

UN PROGRAMME D'ORDINATEUR POUR LA COUPE SYLLABIQUE EN FRANCAIS

Il est naturel qu'après avoir réalisé un programme qui effectue la coupe syllabique automatique en latin et en grec, on ait cherché à obtenir le même outil pour le français.

Dans le récent article de cette même revue (1975, n° 1), où je présentais le programme latin, je disais les raisons de son utilité ; elles sont, en gros, les mêmes pour le français, et il me semble superflu de les répéter.

LE DEGRE DE LISIBILITE

Pour les langues vivantes cependant, on peut mentionner une occasion supplémentaire de recourir au comptage des syllabes.

Ch. Muller, je l'ai dit, a montré dans ses études sur les œuvres de Corneille que le procédé peut être efficacement utilisé en stylistique ; c'est, au fond, le même principe qui est appliqué dans des recherches plus essentiellement utilitaires comme celle qui détermine le degré de lisibilité d'un texte.

Ce degré de lisibilité dépend de nombreux facteurs de natures fort diverses. G. De Landsheere (1) considère que les mesures de volumes, parmi lesquelles le nombre de syllabes, sont de bons prédicteurs. Les spécialistes ont d'ailleurs cherché à établir des formules de lisibilité tenant compte

des actions conjuguées du vocabulaire et de la syntaxe. G. De Landsheere a adapté au français la formule que R. Flesh propose pour l'anglais, et elle comporte trois variables : le nombre de phrases, de mots, de syllabes.

Ces calculs des éléments formels de la langue renseignent indirectement sur la difficulté du style ; il n'est pas douteux, pense le même spécialiste, que "beaucoup de mots savants sont construits à l'aide de suffixes ; la longueur moyenne des mots d'un texte abstrait est donc supérieure à celle d'un texte concret. De plus, les grands artifices de style ne vont guère sans allongement syntaxique." Les conclusions de Ch. Muller prouvent le fait, au moins partiellement.

Ce serait une erreur de croire que l'intérêt de cette notion de lisibilité se limite au domaine de l'éducation. Il est certes de première importance de pouvoir apprécier sur des bases numériques objectives si le style des manuels scolaires et des livres pour enfants et adolescents est adapté aux capacités de leurs lecteurs, mais il est, à d'autres égards, aussi important pour ceux qui rédigent des textes destinés à un très large public et qui ont intérêt à ce qu'ils soient efficaces (articles de propagande, circulaires administratives, questionnaires, annonces publicitaires, de recrutement...) de pouvoir calculer, avec la même objectivité, les chances qu'ils ont d'être lus et bien compris.

Le développement d'une telle technique ne se fera qu'à condition de pouvoir automatiser le calcul des éléments qui entrent dans la formule utilisée. C'est sur cette conviction que G. De Landsheere clôt le chapitre qu'il consacre à la question.

LE PROGRAMME DE SYLLABATION

Si on analyse, en vue de la syllabation automatique, les matériaux à traiter et les règles à respecter, on se trouve, comme pour le programme latin, devant la difficulté d'une part, des options à prendre, de l'autre, des problèmes de programmation.

LES PRINCIPES

La syllabation repose sur la phonétique, non sur la sémantique. C'est là un fait que l'observation d'une langue vivante prouve mieux que n'importe quel argument théorique. Aucun usager du français, même bien conscient de la composition des mots, ne fera les coupes suivantes : *neur/a/sthénie*, *courag/eux*, *in/imitable*. Le rôle que l'étymologie pourrait jouer est forcément encore plus faible en français qu'en latin puisque, pour beaucoup de termes, il dépend d'une connaissance que tous n'ont pas (latin, grec, racines germaniques, évolution phonétique...) et qui, d'ailleurs, ne peut intervenir spontanément dans l'usage normal de la langue. Les raisons que nous avons données en faveur de la coupe du latin basée sur la phonétique acquièrent donc ici plus de force encore.

Il faut admettre cependant que la syllabation reste un fait hybride et qu'elle est parfois incapable de se libérer de la sémantique, ou plutôt je crois, de l'orthographe.

La double connaissance que nous avons d'une langue vivante, connaissance auditive et visuelle, peut nous amener à deux analyses contradictoires d'un même mot. Le cas des groupes *dh, lh, nh, xh* est probant. Puisque l'*h* est muet en français, il est plus juste pour l'oreille de faire *bo/nheur* que *bon/heur, ma/lheur* que *mal/heur, i/nhumain* qu'*in/humain, e/xhaler* qu'*ex/haler*; et aussi satisfaisant pour l'esprit puisqu'on a normalement *bo/nasse, ma/ladroit, i/nimitable, e/xaucer*; mais l'œil s'insurge et pour les cas de ce genre, nous nous sommes pliés à la syllabation traditionnelle.

D'ailleurs, si on se donnait comme but de rendre fidèlement compte de la réalité phonétique, on ne saurait où s'arrêter. Il faudrait rejeter dans la seconde syllabe les consonnes doubles qui ne forment qu'un seul son : *do/nner*. Il faudrait aussi traduire les phénomènes qui affectent à tout moment la forme orale de la langue : l'élision de l'*e* muet intérieur ou final, et les liaisons de tous genres : *ell(e) parl(e)ra, ell(e)-arrivait, les-enfants, un fol-amour, il par(t)-en voiture*.

Seule l'élision de l'*e* muet final devant voyelle se fait avec une régularité qui permettrait un traitement automatique ; les autres phénomènes présentent des modalités si diversifiées et si complexes qu'on s'empêtré bientôt dans un réseau de difficultés insurmontables (2) ; on ne pourrait obtenir qu'un outil capricieux, responsable de beaucoup d'incohérences, d'erreurs et de lacunes. D'ailleurs, ce serait aller au-delà de ce que nous cherchons, ou plutôt faire autre chose, aboutir à une transcription phonétique imprécise plutôt qu'à une syllabation, au sens ordinaire du mot.

Voici, en conséquence, quels résultats il faut attendre de notre programme :

- L'e dit muet, en finale ou à l'intérieur des mots, compte toujours pour une syllabe, sauf naturellement derrière voyelle (*lierai, idée*).
- Il n'est tenu compte d'aucune liaison et chaque mot est une unité indépendante.

L'adoption de ces deux règles fait que, si l'on cherche le nombre de syllabes, c'est celui de la langue écrite que l'on obtiendra, nettement plus élevé que celui des syllabes réellement entendues : *elle parlera* (5,3), *une énorme usine* (8,5), *il y a* (3,2).

- Le rôle de l'apostrophe étant, par essence, de signaler une élision qui entraîne la contraction de deux termes, on a traité ces éléments comme formant un seul mot : *l'hom/me, aujourd'hui*.
- Par contre, on a considéré comme indépendants les termes unis par un trait d'union parce que celui-ci coïncide avec la coupe syllabique ; il n'était pas plus nécessaire de la signaler ici qu'entre deux mots ordinaires. L'intervalle entre deux mots et le trait d'union indiquent donc la fin d'une syllabe en même temps que celle d'un mot ou d'un élément non final d'un mot composé ; c'est pourquoi on ne pouvait conserver, dans le texte syllabé, le trait d'union qui suit le *t* dit euphonique : c'était donner à cette consonne isolée la valeur d'une syllabe ; pour éviter cet inconvénient et signaler le cas particulier, on a transformé le trait d'union en & : *dira-t&il*.

Si la nature des recherches qu'on entreprend exige que chaque terme reste isolé, il suffit de ne perforer ni l'apostrophe, ni le trait d'union, ou l'un des

signes au choix ; mais il est bon de se souvenir que, dans ce cas, les consonnes isolées compteront erronément pour une syllabe : un verre *d'eau*, *pense-t-elle*.

– On ne s'étonnera pas qu'au profit de l'usage normal, on ait sacrifié les exigences de la versification classique qui, dans certains cas, impose une syllabation particulière : *admira/ti/on* au lieu de *admira/tion*.

– Pour certains groupes de consonnes, surtout ceux qui contiennent *s*, les linguistes acceptent différentes coupes (*ob/stiné* ou *obs/tiné*, *in/stinctif* ou *ins/tinctif*) ; j'ai unifié autant que possible les solutions pour les cas similaires et j'ai scindé ces groupes derrière *s* (3), mais on a la plus grande facilité de changer cette option pour tous les groupes ou pour certains d'entre eux.

– Pour les groupes de voyelles, si nombreux en français, plusieurs solutions sont aussi parfois admises mais, dans quelques rares cas, les avis des spécialistes divergent (*sanc/tuai/re* ou *sanc/tu/ai/re*). J'ai suivi les données du *Dictionnaire de la Prononciation française* de L. Warnant, qui a le mérite d'être récent, complet, pratique et de s'appuyer sur des recherches sérieuses. Quand il offre le choix, j'ai préféré la forme la plus favorable au traitement automatique.

– Il arrive qu'un signe appartienne, en fait, à deux syllabes ; dans ce cas, par convention, je l'ai placé au début de la seconde : *pa/ysan*, *e/xaucer*, *e/nivrer*, *ro/yal*.

- Enfin je rappelle que, dans certains cas particuliers, on a respecté l'usage :
mal/heur, ad/héner

LA PROGRAMMATION

Le programme est prévu pour un texte non ponctué, perforé dans la typographie dite riche adoptée par le *Laboratoire d'Analyse statistique des Langues anciennes* pour ses études sur le français (4).

Comme en latin, une voyelle est suffisante mais nécessaire à l'existence d'une syllabe; le principe de base du programme sera donc ici aussi le repérage des voyelles, la nature de ce qui les suit déterminant les points de scission.

Comme en latin, il faut repérer *qu* et *gu* (celui-ci uniquement devant voyelle) puisque le *u* n'y est pas le signe d'une syllabe. Par exception, *pi/qûre* n'aura erronément que deux syllabes.

La voyelle isolée cependant n'offre guère de difficultés, mais les groupes de deux ou trois voyelles sont fréquents et constituent des réalités phonétiques très variées. Il a fallu analyser séparément le comportement de chaque signe vocalique dans les diverses positions; puis, pour la rationalisation du traitement automatique, regrouper les cas où les mêmes instructions étaient applicables. Il est évident que certaines homographies ne s'accommodent pas d'une syllabation identique et ont obligé à aller le plus loin possible dans la particularité des solutions; on comprend aussi qu'elles sont la cause des insuffisances du programme auxquelles on n'a pas toujours pu ou voulu

remédier; citons par exemple :

oue ou *ou/e* : *cloueraï clou/er*

ient ou *i/ent* : *plient cli/ent*

oe ne se scinde que dans le cas du préfixe *co* (*coexister, coefficient...*); on a donc fait un sort spécial au groupe *coe* en tête de mot tout en préservant cependant de la coupe le mot *cœur*, que les essais ont révélé très fréquent.

Par une chance inespérée, les groupes de trois voyelles n'apportent pas de difficulté supplémentaire; la voyelle intérieure réagit comme n'importe quelle deuxième voyelle, et la troisième réagit à son égard comme une seconde à l'égard d'une première.

L'énumération des différents cas paraîtra peut-être fastidieuse; pourtant il me semble qu'elle fera mieux comprendre les raisons et le caractère souvent exceptionnel des failles du programme. Les mots ou groupes de mots annoncés comme erreurs sont ceux qui reçoivent une coupe erronée ou simplement discutable.

E et Y + VOYELLE

Synérèse : *veine, veuve, nageant, nageons, eau, œuf, cerfeuil, payons, ayez*

Erreurs : quelques mots d'origine grecque composés du préfixe *poly* ou comportant *y* précédé de consonne + liquide : *polyandre, polyèdre, Amphitryon, dryade*

A et O + I, U ou E

Synérèse : *maison, paume, œuf, eau, bout.*

+ AUTRE VOYELLE

Diérèse : *a/é/er, bo/a, a/orte, pa/yant, co/opter.*

Erreurs : - *pa/on, ta/on* (sacrifiés au profit de *chaos, aorte, extraordinaire*).
- *alco/ol* (coupé d'ailleurs ainsi d'après certains spécialistes, dont L. Warnant).
- *pa/ye/ra, pa/ye/ment, pa/ye...* (mais L. Warnant considère ces coupes comme admissibles).
- *bo/y/cotter.*

COE à l'initiale veut la diérèse dans les cas les plus fréquents : *co/exister, co/ercition, co/efficient.*

É + VOYELLE

Diérèse : sauf si la voyelle est un *e* muet - *cré/ant, né/ant, ré/unir, agré/é - crée, agréera.*

Mais le signe *e* derrière *é* n'est pas toujours la marque d'un *e* muet ; la difficulté a pu être surmontée dans les deux cas principaux :

— derrière le préfixe *ré* : *ré/examiner* ; *ré/el* en profite, mais non *irréel* où le groupe *rée* n'est pas initial.

— dans les désinences *er* et *ez* des verbes en *éer* : *cré/er, supplé/ez*

Erreurs : *irrél, déesse...* (termes rares et isolés qui ne pouvaient être corrigés qu'au prix d'un repérage individuel).

U + VOYELLE

Synérèse : sauf devant *y*

sauf devant les désinences *ions* et *iez*

lueur, persuader, luire, lue, nuage, actuel, fouet, altruisme, luxueux

- *bru/yère, lou/ions, dévou/iez*

Erreur : *monstrueux* qui, à cause du groupe *tr*, doit faire *monstru/eux*.

I + VOYELLE

Synérèse : sauf préfixe *anti* + voyelle

sauf devant les désinences *ions* et *iez*

sciure, lia, nie, amitié, admiration

- *anti/esclavagiste, li/ions*

Erreurs : *anti/enne* (sacrifié au profit du préfixe *anti*)
archiépiscopal (mot isolé).

GROUPES PARTICULIERS

Certains groupes suivis de voyelle exigent la coupe devant cette voyelle sauf si elle est un *e* muet : *louer* mais *clou/er*, *jardinier* mais *ouvri/er*, *riant* mais *fri/and*, *clou/er* mais *cloueraï*.

Ce sont les groupes : *blu, clu, cru, flu*

blou, brou, clou, crou, flou

bli, bri, cli, cri, dri, gri, fri, pli, pri, tri, vri.

Le problème semble relativement simple, mais il se complique par le fait que ces groupes se rencontrent dans de nombreuses formes verbales.

Pour les autres verbes dont le radical se termine par une voyelle, seules les désinences *ions* et *iez* appellent une coupe particulière : *louer*, *louons*, *lou/ions*.

Mais ici, les fréquentes désinences commençant par *e* exigent la synérèse ou la diérèse selon que cet *e* est muet ou non : *clou/er*, *clou/ez*, *cloue*, *clouent*, *cloueraï*.

Il est évident que le traitement qu'on leur réserve s'applique automatiquement à tous les groupes homographes. La plupart des termes, dérivés ou non des verbes responsables, s'en accommodent : *cri/eur*, *ouvri/er*, *plient* *pli/ure*, *pri/ère* ... ; d'autres, qui auraient été fâcheusement affectés ont pu être épargnés : *flu/et*, *brou/ette*.. ; d'autres, enfin, en restent les victimes : *un client*, *un affluent*.. ; ils semblent heureusement rares.

Pour éviter un alourdissement du programme qui me semblait peu justifié, j'ai appliqué le traitement général aux groupes *glu* et *gru*, très rares ; on rencontrera donc de fautives synérèses dans *gluant*, *s'engluer*, *gruau*.

Les consonnes créent moins de difficultés. Comme pour le latin, il a fallu trier parmi les groupes de deux consonnes, ceux qui ne se scindent pas, et pour les groupes de trois consonnes, ceux qui veulent la coupe non après la première, mais après la seconde : *por/te* mais *peu/ple*, *ef/frayer* mais *onc/tueux*.

Quelques termes assez rares s'écartent des règles générales et recevront une coupe erronée.

Erreurs : - le groupe *gn* dans quelques mots savants : *a/gnostique*, *i/gnifuge*...

- le groupe intérieur *chr* dans quelques mots rares : *anac/hronique*, *isoc/hrome*...

- *et/hnologie*, *lon/gtemps*, *vin/gtaine*, *vin/gième*, *san/gsue*, *as/thme*, *is/thme*... (mots rares, isolés, et où le hasard fait que la ou les consonnes mal distribuées sont muettes)

- *pos/tposer*, *pos/tcolaire* (mais *pos/thume* est correct)

- *trans* suivi d'une autre consonne que *f*, *g*, *p*, *c*, *b* : *tran/slation*, *tran/slucide*, *tran/shumance*, *tran/smettre*, *tran/smuter*, *tran/svaser*, *tran/ssubstantiation*.

La plupart de ces mots peuvent facilement recevoir la solution appropriée, mais on a renoncé à alourdir le programme du cumul des traitements de ces termes assez rares.

Il a fallu considérer à part les fins de mots car un groupe ne se scindera pas à la finale comme à l'intérieur du mot ;

métem/psychose - *temps* *pay/san* - *pays*.

CONCLUSION

Si l'on accepte les principes sur lesquels on a dû le fonder, on peut estimer que le programme français donne des résultats au moins aussi satisfaisants que le programme latin alors qu'au départ, à cause de la complexité plus grande de la langue, on aurait pu craindre le contraire (5).

Si l'on compare sommairement les deux listes de mots qui échappent aux règles, il est difficile de faire une appréciation numérique précise ; mais on se souvient peut-être du problème difficile que constituent en latin quelques formes de certains composés de *ire* ; la liste française, elle, comporte presque exclusivement des mots rares et isolés.

Appliqué à 4 extraits quelconques de textes différents, qui comptent ensemble plus de 1.000 mots plurisyllabiques, le programme n'a fourni que 2 syllabations incorrectes : *client* une fois et *déesse* en 5 exemplaires dans une page consacrée à Athéna. Un tel résultat permet, je crois, de conclure qu'il est efficace.

Université de Liège
Laboratoire d'Analyse statistique
des Langues anciennes

M. DUCHESNE-DEGEY

NOTES

- (1) *Introduction à la Recherche en Education*, p. 207.
Dans un chapitre de cet ouvrage (p. 203 à 217), l'auteur fait un exposé clair et concis de la notion de lisibilité, des facteurs qui la conditionnent, des calculs qui la mesurent, et il signale les principales études qui y ont été consacrées.
- (2) Dans le *Traité de Prononciation française* de P. FOUCHE qui est, il est vrai, une étude exhaustive, 50 pages sont consacrées au seul traitement de l'e muet, et 40 aux liaisons.
- (3) Dans son *Dictionnaire de la Prononciation française*, L. WARNANT n'adopte pas un traitement uniforme ; il fait *ins/tinctif* mais *ob/stiné*.
- (4) La syllabation automatique est possible avec la typographie pauvre, mais l'absence de tréma et la confusion de e et de é provoquent un pourcentage d'erreurs plus élevé, surtout dans certains groupes de formes verbales : *reloue* et *reloué* font chacun normalement 2 syllabes, mais *cloue* en fait une alors que *cloué* en possède 2.
- (5) Ceux qui pratiquent la programmation aimeront peut-être de savoir que ces programmes PL1 comportent pour le latin environ 125 instructions, pour le français, 150.