RÈGLES DU CLASSEMENT ALPHABÉTIQUE EN LANGUE FRANÇAISE ET PROCÉDURE INFORMATISÉE POUR LE TRI

Alain LaBonté
informaticien-conseil
Secrétariat du Conseil du trésor
Gouvernement du Québec

Version revue et corrigée de l'original par l'auteur en janvier 1996 et août 1998

Les lecteurs sont invités à consulter cet autre document du même auteur : Technique de réduction/tris informatiques à quatre clés

---

Table des matières

1. Introduction

Il existe des règles non écrites chez les éditeurs de dictionnaires de langue française pour le tri alphabétique. À l'automne 1986, il nous a été possible d'obtenir ces règles auprès des éditeurs Robert et Larousse, les deux plus grands éditeurs de dictionnaires de la francophonie, par l'intermédiaire de Madame Marie-Éva de Villers, alors terminologue à l'Office de la langue française du Québec, et de Monsieur Jean-Claude Corbeil, auteur du *Dictionnaire thématique visuel*, publié chez Québec-Amérique (aussi publié en anglais sous le titre *Facts on File Visual Dictionary*).

Pour assurer la prévisibilité absolue des tris informatiques, nous avons prévu, pour des clés équivalentes selon les règles du dictionnaire, de les discriminer. La prévisibilité absolue d'un tri informatique est essentielle si on doit comparer deux fichiers triés : les tris doivent toujours donner le même résultat, quel que soit l'ordre des clés de la liste initiale à trier. Si deux clés ne différaient qu'à cause du nombre d'espaces ou de la position de ces espaces, par exemple, il y aurait danger de résultats variables, ce dont nous voulons réduire le risque pratiquement à zéro.

Le présent document a été réalisé en plusieurs phases subséquentes à l'obtention des règles par l'intermédiaire des deux spécialistes en linguistique mentionnés précédemment. En plus de ces deux personnes, nous tenons particulièrement à remercier les intervenants cités dans le court historique qui suit. Ceux-ci représentent respectivement un éditeur, une utilisatrice, un producteur de fichiers spécialisés et un grand constructeur informatique. L'ordre d'intervention de ces personnes est chronologiquement exact, mais tout à fait fortuit ; cependant ce « hasard » qui fait si bien les choses mérite d'être souligné.

À la toute fin de 1986, nous avons consulté l'informaticien responsable de l'élaboration de l'index du *Dictionnaire thématique visuel*, monsieur André Goulet, de Logidec. Celui-ci nous a expliqué les problèmes pratiques qu'il a fallu résoudre pour confectionner l'index de ce dictionnaire, qui ne comporte toutefois pas d'homographes, étant donné que le vocabulaire y est restreint à certains thèmes. Par la suite, un premier document fut produit et distribué publiquement. Madame France Chartrand, du Centre de traduction de Canadien Pacifique à Montréal, a alors soulevé des questions pertinentes quant à la façon de traiter les signes de ponctuation. Pour respecter l'ordre du dictionnaire, il fallait éliminer les tirets et autres signes spéciaux. Mais ceci a alimenté notre réflexion sur l'impact potentiel de cette élimination sur la prévisibilité du tri. Et notre modèle fut à nouveau amélioré.

À la fin de 1987, monsieur John Chandioux, de Montréal, producteur de « dictionnaires électroniques » pour différents logiciels de traitement de textes, a voulu tester concrètement notre méthode en triant par ce moyen son fichier de base et en le comparant avec la nouvelle édition du dictionnaire Robert, comme il doit le faire chaque année à cause de l'introduction de nouveaux mots. Notre méthode a fonctionné avec succès de A à Z, alors que les tris informatiques réalisés précédemment rendaient cet exercice très pénible parce que l'ordre obtenu était toujours systématiquement différent.

Finalement, monsieur Denis Garneau, du Centre technique de soutien des langues nationales d'IBM, à Toronto, a soulevé des problèmes théoriques et pratiques avec certains cas limites, qui nous ont convaincus de la nécessité de confectionner deux clés différentes pour une discrimination cohérente en fonction de la priorité des signes diacritiques et de la hauteur de casse. Auparavant nous croyions qu'une seule et même clé était nécessaire pour établir un ordre de priorité tenant compte à la fois des signes diacritiques et de la hauteur de casse, ce qui ne se vérifie pas avec certains cas très particuliers, mais quand même réels.

Ce travail a été amorcé en vue de l'élaboration d'une norme canadienne de classement informatique. Il a soulevé un intérêt si grand que l'on a jugé bon d'en faire le présent document, qui est complet en soi.

Nous avons découvert récemment que la confection de clés multiples décomposant l'information textuelle selon ses propriétés hiérarchiques essentielles (lettres de l'alphabet latin classique, signes diacritiques, hauteurs de casse et symboles spéciaux) aux fins de classement entraîne des effets secondaires importants : cela pourrait en effet résoudre de vieux problèmes touchant beaucoup d'autres applications, comme par exemple la constitution de clés permanentes (ordonnées malgré la présence d'accents ou de minuscules, de majuscules et de signes spéciaux) servant au repérage d'enregistrements dans les fichiers informatisés à l'aide des mécanismes traditionnels d'indexation, ou encore la comparaison de données accentuées avec des données analogues non accentuées (comme la comparaison directe de la racine « CLE » avec les mots « clef » ou « clé » [l'un avec, l'autre sans accent]) ou vice-versa. Qui plus est, la structure des clés est telle que l'on peut reconstituer textuellement l'information d'origine sans avoir à conserver celle-ci en double, ce qui est un facteur d'économie compensant la complexité apparente de la décomposition.

Pour ceux de nos lecteurs que le classement intéresse plus particulièrement, nous nous devons de les informer qu'il existe une norme de classement alphabétique en France (AFNOR Z.44-001), s'appliquant au classement des noms et dénominations des personnes physiques et morales, figurant dans les annuaires, les répertoires de clients ou de fournisseurs. Cette norme est complexe et non intuitive : on doit savoir d'avance si ce que l'on cherche est un nom de personne physique ou morale, les règles de classement étant différentes dans les deux cas. Il existe aussi une telle norme en Allemagne (DIN 5007), tout aussi complexe, et probablement d'autres ailleurs.

Il serait avantageux de revoir ces normes pour les rendre plus conviviales, plus intuitivement utilisables par le commun des mortels. Quant au dictionnaire, les règles établies sont passablement naturelles à utiliser, une fois les subtilités résolues sans que l'utilisateur ait à s'en préoccuper. Nous sommes convaincus que le classement du dictionnaire devrait être la base minimale sur laquelle devraient être fondés tous les tris informatiques impliquant des données textuelles. La confection de clés spécialisées en fonction d'applications précises (comme les répertoires téléphoniques) pourrait être combinée avec la méthode décrite ici pour produire un tri rationnel, prévisible et en harmonie avec les règles traditionnelles, que nous ne faisons, en définitive, que systématiser.

Il est possible que cette méthode soit encore améliorée comme ce fut le cas dans le passé, mais telle quelle, elle est le fruit de plus de deux ans de réflexion et d'essais pratiques.

2. Règles de classement alphabétique

Voici donc les règles de classement alphabétique conformes aux dictionnaires en usage dans la francophonie, augmentées des dispositions requises pour assurer la **prévisibilité absolue** du classement :

1. Pour les besoins du tri en français, on ne tient pas compte des signes diacritiques ni de la hauteur de casse (majuscules ou minuscules), sauf pour les homographes (voir règle suivante), et le classement s'effectue alors dans une séquence respectant l'ordre habituel des vingt-six lettres de l'alphabet latin (de « a » à « z »).
2. Pour les homographes, c'est-à-dire pour les mots qui s'écrivent exactement de la même façon, si on ignore les signes diacritiques pour les besoins de la cause, l'ordre de préséance des signes d'origine est respecté, avant même de donner priorité aux minuscules sur les majuscules [Dans les dictionnaires français, on n'utilise que des capitales pour les rubriques : aucune règle n'y est donc nécessaire en ce qui concerne l'ordre de préséance des minuscules et des majuscules. Dans les encyclopédies de langue française, aucune règle ne semble être suivie strictement, bien que l'on note une légère tendance à accorder priorité aux majuscules (avec de nombreuses exceptions). Étant donné qu'il semble plus naturel de donner priorité aux minuscules, nous avons choisi cet ordre de préséance, suivi très strictement dans les dictionnaires d'autres origines linguistiques (notamment anglo-saxons).]. La discrimination s'effectue cependant à rebours, *en partant de la fin du mot et en reculant*. L'ordre de priorité des lettres accentuées du français peut facilement être déduit à partir des principaux dictionnaires ; tous les dictionnaires consultés respectent l'ordre suivant :
   aA àÀ âÂ
   cC çÇ
   eE éÉ èÈ êÊ ëË
   iI îÎ ïÏ
   oO ôÔ
   uU ùÙ ûÛ üÜ
   yY ÿŸ
   Note : la lettre « ÿ » est exceptionnellement utilisée en français ex. : « L'Haÿ-les-Roses », nom d'une localité du sud de Paris).
3. Les digrammes soudés (ligatures) comme « æ » et « œ » sont classés avec les lettres doubles correspondantes, en les discriminant toutefois par un indice de priorité particulier, pour assurer la **prévisibilité absolue** du classement.
4. Tous les signes spéciaux, comme les tirets, espaces et apostrophes, sont éliminés du corps du champ temporaire de tri mais sont renvoyés à la fin du champ et précédés d'un indicateur de position, pour assurer la **prévisibilité absolue** du classement.

3. Exemples illustrant ces règles

- Pour la 1ère règle :
  - surélévation
  - sûrement
  - suréminent
  - sûreté

- Pour la 2e règle :
  - cote
  - côte
  - Côte
  - coté
  - Coté
  - côté
  - Côté
  - coter

  Note : La discrimination se fait ici d'abord sur le « e », ensuite sur le « o », ensuite sur la hauteur de casse (minuscules avant les majuscules). Le cas « COTE » est un des exemples les plus subtils d'homographes dans les dictionnaires français.

  - élève
  - élevé

  - gène
  - gêne

  - MÂCON
  - Maçon

  - pèche
  - PÈCHE
  - pêche
  - PÊCHE
  - péché
  - PÉCHÉ
  - pécher
  - pêcher

  - relève
  - relevé
  - révèle
  - révélé

- Pour la 3e règle :
  - cadurcien
  - cæcum
  - caennais
  - cæsium

        - cafard
        - coercitif
        - cœur

- Pour la 4e règle :
    - vice-consul
    - vicennal
    - vice-président
    - vice-roi
    - vicésimal
    - vice versa
    - vice-versa

4. Algorithme de confection des champs de tri

Pour le classement informatisé de texte français, la procédure suivante est proposée (les tables sont en annexe) :

I. Pour chacune des chaînes de caractères à trier :

-1. Initialiser la longueur de la clé de premier ordre (CPO), de second ordre (CSO) et de troisième ordre (CTO) à 0. Initialiser la clé de quatrième ordre (CQO) avec une chaîne contenant le caractère correspondant au nombre binaire 0.

0. En balayant la chaîne de la gauche vers la droite, obtenir le prochain caractère.

1. S'il s'agit d'un symbole qui n'est pas dans la table I (index), ajouter ce caractère à la suite de la clé de quatrième ordre (CQO), en ayant soin de faire précéder le caractère d'un code binaire de longueur fixe indiquant sa position dans le champ d'origine. Reprendre immédiatement la procédure à 0.

2. En indexant la table A avec le caractère d'origine, obtenir un ou deux (dans le cas des digrammes soudés ou ligatures) caractères de remplacement et les ajouter **à la suite** de la clé de premier ordre (CPO).

3. En indexant la table B avec le même caractère d'origine, obtenir un ou deux caractères, constituant l'indice de priorité des signes diacritiques, et les ajouter **devant** la clé de second ordre (CSO), qui joue alors le rôle d'une pile (l'ordre de confection est inversé et pour retrouver la correspondance avec l'information originale, il faut lire la chaîne à rebours).

4. En indexant la table C toujours avec le même caractère d'origine, obtenir un ou deux caractères, constituant l'indice de hauteur de casse, et les ajouter **à la suite** de la clé de troisième ordre (CTO).

5. Tant que la chaîne-source n'est pas épuisée, reprendre la procédure à 0.

6. Concaténer les clés de premier ordre (CPO), de deuxième ordre (CDO), de troisième ordre (CTO) et de quatrième ordre (CQO) pour constituer la clé complète de tri.

II. Une fois effectuée la conversion de toutes les chaînes de la liste à trier, classer selon l'ordre binaire (croissant ou décroissant) du champ temporaire de tri de chacun des éléments de la liste.

Exemple :

LISTE À CLASSER (déjà dans l'ordre)	LISTE CONVERTIE POUR LE TRI
cæcal	*caecal*<16><16><16><22><22><16>* *<08><08><08><08><08><08>*<00>*
caennais	*caennais*<16><16><16><16><16><16><16><16>* *<08><08><08><08><08><08><08><08>*<00>*
C.A.F.	*caf*<16><16><16>*<09><09><09>*<00><01>.<03>.<05>.*
c'est-à-dire	*cestadire<16><16><16><16><18><16><16><16><16>* *<08><08><08><08><08><08><08><08><08>* *<00><01>'<05>-<07>-*

jésus *jesus*<16><16><16><17><16>*
<08><08><08><08><08><00>*
Jésus *jesus*<16><16><16><17><16>*
<09><08><08><08><08><00>*
pêche *peche*<16><16><16><19><16>*
<08><08><08><08><08><00>*
PÊCHE *peche*<16><16><16><19><16>*
<09><09><09><09><09><00>*
péché *peche*<17><16><16><17><16>*
<08><08><08><08><08><00>*
PÉCHÉ *peche*<17><16><16><17><16>*
<09><09><09><09><09><00>*
pechère *pechere*<16><16><18><16><16><16><16>*
<08><08><08><08><08><08><08><00>*
péchère *pechere*<16><16><18><16><16><17><16>*
<08><08><08><08><08><08><08><00>*
vice-légat *vicelegat*<16><16><16><17><16><16><16><16><16>*
<08><08><08><08><08><08><08><08><08>
<00><04>-
vice versa *viceversa*<16><16><16><16><16><16><16><16><16>*
<08><08><08><08><08><08><08><08><08>
*<00><04> *
vice versa *viceversa*<16><16><16><16><16><16><16><16><16>*
<08><08><08><08><08><08><08><08><08>
*<00><04> <05> *
vice-versa *viceversa*<16><16><16><16><16><16><16><16><16>*
<08><08><08><08><08><08><08><08><08>
<00><04>-

Note : L'astérisque (*) n'est pas utilisé pour composer la chaîne de tri mais délimite seulement celle-ci pour montrer la présence d'espaces à l'intérieur du champ à trier et séparer les quatre ordres de clés.

Annexe

	TABLE A	TABLE B	TABLE C
	-------	-------	-------
Caractères d'origine	Caractères de remplacement non accentués	Indices de priorité des diacritiques	Indice de hauteur de casse
0	0	<16>	<08>
1	1	<16>	<08>
2	2	<16>	<08>
3	3	<16>	<08>
4	4	<16>	<08>
5	5	<16>	<08>
6	6	<16>	<08>
7	7	<16>	<08>
8	8	<16>	<08>
9	9	<16>	<08>
a	a	<16>	<08>
A	a	<16>	<09>
à	a	<18>	<08>
À	a	<18>	<09>
â	a	<19>	<08>
Â	a	<19>	<09>
b	b	<16>	<08>
B	b	<16>	<09>
c	c	<16>	<08>
C	c	<16>	<09>
ç	c	<21>	<08>
Ç	c	<21>	<09>
d	d	<16>	<08>
D	d	<16>	<09>
e	e	<16>	<08>
E	e	<16>	<09>
é	e	<17>	<08>
É	e	<17>	<09>
è	e	<18>	<08>
È	e	<18>	<09>
ê	e	<19>	<08>
Ê	e	<19>	<09>
ë	e	<20>	<08>
Ë	e	<20>	<09>
f	f	<16>	<08>
F	f	<16>	<09>

g	g	<16>	<08>
G	g	<16>	<09>
h	h	<16>	<08>
H	h	<16>	<09>
i	i	<16>	<08>
I	i	<16>	<09>
î	i	<19>	<08>
Î	i	<19>	<09>
ï	i	<20>	<08>
Ï	i	<20>	<09>
j	j	<16>	<08>
J	j	<16>	<09>
k	k	<16>	<08>
K	k	<16>	<09>
l	l	<16>	<08>
L	l	<16>	<09>
m	m	<16>	<08>
M	m	<16>	<09>
n	n	<16>	<08>
N	n	<16>	<09>
o	o	<16>	<08>
O	o	<16>	<09>
ô	o	<19>	<08>
Ô	o	<19>	<09>
p	p	<16>	<08>
P	p	<16>	<09>
q	q	<16>	<08>
Q	q	<16>	<09>
r	r	<16>	<08>
R	r	<16>	<09>
s	s	<16>	<08>
S	s	<16>	<09>
t	t	<16>	<08>
T	t	<16>	<09>
u	u	<16>	<08>
U	u	<16>	<09>
ù	u	<18>	<08>
Ù	u	<18>	<09>
û	u	<19>	<08>
Û	u	<19>	<09>
ü	u	<20>	<08>
Ü	u	<20>	<09>
v	v	<16>	<08>
V	v	<16>	<09>
w	w	<16>	<08>

W	w	<16>	<09>
x	x	<16>	<08>
X	x	<16>	<09>
y	y	<16>	<08>
Y	y	<16>	<09>
ÿ	y	<20>	<08>
Ÿ	y	<20>	<09>
z	z	<16>	<08>
Z	z	<16>	<09>
æ	ae	<22><22>	<08><08>
Æ	ae	<22><22>	<09><09>
œ	oe	<22><22>	<08><08>
Œ	oe	<22><22>	<09><09>
<NBSP>	<espace>	<16>	<08>

Notes :

1. Le caractère NBSP (espace insécable) est défini dans la norme ISO/CEI 8859-1 ; l'espace insécable est essentiel si on veut vraiment séparer, par exemple, les noms des prénoms, dans une liste informatique : il est plus naturel de classer « Dion Zacharie » avant « Dionne Herménégilde », ce qui ne serait pas le cas si on éliminait l'espace. Les espaces insécables sont nécessaires pour les distinguer des espaces non significatifs, comme dans les noms de famille écrits en plusieurs mots (voir celui de l'auteur), espaces qui, dans ces cas, doivent être déplacés pour le tri.
2. Forcément, les indices devront être codés en binaire dans une application informatique. On prendra grand soin à ce que l'ensemble de codes utilisés pour représenter les indices n'ait pas d'intersection avec l'ensemble des codes utilisés pour les caractères de remplacement non accentués, ce qui pourrait nuire à l'ordre de classement de champs de longueurs variables, comme dans l'exemple suivant :

Champ à trier	Clé intermédiaire
A1	a1<16><16><09><08><00>
à	a<18><08><00>

Si, dans cet exemple, <18> avait par hasard un code correspondant au caractère « 4 », il y aurait intersection avec l'ensemble des caractères de remplacement et le classement serait erronné, puisque la chaîne « à », de taille plus petite, devrait se retrouver en premier. L'ensemble servant à confectionner les indices doit comprendre les nombres binaires les plus petits.

Dans le même ordre d'idées, on recommande de réserver les codes <01> à <15> pour les indices de hauteur de casse (même si on n'en utilise ici que deux) et <16> à <31> pour les indices de priorité des diacritiques, ceux-ci devant être plus grands que les premiers, si on utilise des champs de longueurs variables. Les codes inutilisés pourraient être utiles pour généraliser la méthode avec d'autres langues que le français. Le code <00> doit être réservé pour délimiter la clé de quatrième ordre si on utilise des champs de longueurs variables.

Ces règles ont été publiées précédemment sur papier en 1988 sous le même titre. Référence bibliographique : ISBN 2-550-19046-7 (dépot légal - quatrième trimestre 1988 - Bibliothèque nationale du Québec)

Elles ont donné lieu à une norme préliminaire canadienne rédigée par l'auteur du présent texte et publiée en 1992, confirmée comme norme nationale du Canada en 1996 (norme CAN-CSA Z243.4.1-1996). Cette dernière précise aussi les indices de classement attribués à un certain nombre de caractères spéciaux, et elle complète les tables de classement pour tous les autres caractères latins utilisés dans les langues européennes, en fonction d'un ordre qui reprend rigoureusement les règles énoncées ici. Ces règles ont été initialement conçues pour harmoniser l'ordre de classement des langues suivantes (sans en exclure d'autres) : français, anglais, allemand, portugais, italien et néerlandais. L'espagnol a un ordre légèrement différent et une adaptation des tables serait nécessaire dans ce cas, comme dans le cas des langues scandinaves ou d'autres langues européennes. La méthode resterait toutefois la même.

Pour complémenter cette méthode, les outils de programmation FRANCIS ont fait au Québec l'objet d'une commercialisation publique. Ces outils permettent d'une part d'accentuer automatiquement et correctement des données nominatives et toponymiques (noms connus au Québec ; adaptations requises pour d'autres environnements) comportant ou ne comportant pas déjà d'accents, du micro-ordinateur à l'ordinateur de grande puissance.

Ces outils permettent d'autre part de trier correctement les données alphabétiques, de les comparer, de déterminer des équivalences de recherche et plus généralement de les manipuler de différentes manières en fonction des propriétés multiples de la méthode décrite ici, en conformité avec la norme canadienne.

Au moment où ces règles sont rééditées pour le réseau Internet, l'auteur est à rédiger un projet de norme internationale de classement (projet ISO/CEI 14651) pour l'ensemble des caractères du jeu universel de caractères codés sur plusieurs octets (norme ISO/CEI 10646-1:1993, correspondant au standard UNICODE). Il s'agit d'une extension des concepts qui sont présentés ici.

Référence bibliographique de l'édition originale sur papier :

Gouvernement du Québec
Dépôt légal - 4^{e} trimestre 1988
Bibliothèque nationale du Québec
ISBN 2-550-19046-7