A TWO-LEVEL MORPHOLOGICAL ANALYSIS OF FRENCH

S. Lun

1. Introduction

This project is intended to deal with the orthographical changes in French nouns and adjectives when they are suffixed by feminine marker -e and/or plural marker -s. Ten finite-state automata and a small lexicon were written for this purpose. The following sections will be devoted to discussing the automata; the lexicon and examples will also be given.

2. Lexicon

The schematic French lexicon is contained in Appendix B. We have ALTERNATIONS and various lexical items in this French lexicon, and it can be expanded easily to incorporate any French nouns and adjectives. However, exceptions are not dealt with here, although they can be given different classifications (as in the English lexicon in Karttunen and Wittenburg’s paper).

Some explanation of the symbols seems needed here. I use /GN to denote animate nouns that can have either gender. /NM and /NF are masculine and feminine nouns respectively. /A denotes the class of adjectives, which can be subdivided into AM (masculine adjective) and AF (feminine adjective). The adverbial ending -ment, which can follow a feminine adjective, is not considered here.

The Root LEXICON consists of nouns and adjectives in their dictionary forms followed by a continuation class and an English translation. ALTERNATIONS state what alternations the continuation class can have. The remaining LEXICONs describe which suffixes are allowed in which order.
3. Automata

The two-level model initiated by Kimmo Koskenniemi makes use of finite-state automata which work like filters. Underlying characters (which can be morphological symbols as well as phonemes or graphemes) are paired with surface characters. I have constructed ten automata for this project and two-level rules are stated for them (using a modified version of Kimmo’s formalism). A PASCAL version is given in the appendices, although the working file is in LISP.

The first automaton is needed for technical reasons; the others correspond to the following rules. For the rest, I will number them from 2 to 10 in the same order as in the working file.

(2) (+ . 0) <---> S(s . 0)
(s . 0) <---> S(+ . 0)
(+ . 0) <---> e(+ . 0)
(e . 0) <---> e(+ . 0)
(e.g. nez+s --> nez; malade+e --> malade)

(3) (e . E) <---> r(+ . 0)e
(e.g. cher+e --> chErE, E is the e with a grave accent)

(4) (+ . h) <---> nc e
(c . q) <---> (H . u)e
(+ . u) <---> (c . q)e
(e.g. franc+e --> franche; public+e --> publique)

(5) (e . $) <---> gu(+ . 0)
($) represents e with two dots on top. It is so rare that probably it should be considered as an exception, e.g. aigu+e --> aigu$)

(6) (x . s) <---> (+ . 0)e
(f . v) <---> (+ . 0)e
(e.g. heureux+e --> heureuse; vif --> vive)

(7) (s . x) <---> V{u/(1 . u)}(+ . 0)
(when it (1 . u), V must be a)
(e.g. chou+s --> choux, jeu+s --> jeux)

(8) (1 . 0) <---> a(1 . u)(+ . 0)(s . x)
(1 . u) <---> a[(1 . 0)](+ . 0)(s . x)
(e.g. travail+s --> travaux, principal+s --> principaux)
(9) \( (r . s) \leftrightharpoons Ceu_+(+ . 0)e \)
    \( \text{(where } C \text{ cannot be } t) \)
(\( e \cdot r \) \( \leftrightharpoons t(u . i)(r . c)(+ . 0)e \)
(\( u . i \) \( \leftrightharpoons t(e . r)(r . c)(+ . 0)e \)
(\( r . c \) \( \leftrightharpoons t(e . r)(r . c)(+ . 0)e \)
\( \text{e.g. danseur+e } \rightarrow \text{ danseuse;} \)
\( \text{chanteur+e } \rightarrow \text{ chantrice} \)

(10) \( (+ . l) \leftrightharpoons \{ e/i\hat{l}e \)
    \( (+ . n) \leftrightharpoons \{ a/e/o\hat{n}e \)
    \( (+ . t) \leftrightharpoons \{ a/e/o\hat{t}e \)
\( \text{e.g. exceptionnel+e } \rightarrow \text{ exceptionnelle;} \)
\( \text {lion+e } \rightarrow \text{ lionne, chien+e } \rightarrow \text{ chiennne;} \)
\( \text{sot+e } \rightarrow \text{sotte} \)

4. Discussion

The automata I wrote up could be made more compact. One thing I didn't do is to block some orthographically prohibited sequences of feasible pairs, like \( g \) followed by \( g \) in State 2, Automata 5. The generator will thus allow more unrealistic words if they are not impossible at all. Moreover, a stronger claim for the possible orthographical sequences will block many combinations and helps reduce the enormous number of theoretically possible states when these automata are merged into one big automaton using MERGEAUTOMATA (96 actual states out of 70761600 possible states).

The automata were set up to work. Although linguistic evidence is not required to justify the merging together of some orthographic rules, I did it because it would reduce the number of automata. An evaluation measure is thus needed to decide whether we want the automata to have psychological reality or only efficiency counts.

The same feasible pair sometimes appears in many automata and careless blocking did occur when I wrote up the automata due to the fact that all automata interlock closely with each other. This made the job tedious and revision frequent.

Morphological symbols besides + seem unavoidable in French orthography. For instance, \(-e\) is a feminine marker but can also be third person singular present and others as well. Therefore, although the morphophoneme may be the same \(-e\), the graphemes used are different, e.g. \( net+e \rightarrow nette, achet*e \) (from \( acheter \) "buy") becomes \( achete \) with a grave accent above the first \( e \).
A. French Automata

ALPHABET
   a  b  c  d  e  E  $  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y  z
   +
   NULL 0
   ANY =
   SUBSET S  s  x  z
   SUBSET V  a  e  E  $  i  o  u
   SUBSET X .
   END

"Surface characters"  1 30
   a  b  c  d  e  E  $  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y  z =
   a  b  c  d  e  E  $  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y  z =
   1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

"Deletion (S+s --> S; e+e --> e)"  5 6
   S  e  +  s  e =
   S  e  0  0  0  =
   1: 2 3 1 0 0 1
   2: 1 1 4 0 0 1
   3: 2 1 5 0 0 1
   4: 0 1 0 1 0 1
   5: 1 0 0 0 1 0

"Grave Vowel Formation (er+e --> Ere)"  7 5
   e  r  +  e =
   E  r  0  e =
   1: 2 1 1 5 1
   2: 0 3 0 0 0
   3: 0 0 4 0 0
   4: 0 0 0 1 0
   5: 0 6 1 1 1
   6: 2 1 7 5 1
   7: 2 1 0 0 1
"c Change(nc+e --> nche; c+e --> que)"  8 8
 n c c + + + e =
 n q c u h 0 e =
 1: 2 5 7 0 0 1 1 1
 2: 1 0 3 0 0 1 1 1
 3: 1 0 1 0 4 6 1 1
 4: 0 0 0 0 0 0 1 0
 5: 0 0 0 4 0 0 0 0
 6: 1 0 1 0 0 0 0 1
 7: 1 0 1 0 0 8 1 1
 8: 1 0 1 0 0 0 0 1

"gu Change (gu+e --> gu$)"  4 7
 g u + e e s =
 g u 0 $ e x =
 1: 2 1 1 0 1 1 1
 2: 1 3 1 0 1 0 1
 3: 1 1 4 0 1 0 1
 4: 1 1 0 1 0 0 1

"xf Change (xf+e --> se; f+e --> ve)"  5 7
 x f + e x f =
 s v 0 e x f =
 1: 2 2 1 1 4 4 1
 2: 0 0 3 0 0 0 0
 3: 0 0 0 1 0 0 0
 4: 0 0 5 1 1 1 1
 5: 0 0 0 0 1 1 1

"Plural-x (au, eau, eu, ou+s --> aux, eaux, eux, oux)"  9 7
 g u + s s l =
 g u 0 x s l =
 1: 7 2 1 1 4 5 1
 2: 1 1 3 0 1 1 1
 3: 1 1 0 1 0 1 1
 4: 1 1 1 0 1 1 1
 5: 1 1 6 0 1 1 1
 6: 1 1 0 0 1 1 1
 7: 1 8 1 0 1 1 1
 8: 1 1 9 0 1 1 1
 9: 1 1 0 0 1 1 1
"Plural-aux (al, ail+s --> aux)" 13 12
  g a i l l u + s s e =
  g a 0 i u l u 0 x s e =
1:  8 2 0 1 0 1 6 1 0 1 1 1
2:  1 1 3 1 1 4 1 2 6 1 0 1 1 1
3:  0 0 0 0 4 0 0 0 0 0 0 0
4:  0 0 0 0 0 0 0 5 0 0 0 0
5:  1 1 0 1 0 1 1 0 1 0 0 0
6:  1 1 0 1 0 1 1 7 0 1 1 1
7:  1 1 0 1 0 1 1 1 1 1 1 1
8:  1 1 0 1 0 1 1 0 1 1 1 1
9:  1 1 0 1 0 1 1 0 0 1 1 1
10: 1 1 0 1 0 1 1 0 0 1 0 1
11: 1 1 0 1 0 1 2 1 1 0 1 1 1
12: 1 1 0 1 0 1 1 1 3 0 1 1 1
13: 1 1 0 1 0 1 1 0 0 0 1 1

"(t)eur Change (eur+e --> euse; teur+e --> trice)" 12 10
  t e u r e u r + r =
  t e u s r i c 0 r =
1:  2 7 1 0 0 0 0 1 1 1
2:  1 1 2 1 0 3 0 0 1 1 1
3:  0 0 0 0 0 4 0 0 0 0
4:  0 0 0 0 0 0 5 0 0 0
5:  0 0 0 0 0 0 6 0 0
6:  0 1 0 0 0 0 0 0 0
7:  1 1 8 0 0 0 0 1 1 1
8:  1 1 1 9 0 0 0 1 1 0
9:  0 0 0 0 0 0 6 0 0
10: 1 1 1 0 0 0 0 1 1 1
11: 1 0 1 0 0 0 0 1 1
12: 1 1 1 2 0 0 0 0 1 1 1

"Gemination (el, eil, en, at, et, ot, on+e etc.)" 9 12
  a o e i l n t + t + + + =
  a o e i l n t 1 n t 0 =
1:  3 3 4 1 1 1 1 0 0 0 1 1
2:  0 0 1 0 0 0 0 0 0 0 0
3:  1 1 1 1 1 6 7 0 0 0 1 1
4:  1 1 1 5 8 6 7 0 0 0 1 1
5:  1 1 1 1 8 1 1 0 0 0 1 1
6:  1 1 1 1 1 1 1 0 2 0 9 1
7:  1 1 1 1 1 1 1 0 0 2 9 1
8:  1 1 1 1 1 1 1 2 0 0 9 1
9:  1 1 0 1 1 1 1 0 0 0 0 1

END
B. French Lexicon

ALTERNATIONS

( /GN = /GN )
( /NM = /NM )
( /NF = /NF )
( /A = /A )
( AM = AM )
( AF = AF )
( # = # )

END

LEXICON /GN 0 /NM ""; +e /NF ""
LEXICON /NM 0 # " N M SG"; +s # " N M PL"
LEXICON /NF 0 # " N F SG"; +s # " N F PL"
LEXICON /A 0 AM ""; +e AF ""
LEXICON AM 0 # " ADJ M SG"; +s # " ADJ M PL"
LEXICON AF 0 # " ADJ F SG"; +s # " ADJ F PL"
<table>
<thead>
<tr>
<th>Word</th>
<th>Type</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>chat</td>
<td>GN</td>
<td>&quot;cat&quot;</td>
</tr>
<tr>
<td>garCon</td>
<td>NM</td>
<td>&quot;boy&quot;</td>
</tr>
<tr>
<td>femme</td>
<td>NF</td>
<td>&quot;woman&quot;</td>
</tr>
<tr>
<td>ami</td>
<td>GN</td>
<td>&quot;friend&quot;</td>
</tr>
<tr>
<td>nez</td>
<td>NM</td>
<td>&quot;nose&quot;</td>
</tr>
<tr>
<td>avis</td>
<td>NM</td>
<td>&quot;notice, advice&quot;</td>
</tr>
<tr>
<td>noix</td>
<td>NF</td>
<td>&quot;nut&quot;</td>
</tr>
<tr>
<td>chou</td>
<td>NM</td>
<td>&quot;cabbage&quot;</td>
</tr>
<tr>
<td>jeu</td>
<td>NM</td>
<td>&quot;game&quot;</td>
</tr>
<tr>
<td>journal</td>
<td>NM</td>
<td>&quot;newspaper&quot;</td>
</tr>
<tr>
<td>eau</td>
<td>NF</td>
<td>&quot;water&quot;</td>
</tr>
<tr>
<td>travail</td>
<td>NM</td>
<td>&quot;job&quot;</td>
</tr>
<tr>
<td>noyau</td>
<td>NM</td>
<td>&quot;stone&quot;</td>
</tr>
<tr>
<td>marchand</td>
<td>GN</td>
<td>&quot;merchant&quot;</td>
</tr>
<tr>
<td>Etudiant</td>
<td>GN</td>
<td>&quot;student&quot;</td>
</tr>
<tr>
<td>laitier</td>
<td>GN</td>
<td>&quot;milkman/milkmaid&quot;</td>
</tr>
<tr>
<td>danseur</td>
<td>GN</td>
<td>&quot;dancer&quot;</td>
</tr>
<tr>
<td>admirable</td>
<td>A</td>
<td>&quot;admirable&quot;</td>
</tr>
<tr>
<td>principal</td>
<td>A</td>
<td>&quot;principal&quot;</td>
</tr>
<tr>
<td>actif</td>
<td>A</td>
<td>&quot;active&quot;</td>
</tr>
<tr>
<td>vif</td>
<td>A</td>
<td>&quot;alive&quot;</td>
</tr>
<tr>
<td>amer</td>
<td>A</td>
<td>&quot;bitter&quot;</td>
</tr>
<tr>
<td>aigu</td>
<td>A</td>
<td>&quot;sharp&quot;</td>
</tr>
<tr>
<td>exceptionnel</td>
<td>A</td>
<td>&quot;exceptional&quot;</td>
</tr>
<tr>
<td>bon</td>
<td>A</td>
<td>&quot;good&quot;</td>
</tr>
<tr>
<td>ancien</td>
<td>A</td>
<td>&quot;ancient&quot;</td>
</tr>
<tr>
<td>malade</td>
<td>A</td>
<td>&quot;sick&quot;</td>
</tr>
<tr>
<td>cruel</td>
<td>A</td>
<td>&quot;cruel&quot;</td>
</tr>
<tr>
<td>net</td>
<td>A</td>
<td>&quot;net, clean&quot;</td>
</tr>
<tr>
<td>heureux</td>
<td>A</td>
<td>&quot;delighted, happy&quot;</td>
</tr>
<tr>
<td>public</td>
<td>A</td>
<td>&quot;public&quot;</td>
</tr>
<tr>
<td>blanc</td>
<td>A</td>
<td>&quot;white&quot;</td>
</tr>
<tr>
<td>Epais</td>
<td>A</td>
<td>&quot;thick&quot;</td>
</tr>
<tr>
<td>sot</td>
<td>A</td>
<td>&quot;silly, stupid&quot;</td>
</tr>
<tr>
<td>moqueur</td>
<td>A</td>
<td>&quot;mocking&quot;</td>
</tr>
<tr>
<td>protecteur</td>
<td>A</td>
<td>&quot;protective&quot;</td>
</tr>
<tr>
<td>protecteur</td>
<td>GN</td>
<td>&quot;protector&quot;</td>
</tr>
</tbody>
</table>