STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
REKENAFDELING

Automatisch scheiden van lettergrepen in Romaanse talen

door

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

NR 4

januari 1968

§1. Inleiding

Dit rapport behandelt de automatische splitsing in spellingslettergrepen van woorden in vijf Romaanse talen: Frans, Italiaans, Spaans, Portugees en Roemeens. In tegenstelling tot soortgelijke programma's voor Nederlands [1] en Duits [2] ligt hier geen grondige analyse van deze talen aan ten grondslag. Het bleek namelijk dat voor de Franse taal [3] een programma dat alleen gebaseerd was op de medeklinkercombinaties die een lettergreep kunnen beginnen, en dat dus geen gebruik maakte van lijsten met voor- en achtervoegsels en maatregelen voor speciale gevallen, zeer bevredigende resultaten gaf. Het lag daarom voor de hand te bezien in hoeverre dit eenvoudige programma ook voor andere Romaanse talen te gebruiken was. Het bleek dat hiervoor alleen een steeds andere vulling van de lijst van medeklinkercombinaties die een lettergreep kunnen beginnen noodzakelijk was. Wij geven dus een enkel programma voor alle vijf Romaanse talen. Door vooraf een getal in te lezen wordt dan de Franse (ingelezen getal = 1), Italiaanse (getal = 2), Spaanse (3), Portugese (4) of Roemeense (5) splitsing uitgevoerd.
Moreau gaf voor de Romaanse talen regels in [4]; onze oplossing wijkt op een aantal punten van de zijne af. Voor Italiaans bestaat een publicatie van Sipala [5].
Doordat, evenals bij Frans, tussen direct opvolgende klinkers niet wordt gesplitst, is het programma niet direct bruikbaar voor het tellen van lettergrepen.
In §2 beschrijven we de door ons gekozen oplossing en geven we de vullingen van het array MEERCONS in letters. §3 bevat de tekst van het ALGOL-programma en de numerieke vullingen van de in te lezen arrays. §4 bespreekt de resultaten van de toepassingen van het programma op een viertal teksten in het Italiaans, Spaans, Portugees en Roemeens om de mate van succes te meten.

§2. Beschrijving van het programma.

Het programma is haast identiek met dat voor Frans uit [3].Alleen werden de drieletterige medeklinkercombinaties, in de Romaanse talen zeer zeldzaam, veronachtzaamd. De twee dimensies van het array MEERCONS krijgen nu een andere betekenis: de eerste index geeft de taal aan, de tweede bevat de tweeletterige medeklinkercombinaties (het aantal daarvan is weer op de nulde plaats opgeborgen).
In de volgende tabel zijn de vijf vullingen van MEERCONS af te lezen (de eerste 12 hebben de vijf talen gemeen).

taal	S eerste	H tweede	L tweede	N tweede	R tweede	S tweede
alle						
1 t/m 5		CH	BL FL		BR CR DR	
			GL PL		FR GR PR	
					TR	
Frans		DH KH PH	CL KL VL	GN MN	VR	CS KS
1		RH TH				PS TS
Italiaans	SC SP	GH LH PH	CL	GN PN	VR	
2	ST SV					
Spaans			CL LL		RR	PS
3						
Portugees		LH NH	TL			
4						
Roemeens	(SL)	GH	CL SL TL	GN PN		
5						

Zoals DH voor Frans werd opgenomen i.v.m. "ajour-d'hui", zo wordt voor Italiaans LH opgenomen i.v.m. "del-l'h ...".
In het Roemeens wordt bij een woord dat op een "i" eindigt, die direct na een medeklinker staat, geen laatste lettergreep met die "i" als klinker afgesplitst.
De taalscheiders worden gevolgd door het getal dat de taal van de volgende woorden aangeeft. Het getal nul beëindigt het programma.
Voor details omtrent procedures en code zie men [3].

§3. ALGOL-programma en vulling van de arrays.

```
begin integer k, n, sym, koppelteller,taal;
  integer array STANDAARD[0:127], W[1:60], MEERCONS[1:5,0:27],
  koppel[0:25];


  procedure nextsymbol;
  begin switch SW:= NIETTOEGELATEN, WOORDSCHEIDER, TAALAFSLUITER;
  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 j:= 1 step 1 until 12 do
    begin aantal:=read; for k:=1,2,3,4,5 do MEERCONS[k,j]:=aantal end;
    for k:= 1,2,3,4,5 do
    begin aantal:= MEERCONS[k,0]:= read;
      for j:= 13 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=y then
  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;
    if 2 < aantalcons then
    begin meercons:= w[tweedeklinker - 1] × 50 + w[tweedeklinker - 2];
      for t2:= MEERCONS[taal,0] step - 1 until 1 do if meercons =
      MEERCONS[taal,t2] then splitsaf(tweedeklinker - 3)
    end;
    if w[eersteklinker + 1] = x then
    begin if aantalcons = 1 then goto RESTWOORD end;
    if taal=5 then begin if w[tweedeklinker]=i then begin if
    tweedeklinker=woordeind then begin if w[tweedeklinker-1]=W[n-1]
    then goto RESTWOORD end end end Roemeens woord op Ci;
    splitsaf(tweedeklinker - 2);
  AFWERKING: koppel[0]:= koppelteller;
    koppel[koppelteller + 1]:= - 100
  end splits;


  vul;
  taal:= read;
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;
TAALAFSLUITER: taal:= read; if taal≠0 then
  begin NEWPAGE;goto VOORWOORD end
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:

twaalf elementen voor alle talen:

403	602	606	607	616
902	903	904	906	907
916	920			

27 elementen voor Frans

404	411	416	418	420
603	611	622	707	713
922	953	961	966	970

23 elementen voor Italiaans

169	407	412	416	603
707	716	819	922	1019
1119				

16 elementen voor Spaans

603	612	918	966

15 elementen voor Portugees

412	414	620

18 elementen voor Roemeens

407	603	619	620	707
716				

§4. Resultaten

Voor de resultaten met het Franse programma zie men [3]. Voor de andere talen werden in vier teksten [6] met totaal 3500 woorden alle woorden gesplitst. Het bleek dat in het Portugees geen, in het Spaans 1, het Italiaans 2 en het Roemeens 5 fouten werden gemaakt. Dit gemiddelde foutenpercentage van 0,2% is niet het foutenpercentage bij toepassing in automatisch zetten, daar aan het eind van een regel lange woorden meer kans hebben gesplitst te moeten worden, maar is zo laag dat de programma's voor automatisch zetten bruikbaar zijn.

Summary

A program for the automatic division into spelling syllables of words in five Romance languages is given in ALGOL 60. No exception lists are used. The program is essentially the one developed earlier for French but with different inputs of consonant combinations for each of the languages: French, Italian, Spanish, Portuguese and Rumanian. The error rate is low enough for use in automatic typesetting of texts in the Romance languages.

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

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

[3] H. Brandt Corstius en E.G.M. Broerse, Automatisch scheiden van Franse lettergrepen, Mathematisch Centrum, NR 3, 1967.

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

[5] P. Sipala, Sistemi automatici per la composizione tipografica, L'Elettrotecnica n. 4 vol. LIII (1966), p. 1-6.

[6] Italiaans: Atti della Accademia Ligure di Scienze e Lettere, Genova, vol. XX (1964), pagina's 21 en 41.
Spaans: Revista de la Real Academia de Ciencias de Madrid, col LXI (1967), pagina's 285 en 383.
Portugees: Gazeta de Matemática, Lisboa, vol XXVI (1966), pagina's 17 en 33.
Roemeens: Communicarile Academiei Republicii Populare Romîne, vol XIII (1963), pagina's 65 en 72.