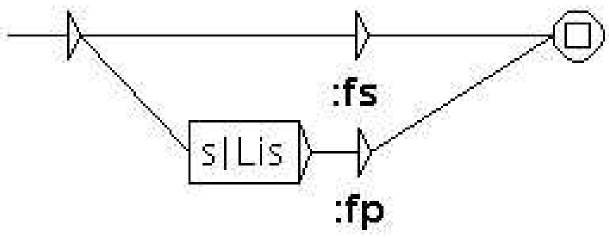


Fig. I.1 - Inflection Class "N301_A311"; Lemma: "abertura| contratual"

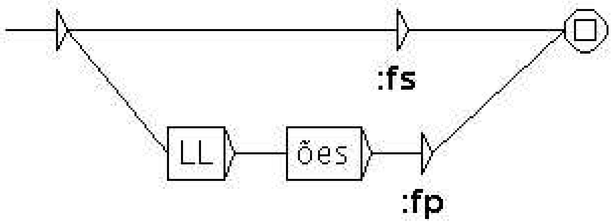


Fig. I.2 - Inflection Class "N308"; Lemma: "acção| de boicote"

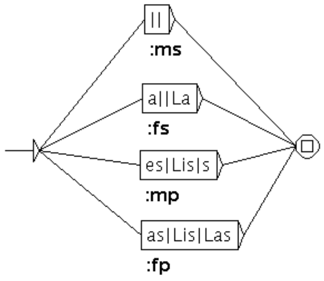


Fig. I.3 - Inflection Class "N005_A111_A001"; Lemma: "director| regional| adjunto"

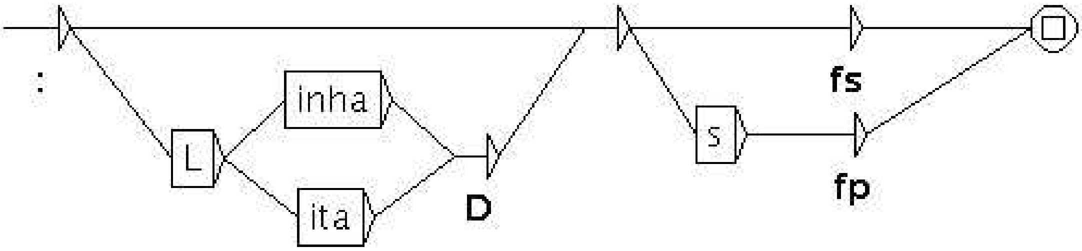


Fig. I.4 - Inflection Class ”N301_dht301”; Lemma: ”mesa|-de-cabeceira”

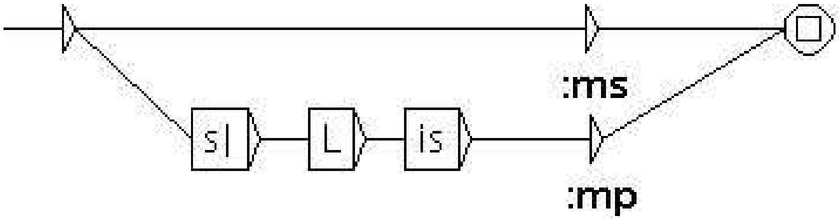


Fig. I.5 - Inflection Class ”N201_A211”; Lemma: ”percurso| pedonal|”

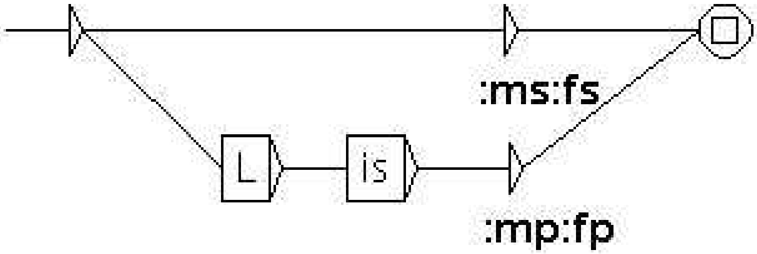


Fig. I.6 - Inflection Class ”N111”; Lemma: ”qual|quer”

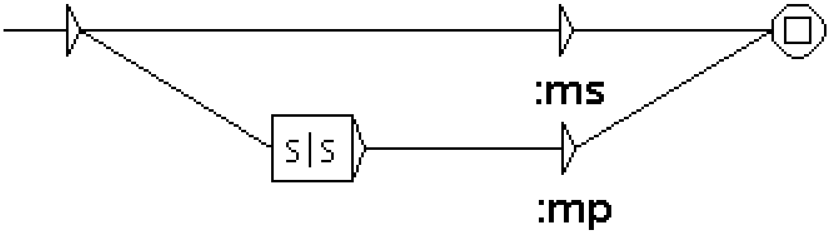


Fig. I.7 - Inflection Class ”N201_A201”; Lemma: ”sorriso| amarelo”

Unitex – Inflect

Anexe II - Multi-Point Inflection (Hyphen:Space)

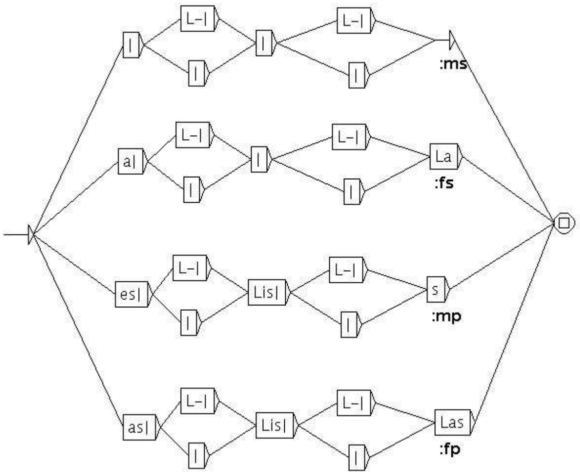


Fig. II.1 - Inflection Class "N005_A111_A001_sh"

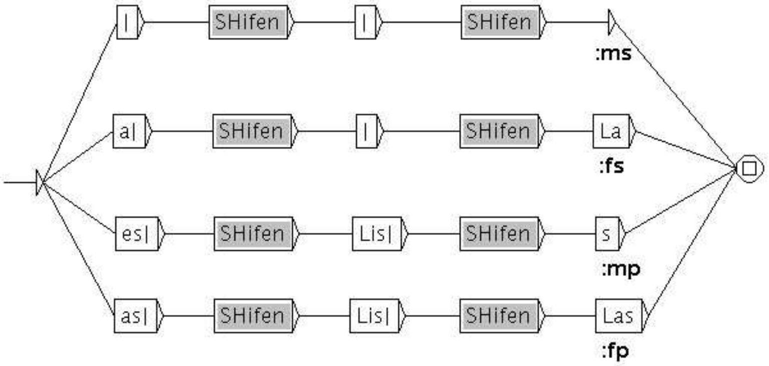


Fig. II.2 - Inflection Class "N005_A111_A001_sh_v2"
Variant of the graph "N005_A111_A001_sh_v2" with sub-graphs

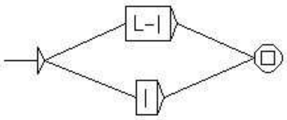


Fig. II.3 - Sub-graph "SHifen"