STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
REKENAFDELING

Automatisch scheiden van Franse lettergrepen

door

H. Brandt Corstius
en
E.G.M. Broerse

NR 3

november 1967

§1. Inleiding

Dit rapport behandelt de automatische splitsing van Franse woorden in spellingslettergrepen. Ofschoon Frans waarschijnlijk de eerste taal is waarvoor dit probleem in verband met het automatisch zetten gesteld werd [1], hebben wij geen volledig gepubliceerd programma kunnen vinden. De uitvoerigste publikatie is die van Moreau [2]; onze oplossing wijkt op een aantal punten van de zijne af. Als autoriteit op het gebied van de Franse lettergreepsplitsing namen we "Le bon usage" [3] en Larousse [4].

Doordat het Frans, in tegenstelling tot b.v. Nederlands en Duits, bijna geen samengestelde woorden kent, en doordat de spelling, in tegenstelling tot die van het Engels, vrij logisch is, is een bijna volmaakt programma van geringe lengte mogelijk. De enige moeilijkheid zou zich voordoen bij het splitsen tussen twee direct aangrenzende klinkers. Maar die splitsing is nu juist bij het afbreken aan het eind van een regel verboden [3, §89]. Voor het tellen van Franse lettergrepen is ons programma daardoor niet direct bruikbaar.

In §2 beschrijven we de door ons gekozen oplossing en geven we de vulling van het array "MEERCONS" in letters. §3 bevat de tekst van het ALGOL-60 programma en de numerieke vulling van de in te lezen arrays. §4 bespreekt de resultaten van de toepassing van het programma op een frequentielijst en op een aantal in kranten gesplitste woorden om de mate van succes te meten.

§2. Beschrijving van het programma

De comprimatietabel, zoals die in het Nederlandse [5] en Duitse [6] geval noodzakelijk was voor het vaststellen van de meerletterige klinkers, is in het Franse geval niet nodig.

Een rij mogelijke vóór- en achtervoegsels bleek bij successieve testing meer fouten te veroorzaken dan te vermijden. Er werden dus in het geheel geen voor- en achtervoegsels toegepast.
In de lijst van de 2-medeklinker-combinaties die een lettergreep beginnen, namen we op:
ch, dh, kh, ph, rh, th
dj
bl, cl, fl, gl, kl, pl, vl,
gn, mn,
br, cr, dr, fr, gr, pr, tr, vr,
cs, ks, ps, ts,
cz, tz

In de lijst van de 3-medeklinker-combinaties die een lettergreep beginnen namen we op:

sph, chr, phr, thr, pht.

Volgens [4] werd bij een "x" tussen twee klinkers niet gesplitst.

Het nu volgende programma spreekt, als vereenvoudiging van de programma's in [5] en [6], voor zich zelf. De procedures "nextsymbol" en "drukaf" zorgen voor de invoer en de uitvoer van de letters, de procedure "vul" vult de arrays STANDAARD en MEERCONS, terwijl de procedure "splits(n)" de posities van mogelijke koppeltekens in het woord van n letters uit array W bepaalt. NLCR geeft regelopvoering, PRSYM (65) drukt het koppelteken af. Het array STANDAARD zorgt dat de letters in het programma de codering a = 1 ... z = 26 krijgen.

§3. ALGOL-programma en vulling van de arrays

```
begin integer k, n, sym, koppelteller;
  integer array STANDAARD[0:127], W[1:60], MEERCONS[2:3,0:30],
  koppel[0:25];

  procedure nextsymbol;
  begin switch SW:= NIETTOEGELATEN, WOORDSCHEIDER, TEKSTAFSLUITER;
  NIETTOEGELATEN: sym:= STANDAARD[RESYM];
    goto if sym < 0 then SW[ - sym] else WOORDEENHEID;
  WOORDSCHEIDER:
  end nextsymbol;


  procedure drukaf(element); value element; integer element;
  if element ≠ 0 then PRSYM(if element < 27 then element + 9 else
  if element < 29 then element + 99 else 120);


  procedure vul;
  begin integer k, j, aantal;
    for k:= 0 step 1 until 127 do STANDAARD[k]:= read;
    for k:= 2, 3 do
    begin aantal:= MEERCONS[k,0]:= read;
      for j:= 1 step 1 until aantal do MEERCONS[k,j]:= read
    end
  end vul;


  procedure splits(n); value n; integer n;
  begin integer a, e, i, o, u, x, y, letter, volgendeletter,
    apostrof, eersteklinker, tweedeklinker, woordbegin, woordeind,
    koppelteller, aantalcons, meercons, t1, t2, klinkerteller,
    totklinkers;
    boolean klinker;
    integer array VERWIJZING, w[1:50], KLINKER[1:25];
```

```
  procedure splitsaf(letternr); value letternr; integer letternr;
  begin koppelteller:= koppelteller + 1;
    koppel[koppelteller]:= VERWIJZING[letternr];
    woordbegin:= letternr + 1; goto RESTWOORD
  end splits af;


  boolean procedure klinkers;
  begin if klinkerteller = 0 then klinkers:= false else
    begin klinkers:= true; klinkerteller:= klinkerteller - 1;
      eersteklinker:= tweedeklinker;
      tweedeklinker:= KLINKER[totklinkers - klinkerteller]
    end
  end klinkers;

  t1:= t2:= klinkerteller:= koppelteller:= 0; woordbegin:= a:= 1;
  apostrof:= 29; e:= 5; i:= 9; o:= 15; u:= 21; x:= 24; y:= 25;
  klinker:= false;
COMPR: t1:= t1 + 1;
COMPR1: t2:= t2 + 1; if t2 > n then
  begin if klinker then
    begin w[t1]:= letter; VERWIJZING[t1]:= t2 - 1;
      klinkerteller:= klinkerteller + 1;
      KLINKER[klinkerteller]:= t1
    end;
    goto EINDCOMPR
  end;
  letter:= W[t2]; if letter = apostrof then goto COMPR1;
  if letter=a\/letter=e\/letter=i\/letter=o\/letter=u\/letter=ythen
  begin klinker:= true; goto COMPR1 end;
  if klinker then
  begin klinker:= false; klinkerteller:= klinkerteller + 1;
    KLINKER[klinkerteller]:= t1; w[t1]:= W[t2 - 1];
    VERWIJZING[t1]:= t2 - 1; t1:= t1 + 1
  end;
```

```
    w[t1]:= letter; VERWIJZING[t1]:= t2; goto COMPR;
  EINDCOMPR: woordeind:= t1; totklinkers:= klinkerteller;
    tweedeklinker:= 0; klinkers;
  RESTWOORD: if ¬klinkers then goto AFWERKING;
  RESTWOORD1: aantalcons:= tweedeklinker - eersteklinker - 1;
    for t1:= 3, 2 do if t1 ≤ aantalcons then
    begin meercons:= if t1 = 3 then w[tweedeklinker - 1] × 2500 +
      w[tweedeklinker - 2] × 50 + w[tweedeklinker - 3] else
      w[tweedeklinker - 1] × 50 + w[tweedeklinker - 2];
      for t2:= MEERCONS[t1,0] step - 1 until 1 do if meercons =
      MEERCONS[t1,t2] then splitsaf(tweedeklinker - t1 - 1)
    end;
    if w[eersteklinker + 1] = x then
    begin if aantalcons = 1 then goto RESTWOORD end;
    splitsaf(tweedeklinker - 2);
  AFWERKING: koppel[0]:= koppelteller;
    koppel[koppelteller + 1]:= - 100
  end splits;


  vul;
VOORWOORD: n:= 0;
VOORWOORD1: nextsymbol; goto VOORWOORD1;
WOORDEENHEID: n:= n + 1; W[n]:= sym; nextsymbol;
  splits(n);
  koppelteller:= 1; NLCR;
  for k:= 1 step 1 until n do
  begin drukaf(W[k]); if k = koppel[koppelteller] then
    begin PRSYM(65); koppelteller:= koppelteller + 1 end
  end uitvoer gesplitste woord;
  goto VOORWOORD;
TEKSTAFSLUITER:
end
```

het array STANDAARD:

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	-1	1	2	3
4	5	6	7	8	9	10	11	12	13
14	15	16	17	18	19	20	21	22	23
24	25	26	-1	-1	-2	-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	-2	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-2	-2
29	-1	-3	-1	-1	-1	27	28		

het array MEERCONS:

30 elementen van twee letters

403	404	411	416	418
420	504	602	603	606
607	611	616	622	707
713	902	903	904	906
907	916	920	922	953
961	966	970	1303	1320

5 elementen van drie letters

45403	45416	45420	50416	20819

§4. Resultaten

Alle tellingen van geschreven Frans gaan terug op het French Word Book [7] dat grotendeels gebaseerd is op literatuur uit de 19e eeuw, en voor ons doel daarmee ongeschikt was. Een moderne telling uit 1964 [8] heeft betrekking op gesproken Frans. De lijst van 1063 woordtypen daaruit, die 91,4% van de 312000 woordtekens uitmaakt, werd met het programma in §3 gesplitst. Alle splitsingen bleken correct.

Om ook inzicht te krijgen in de toepassing bij het automatisch zetten werden 400 woorden uit zes Franse kranten [9], die daar aan het eind van de regel afgebroken waren, door het programma in lettergrepen gesplitst. Bij alle woorden bleek het programma dezelfde splitsing te geven als de zetter had aangebracht.

De gemiddelde rekentijd op de ELECTROLOGICA X8 bedroeg 25 millisec. per woord.

Het is zeker mogelijk om Franse woorden te vinden (bijvoorbeeld: dés-armer, stag-nant) die door het programma verkeerd gesplitst worden (namelijk: dé-sarmer, sta-gnant), maar hun frequentie is zodanig dat ze in onze testlijsten niet voorkwamen. Het programma kan dus toegepast worden bij het automatisch zetten van Franse tekst.

[1] Bafour, Blanchard, Raymond, Le procédé B.B.R. de composition automatique des textes. Principes du procédé et expériences préliminaires. Imprimerie Nationale, Paris, 1958.

[2] R. Moreau, Une méthode de décomposition syllabique automatique, Études de linguistique appliquée, 4(1966), p. 65-78.

[3] M. Grevisse, Le bon usage, 8ième edition, Gembloux 1964.

[4] A.V. Thomas, Dictionnaire des difficultés de la langue française, Larousse 1956.

[5] H. Brandt Corstius, Automatisch tellen en scheiden van Nederlandse lettergrepen, Mathematisch Centrum, MR 67, 1964.

[6] H. Brandt Corstius en E.G.M. Broerse, Automatisch scheiden van Duitse lettergrepen, Mathematisch Centrum, NR 2, 1967.

[7] G.E. Vander Beke, French Word Book, New York, MacMillan 1935.

[8] G. Gougenheim, P. Rivenc, R. Michéa, A. Sauvageot, L'élaboration du français fondamental, Paris 1964.

[9] Le Monde, France-Soir, L'Aurore, Le Figaro, Le Parisien libéré, Paris-Jour, allen van 16 oktober 1967.

Summary

A program for the automatic division into spelling syllables of French words is given in ALGOL 60. No exception list is used. The 1063 most frequent words in spoken French, together 91.4% of the spoken language, were rightly hyphenated by the program. From six French newspapers 400 words broken at the end of a line were collected. All words were hyphenated by the program in the same place as the typesetter had done. The program is therefore acceptable for application in the automatic typesetting of French text.

Résumé

On donne un programma en ALGOL 60 pour la division automatique des mots français en syllabes. Il n'y a pas besoin d'une liste d'exceptions. Les 1063 mots les plus fréquents du français parlé, comportant 91.4% du langage parlé, furent correctement décomposés en syllabes par ce programme. Dans six journaux français on a rassemblé 400 mots coupés à la fin d'une ligne. Tous ces mots furent coupés par le programme au même endroit où le compositeur l'avait fait.
Par conséquent le programma peut être adopté pour la composition automatique de textes français.