# Enhancing a large scale dictionary with a two-level system

**David Clemenceau & Emmanuel Roche**
LADL: Laboratoire d'Automatique Documentaire et Linguistique
Université Paris 7; 2, place Jussieu, 75251 Paris cedex 05, France
e-mail: roche@max.ladl.jussieu.fr

## 1 Introduction

We present in this paper a morphological analyzer and generator for French that contains a dictionary of 700,000 inflected words called DELAF[1], and a full two-level system aimed at the analysis of new derivatives. Hence, this tool recognizes and generates both correct inflected forms of French simple words (DELAF look-up procedure) and new derivatives and their inflected forms (two-level analysis). Moreover, a clear distinction is made between dictionary look-up processes and new words analyses in order to clearly identify the analyses that involve heuristic rules.

We tested this tool upon a French corpus of 1,300,000 words with significant results (Clemenceau D. 1992). With regards to efficiency, since this tool is compiled into a unique transducer, it provides a very fast look-up procedure (1,100 words per second) at a low memory cost (around 1.3 Mb in RAM).

## 2 Description of the analyzer

We first built the transducer representing all the entries of DELAF along with their inflectionnal code. Each entry defines a partial function, as in:

*inculpons* ↔ *inculper,V&P1p*

which corresponds to the first person plural in the present tense of the verb *inculper* (*to charge someone*). The union of these 700,000 partial functions leads to the transducer *DELAF* stored in 1Mb with a look-up procedure of 1,100 words per second.

The 70 two-level rules that describe the way characters are changed when prefixes or suffixes are added to words are themselves transducers (Karttunen *et al.*, 1992). The two following two-level rules generate the two surface forms *coïnculper* and *co-inculper* when adding the prefix *co-* to the verb *inculper*:

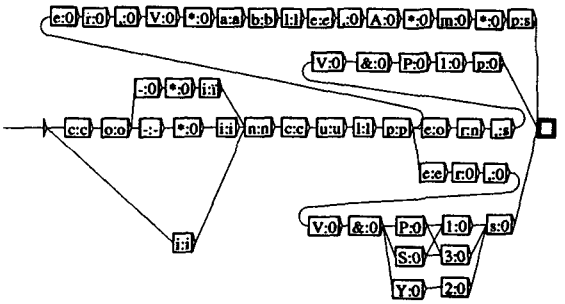
i:ï ⇔ c o -:0 * __
i:i ⇐ c o -:- * __

These 70 transducers have been merged into the transducer *Rules* by performing an intersection.

The two transducers above have been merged with four different DAGs, *Pref*, *Suf*, *DELAS* and *DELAF_A*, representing respectively a list of prefixes, a list of suffixes, the list of canonical forms (infinitive form of a verb for instance) and the whole list of the 700,000 inflected forms appearing in DELAF through the following formula:

[Trans((*Pref* • *DELAS*) ∩ *Suf*) ∩ *Rules*] ∪ [(Trans(*Pref* • *DELAF_A*) ∩ *Rules*) o (Id(*Pref* • *DELAF*)][2]

This operation leads to the transducer of 1.3Mb with a look-up procedure of 1,100 words per second, a sample of which is given in the following figure:



## 3 Results

We tested this transducer on a 1,300,000 words corpus containing 58,000 different graphical forms. Our transducer analyzed 75% of these graphical forms, which is 3% more that the transducer of DELAF alone, at a speed of 1,100 words per second. Hence, more than 97% of the word occurrences of our corpus have been analyzed in the following way:

*algorithmisation* ↔ *algorithme,N*iser,V*Ation,N*.f*.s*

## References

[Clemenceau, 1992] David Clemenceau. *Dictionary completeness and corpus analysis.* COMPLEX 92, pp. 91-100. Linguistics Institute, Hungarian Academy of Sciences, Budapest.

[Courtois, 1990] Blandine Courtois. *Un système de dictionnaires électroniques pour les mots simples du français* in Langue Française n°87, *Dictionnaires électroniques du français.* Larousse, Paris.

[Karttunen *et al.*, 1992] Lauri Karttunen, Ronald M. Kaplan, Annie Zaenen. *Two-level morphology with composition.* COLING 92, pp. 141-148.

[Koskenniemi, 1984] Kimmo Koskenniemi. *A general computational model for word-form recognition and production.* COLING 84, pp. 178-181.

[1]DELAF stands for Electronic Dictionary of Inflected Forms of the LADL (Courtois, 1990).

[2]*Trans* takes a DAG *A* and builds the transducer *Trans(A)* whose language is $L(A)x\mathcal{A}^*$. *Id* takes a DAG *A* and builds the identity function restricted to $L(A)$. The operators • and o respectively stand for concatenation and composition.