La syllabisation automatique du français

par

Nina CATACH

Vincent MEISSONNIER

L.I.S.H. - C.N.R.S. - PARIS

La syllabisation intervient dans de multiples opérations que l'on peut s'efforcer de réaliser sur les textes. La plus connue, absolument nécessaire pour les coupures en fin de ligne et la justification à droite en cas de composition non manuelle, est celle qui est actuellement utilisée dans l'édition (des journaux en particulier). Elle se fait en emmagasinant (déjà coupés en syllabes), les mots d'un lexique complet, par exemple le contenu d'un ou plusieurs dictionnaires. Pour l'anglo-américain, cette opération est si difficile que les correcteurs utilisent constamment le *Webster*, où toutes les coupures syllabiques de mots sont déjà indiquées, d'où perte de temps considérable.

En français, la syllabisation est de type essentiellement *phonologique*; ce qui nous a permis à la fois d'en faire une analyse systématique, avec relativement peu d'exceptions, et de réaliser conjointement deux types de coupures automatiques également importantes : coupures *graphiques* (d'une chaîne *écrite*) et coupures *phoniques* (d'une chaîne au préalable phonétisée).

Les applications de ces deux types de coupures viennent immédiatement à l'esprit : analyse de vérifications de fichiers, recherches linguistiques, littéraires, poétiques, de rythme, d'intonation, de fréquences de groupes de lettres ou de sons, enseignement assisté sur ordinateur (avec couplage des chaînes graphiques et phonétiques correspondantes et éventuellement des courbes mélodiques), conversion automatique graphème-phonème, en vue du passage à la synthèse de la parole, etc. C'est essentiellement en vue de cette dernière application que nous avons accompli ce travail.

Les recherches sur les syllabes, du point de vue acoustique-articulatoire, sont loin de pouvoir être considérées comme définitives : au contraire, les travaux de ces dernières années ont fait apparaître de nombreux points de discordance entre les linguistes à ce sujet.

Sur le plan graphique, les seules expériences que nous pouvons citer sont des recherches de fréquences des syllabes dans les listes de mots et dans les textes (Chavasse, 1948; Verglas, 1962; Baudot, 1968; R. Moreau, 1972; voir Bibliographie).

Pour notre part, nous nous en sommes tenus pour l'instant et dans ses grandes lignes au découpage traditionnel, tel qu'il se pratique couramment, en tenant compte à l'écrit des habitudes de l'usage, à l'oral des descriptions faites jusqu'ici pour le français standard par P. Léon, Delattre, etc. Ceci, pour des raisons d'applicabilité immédiate.

I. FONDEMENTS LINGUISTIQUES

A. Modèle général de description

Si nous avons respecté dans ses grandes lignes le découpage tel qu'il se pratique, il faut souligner que notre programme ne constitue pas une simple liste de mots coupés, ni un assemblage de règles trouvées uniquement par l'expérience, mais que nos règles s'efforcent de relever d'une théorie générale des rapports entre graphèmes et phonèmes en français, et du fonctionnement de notre système d'écriture (voir N. Catach, 1979 et 1980).

Brièvement, disons qu'à notre avis, et malgré les apparences, ce système est *essentiellement phonologique*, ce que confirme d'ailleurs le fonctionnement relativement régulier de notre découpage syllabique. Certes, le français connaît des coupures de type morphologique (sur les préfixes vivants en particulier). Mais il n'y a là rien de comparable aux coupures étymologiques, morphologiques et sémantiques de l'anglais par exemple.

Exemple : *té-lé-spec-ta-teur* (à cause de *télévision, téléreportage*, etc.;
mais *té-les-co-pe* (le sens du préfixe s'est perdu);
ra-dio-scrip-teur, mais *des-crip-tion, res-tric-tion*, etc.;
dé-struc-tu-rer, mais *des-truc-tion*, etc.

Deuxièmement, l'unité pour nous n'est pas la lettre, mais le *graphème*, formé soit d'une lettre, soit d'un groupe de lettres, assurant la correspondance avec une unité orale.

Nous avons ainsi dégagé 70 unités graphiques environ, dont l'ensemble est tellement régulier qu'il peut fonctionner sans contexte, par simple appel à une table d'équivalences. Cette table se substitue, dans 80 à 90 % des cas, à tout examen particulier des rencontres de lettres.

Exemple : *ch, gu, gn, ph, qu* seront appelées des *consonnes*,
comme *b, c, d, f,* etc.

ai, au, eau, ei, eu seront appelées des *voyelles*,
comme *a, e, i, o*, etc. Les voyelles accentuées également.

Nous avons représenté chaque classe de graphèmes ainsi déterminés, les accents, les blancs, la ponctuation, par des entités et par des symboles, ce qui nous permet de formaliser au maximum nos règles :

Exemple :	C (consonnes) est noté	"|"
	V (voyelles) est noté	"—"
	B (blancs) est noté	"␣"
	les accents sont notés	"&" en chaîne graphique
	la ponctuation est notée	"."

B. Cas particuliers

Nous avons accordé un soin tout particulier aux problèmes que posent malgré tout les coupures syllabiques du français, à l'écrit et à l'oral.

Ces problèmes sont essentiellement : les coupures morphologiques des préfixes vivants, les groupes consonnes +liquides, les semi-voyelles et groupes de voyelles, les voyelles nasales, les voyelles muettes finales (*e* caduc). De plus, il y a un certain nombre d'exceptions et enfin, naturellement, des ambiguïtés.

Nous avons parlé des préfixes. En ce qui concerne les groupes consonnes +liquides, qui ne sont jamais séparables (*cer-cle, prê-tre*, etc.), il nous a fallu ne pas inclure les liquides L et R parmi les

consonnes. La question se pose surtout pour des mots comme *loua/coua, liant/client, lier/crier,* où la présence d'un groupe consonne +liquide change la syllabisation de la semi-voyelle, à l'écrit et à l'oral.

En ce qui concerne les groupes de voyelles, en général seule la présence d'un accent ou d'un tréma permet de couper entre les voyelles : *a-è-de, a-é-rien,* etc.

Les voyelles nasales ont cette caractéristique qu'elles forment une seule syllabe devant consonne ou blanc, et deux syllabes devant une voyelle (elles se dénasalisent) :

Exemple : *an / â-ne*
moins / moi-ne
un / u-ne, etc.

Les finales *e, es, ent (règne, règnes, règnent,* etc.) posent, comme les *e* caducs internes, de sérieux problèmes de syllabisation et de phonétisation, qu'il nous a fallu résoudre. De plus, il faut couper différemment des mots comme *cli-ent* et ils *plient, affluent* et ils *affluent*, etc. (ambiguïtés).

C. Règles générales de syllabisation

Une fois traités ces problèmes, on remarque que la conformation standard d'une syllabe en français est, à plus de 80 %, fondamentalement de type : C + V (consonne + voyelle). Les règles principales sont les suivantes (à l'écrit) : - Une voyelle seule peut constituer une syllabe - Elle n'a besoin en principe que d'une seule consonne avant elle pour la soutenir, les autres faisant partie de la syllabe précédente : exemple : *exemp-ter, promp-ti-tu-de*. C'est la *loi de la 1ère consonne* (exception : les groupes consonnes + liquides) - Les groupes de consonnes à l'initiale ou à la finale ne sont jamais coupés - En règle générale, on coupe entre deux consonnes - Y et X (et ILL = yod) ne sont en principe jamais coupés (mais on peut couper *yo-yo, deu-xiè-me*), etc.

Au total, nous pouvons dire que notre programme, qui comprend 120 règles seulement (dont 6 exceptions et 24 règles de préfixes), permet d'obtenir des coupures écrites de type classique (coupures entre les mots graphiques, mais cependant pas de coupure sur l'apostrophe) tout à fait satisfaisantes, *sur n'importe quel texte français.*

D. La syllabisation phonique

A partir du texte phonétisé que nous obtenons avec notre programme de phonétisation automatique, nous pouvons, sur des bases légèrement différentes, et suivant une même procédure de règles de réécriture, comprenant à gauche une transcription phonétique, découper cette suite de phonèmes.

Les principales différences entre syllabe écrite et syllabe orale, en français standard, sont les suivantes :

- Les syllabes comprenant un *e* muet sont supprimées, exemple :

syllabes graphiques	syllabes phoniques
cè-de(s, ent)	SϵD (ϵ = e ouvert)
rè-gne(s, ent)	Rϵ% (% = gn)
rè-gle(s, ent)	RϵGL
2 SYLLABES	1 SYLLABE

- *Les coupures entre les mots ne sont pas régulières* : elles peuvent apparaître ou disparaître selon les cas (nous avons prévu les deux possibilités). Les liaisons sont rejetées au début du mot qui suit, exemple :
  - les yeux ⟶ L& ZYE (& = e fermé)
  - petit enfant ⟶ PETI T*F*' (* = AN)
  - bien-aimé ⟶ BYϵ Nϵ M&

- *X, Y, ILL peuvent être coupés*, exemple :
  - hexa-go-ne ⟶ ϵG-ZA-GON
  - taxi ⟶ TAK-SI
  - noya ⟶ NWA-YA
  - billet ⟶ BI-Yϵ

II. LES DIFFERENTES ETAPES DU PROGRAMME

A. Les trois étapes de la syllabisation

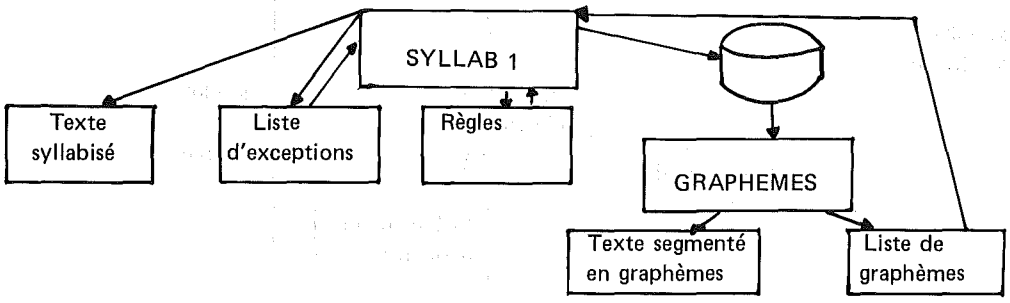
Nous sommes passés par un certain nombre d'étapes que nous appelons SYLLAB.1, SYLLAB.2, SYLLAB.3.

- L'étape SYLLAB1 présentait les caractéristiques suivantes :
  1) Elle suivait de très près l'analyse linguistique mentionnée plus haut;
  2) Elle fonctionnait de droite à gauche;
  3) Elle utilisait un certain nombre de codes et d'entités :

     Exemple : S& : syllabe formée par concaténation du contexte;
     S¬ : syllabe formée par disjonction du contexte.
  4) Elle fonctionnait sur les graphèmes réguliers classés par catégories (voyelles accentuées, non accentuées, consonnes, semi-voyelles, groupes consonnes + liquides, etc.).

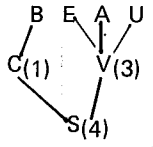
Cette étape analytique a été nécessaire pour affiner, enrichir et vérifier nos règles : ainsi, nous nous

sommes aperçus qu'en dehors des catégories ci-dessus, la catégorie de coupure syllabique elle-même (les blancs) pouvait servir à désambiguïser certains contextes. On peut représenter notre vérification d'hypothèses de la façon suivante :



D'autre part, elle nous a permis de passer plus aisément à la phonétisation automatique, dont la syllabisation graphique constituait une étape.

Exemple :

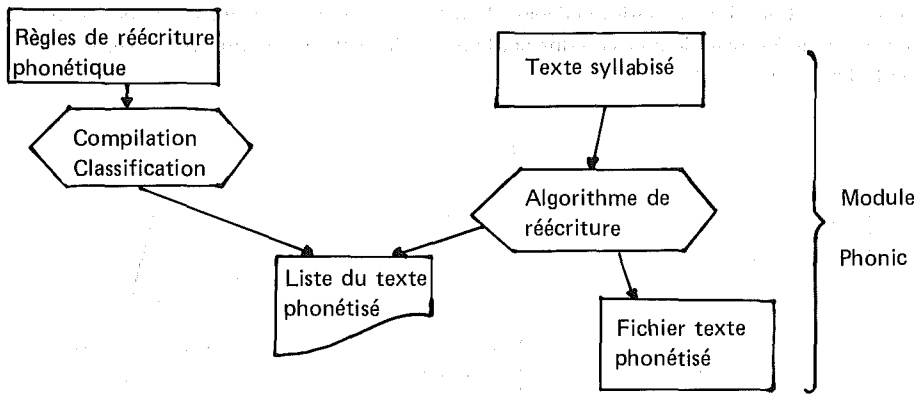


VC → S& donne alors
←

- L'étape SYLLAB2 a profité de notre expérience de phonétisation automatique. Elle présente les caractéristiques suivantes :

1) Comme la précédente, elle fonctionne sur graphèmes réguliers, catégories et un certain nombre de règles linguistiques de base;
2) La mise au clair des règles permet leur modification ou enrichissement sans modification de l'algorithme;
3) Cette étape repose sur un format unique de règles de réécriture, le même que celui utilisé pour la phonétisation, également applicable à d'autres types de recherches;
4) Elle fonctionne de gauche à droite.

Les rapports entre le module de syllabisation et celui de phonétisation (Phonic) peuvent se représenter de la façon suivante :



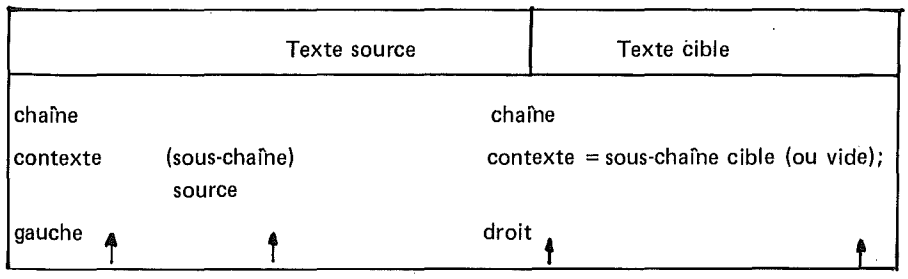
- L'étape SYLLAB3 est constituée par le passage à la syllabisation phonique à partir du texte phonétisé, et à l'aide des mêmes règles de réécriture.

On peut schématiser ces trois étapes de la façon suivante :

	Texte source	Texte cible	Résultat
SYLLAB2	texte graphique	texte graphique syllabisé	syllabes graphiques
PHONIC	texte graphique syllabisé	texte phonétisé	chaîne phonétique
SYLLAB3	texte phonétisé	texte phonétisé syllabisé	syllabes phoniques

B. Principe de la règle de réécriture

On peut résumer le principe de ce que nous appelons *règle de réécriture* de la façon suivante : une sous-chaîne dans une configuration de caractères déterminés d'un texte source est remplacée par une sous-chaîne s'inscrivant dans la séquence de caractères du texte-cible :



(N.B.- Les éléments se trouvant au-dessus des flèches sont des éléments formels séparateurs de constituants).

Au niveau de SYLLAB2, on peut avoir dans la partie gauche soit des lettres, soit des graphèmes, soit des entités graphiques (voyelles, consonnes, accents, blancs, etc.).

Au niveau de SYLLAB3, on peut avoir dans la partie gauche soit des phonèmes, soit des entités phonétiques (voyelles, consonnes, etc.).

Ces entités changent donc de contenu après la phonétisation : ainsi, *i* + voyelle, considéré comme voyelle dans un contexte graphique, sera considéré comme consonne dans un contexte phonétique, et sera traité comme tel dans la syllabisation phonique,

Exemple : iode YOD

L'ensemble des règles est classé pour traiter les chaînes les plus longues avant les plus courtes, et les unités particulières avant les symboles (pour aller du particulier au général). Cette disposition à la fois alphabétique et synthétique permet d'accéder le plus rapidement possible dans un point du texte source au sous-ensemble de règles susceptibles de s'apparier. Prenons un exemple :

Traitement du mot SURESTIMER explication

- Syllabisation (c'est l'élément entre parenthèses qui est traité)

(SUR) - = SUR␣;	SUR	Préfixe
(E) | = *;	E	Recopie simple (étoile)
(|) | - = *;	S	"
(|) - =␣*;	␣T	Blanc + recopie (étoile)
(-) = *;	|	Recopie (étoile)
(|) - =␣*;	␣M	Blanc + recopie (étoile)
(-) = *;	E	Recopie (étoile)
(|)␣= *;	R	Recopie (étoile).

- Phonétisation

S = S;
U = U;
R = R;
(E) |ـ = ϵ; (e ouvert)
(S) = S;
(T) = T;
| = |;
M = M;
(ER) = &; (e fermé)

Pour conclure, et même si (ce dont nous sommes d'accord) on peut faire un certain nombre de reproches à ce type de découpage, fondé sur un sensemble de conventions et d'usages qui seraient à revoir, nous pensons que nos programmes, reposant sur une bonne analyse de la langue écrite et orale, peut d'ores et déjà rendre de grands et réels services.

BIBLIOGRAPHIE

BAUDOT, J.A. (1968), "Information, redondance et répartition des lettres et des phonèmes en français", mémoire non publié.

CATACH, N. (1977), "Conversion automatique graphème-phonème", *Ambiguïtés de la langue écrite*, compte rendu des Actes de la Table Ronde du 11 janvier 1977, éd. CNRS - I.L.F.

CATACH, N. (1979), *L'orthographe*, Presses Universitaires de France, Collection "Que sais-je ? ".

CATACH, N., GRUAZ, C., DUPREZ, D. (1980), *L'orthographe française, traité théorique et pratique*, F. Nathan.

CATACH, N., MEISSONNIER, V. (1979), "Pour une meilleure formalisation de la conversion automatique graphème-phonème", *Xe Journées d'Etudes sur la Parole* (JEP), Grenoble, juin 1979, pp. 173-182.

CATACH, N., MEISSONNIER, V. (1980), "Formalisation et conversion automatique graphème-phonème en français", *Deuxième Conférence internationale sur les bases de données dans les Humanités et les Sciences Sociales,* Madrid, juin 1980.

CHAVASSE (1948), "La fréquence des lettres en français", *Annales des télécommunications*, tome III, n. 1, janvier.

MOREAU, R. (1972), "Recherches sur la fréquence des lettres, des suites de lettres, des syllabes et des mots en français écrit", brochure IBM intitulée *Quelques applications de l'informatique en linguistique.*

VERGLAS, A. (1962), "Remarques sur la relation entre rang et fréquence des lettres en français", *Bulletin d'information du Laboratoire d'analyse lexicologique*, n. 6.