NORME INTERNATIONALE

ISO/IEC 14651

Quatrième édition
2016-02-15

Technologies de l'information — Classement international et comparaison de chaînes de caractères — Méthode de comparaison de chaînes de caractères et description du modèle commun et adaptable d'ordre de classement

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

Numéro de référence
ISO/IEC 14651:2016(F)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Sommaire

Page

Avant-propos

L'ISO (Organisation internationale de normalisation) et l'IEC (Commission électrotechnique internationale) forment le système spécialisé de la normalisation mondiale. Les organismes nationaux membres de l'ISO ou de l'IEC participent au développement de Normes internationales par l'intermédiaire des comités techniques créés par l'organisation concernée afin de s'occuper des domaines particuliers de l'activité technique. Les comités techniques de l'ISO et de l'IEC collaborent dans des domaines d'intérêt commun. D'autres organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO et l'IEC, participent également aux travaux. Dans le domaine des technologies de l'information, l'ISO et l'IEC ont créé un comité technique mixte, l'ISO/IEC JTC 1.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'attention est appelée sur le fait que certains des éléments du présent document peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO et la CEI ne sauraient être tenues pour responsables de ne pas avoir identifié de tels droits de propriété et averti de leur existence. Les détails concernant les références aux droits de propriété intellectuelle ou autres droits analogues identifiés lors de l'élaboration du document sont indiqués dans l'Introduction et/ou dans la liste des déclarations de brevets reçues par l'ISO (voir www.iso.org/brevets).

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'OMC concernant les obstacles techniques au commerce (OTC), voir le lien suivant: Avant-propos — Informations supplémentaires.

Le comité chargé de l'élaboration du présent document est l'ISO/CEI/JTC 1, *Technologies de l'information*, sous-comité SC 2, *Jeux de caractères codés*.

Cette quatrième édition annule et remplace la troisième édition (ISO/CEI 14651:2011), qui a fait l'objet d'une révision technique.

La présente version française de l'ISO 14651:2016 correspond à la version anglaise corrigée du 2016-05-01.

Introduction

La présente Norme internationale fournit une méthode universelle de mise en ordre des données textuelles. Elle fournit également une table-modèle commune qui, lorsque adaptée, peut satisfaire aux exigences de tri d'une langue donnée, tout en triant de manière raisonnable les autres écritures.

La table-modèle commune est conçue de telle sorte qu'une adaptation s'avère nécessaire pour chaque environnement local. C'est pourquoi la conformité à la présente Norme internationale requiert que les modifications à cette table commune, appelées «deltas», soient déclarées de manière à documenter les différences dans les résultats.

La présente Norme Internationale décrit une méthode pour classer l'information textuelle de manière indépendante du contexte.

L'ISO/CEI/TR 30112 contient des dispositions complémentaires pour le tri à celle de la présente Norme internationale; on y trouvera aussi des renseignements complémentaires sur les mots-clés de tri définis dans la présente Norme internationale.

Technologies de l'information — Classement international et comparaison de chaînes de caractères — Méthode de comparaison de chaînes de caractères et description du modèle commun et adaptable d'ordre de classement

1 Domaine d'application

La présente Norme internationale définit ce qui suit.

— Une méthode de référence pour la comparaison de deux chaînes de caractères ayant pour but de déterminer leur ordre de classement dans une liste triée. La méthode s'applique à des chaînes utilisant le répertoire complet de l'ISO/CEI 10646, des sous-répertoires tels que ceux des divers jeux normalisés ISO/CEI à 8 bits ou tout autre jeu de caractères, et permet de produire des résultats de tri valables (après adaptation) pour un ensemble de langues de chaque système d'écriture. Cette méthode de référence utilise des tables de tri dérivées soit de la table-modèle commune de classement définie dans la présente Norme internationale, soit d'une de ses adaptations. La méthode procure un format de référence de la table-modèle commune. Ce format est décrit en notation BNF (forme de Backus-Naur). Son emploi est normatif *dans* la présente Norme internationale.

— Une table-modèle commune de classement utilisée par la méthode de référence. Cette table décrit un ordre de base pour tous les caractères du standard Unicode 8.0 compris dans l'ISO/CEI 10646:2014 et son amendement 1. Tout cela permet de spécifier un ordre complètement déterministe. Cette table constitue le point de départ permettant de préciser un ordre de classement adapté aux règles de classement locales, sans qu'il soit nécessaire de connaître tous les systèmes d'écriture repris dans le jeu universel de caractères codés (JUC).

NOTE 1 Cette table-modèle commune de classement est destinée à être modifiée pour satisfaire aux besoins d'environnements locaux. L'avantage principal de cette pratique, sur le plan mondial, réside dans le fait que, pour d'autres systèmes d'écriture que celui de l'utilisateur, aucune modification n'est nécessaire et cet ordre demeurera aussi cohérent que possible et prévisible dans un contexte international.

NOTE 2 Le répertoire de caractères utilisé dans la présente Norme internationale est équivalent à celui du standard Unicode, version 6.0.

— Un nom de référence représentant cette version particulière de la table-modèle commune, à utiliser comme point de départ à toute adaptation. Ce nom implique notamment que la table est liée à un stade de développement particulier du jeu universel de caractères codés (ISO/CEI 10646).

— Des exigences pour la déclaration de différences (delta) entre une table de tri et la table-modèle commune.

La présente Norme internationale *ne* prescrit *pas* ce qui suit.

— Une méthode particulière de comparaison; toute méthode équivalente conduisant aux mêmes résultats est acceptable.

— Un format précis pour décrire ou pour adapter les tables dans une mise en œuvre donnée.

— Des symboles précis à utiliser par les mise en œuvre, sauf pour ce qui est du nom de la table-modèle commune de classement.

— Une interface utilisateur particulière destinée à choisir les options.

— Un format interne particulier pour les clés intermédiaires utilisées dans les comparaisons ou pour la table de tri. L'utilisation de clés numériques n'est pas prescrite non plus.

— Un ordre dépendant du contexte.

— Un prétraitement particulier des chaînes de caractères avant comparaison.

NOTE Bien que ceci ne soit pas prescrit par la présente Norme internationale, il s'avère souvent nécessaire de préparer les chaînes de caractères avant leur comparaison (cf. l'Annexe C).

Bien que l'on ne prescrive aucune interface utilisateur destinée à choisir les options ou à adapter la table-modèle commune, la conformité exige de toujours déclarer un delta, c'est-à-dire l'ensemble des différences par rapport à cette table. Il est fortement recommandé que l'application présente à l'utilisateur les options et adaptations disponibles.

2 Conformité

Un processus est conforme à la présente Norme internationale s'il produit des résultats identiques à ceux qui résultent de l'application des spécifications décrites en 6.2 à 6.5.

Toute déclaration de conformité à la présente Norme internationale doit être accompagnée, directement ou par référence, d'une déclaration de ce qui suit.

— Le nombre de niveaux de tri que le processus peut utiliser; ce nombre doit être égal ou supérieur à trois.

— Si le paramètre de traitement forward,position est permis.

— Si le paramètre de traitement backward est permis et à quel niveau.

— Le *delta* d'adaptation décrit en 6.4 et le nombre de niveaux définis dans ce delta.

— Si un processus de prétraitement est utilisé, la méthode utilisée doit être déclarée.

Il incombe au producteur de montrer en quoi sa déclaration de delta est reliée à la syntaxe de la table décrite en 6.3, et comment la méthode de comparaison utilisée, si elle est différente de celle mentionnée à l'Article 6, peut être considérée comme produisant les mêmes résultats que ceux prescrits par la méthode décrite à l'Article 6. L'usage d'un processus de prétraitement est optionnel et ses détails ne sont pas précisés dans la présente Norme internationale.

3 Références normatives

Les documents de référence suivants sont indispensables pour l'application du présent document. Pour les références datées, seule l'édition citée s'applique. Pour les références non datées, la dernière édition du document de référence s'applique (y compris les éventuels amendements).

ISO/CEI 10646:2014, *Technologies de l'information — Jeu universel de caractères codés (JUC)*

ISO/CEI 10646:1/Amd 1:2015, *Technologies de l'information — Jeu universel de caractères codés (JUC)/ Amendment 1.*

4 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

4.1
chaîne de caractères
suite de caractères considérée comme un objet simple

4.2
symbole de tri
symbole (4.12) utilisé pour préciser les poids attribués à un *élément de tri* (4.4)

4.3
table (de poids) de tri
table reliant les *éléments de tri* (4.4) aux *éléments de poids* (4.14)

4.4
élément de tri
suite constituée d'un ou de plusieurs caractères considérés comme une seule entité aux fins du *tri* (4.7)

4.5
delta
liste des différences que présente une *table de classement* (4.3) donnée par rapport à une autre

Note 1 à l'article: Une table de tri donnée associée à un delta donné forme une nouvelle table de tri.

Note 2 à l'article: Sauf mention contraire, le terme «delta» désignera les différences par rapport à la table-modèle commune définie dans la présente Norme internationale.

4.6
niveau (de tri)
numéro d'une *sous-clé* (4.11) dans la série de sous-clés formant une clé

4.7
tri
procédé par lequel on détermine si, de deux chaînes, la première est plus petite, égale ou plus grande que la seconde

4.8
clé de tri
série de *sous-clés* (4.11) utilisée pour déterminer un ordre

4.9
prétraitement
procédé par lequel des *chaînes de caractères* (4.1) sont transformées en d'autres chaînes avant le calcul de la *clé de tri* (4.8) de chaque chaîne

4.10
méthode de comparaison de référence
méthode de détermination de l'ordre relatif de deux *clés de tri* (4.8)

Note 1 à l'article: Voir l'Article 6.

4.11
sous-clé
suite de poids calculée pour une *chaîne de caractères* (4.1)

4.12
symbole
élément de tri (4.4)

4.13
poids (de tri)
entier positif, utilisé dans les *sous-clés* (4.11), pour indiquer l'ordre relatif des éléments de tri

4.14
élément de poids
liste d'un certain nombre de poids séquentiellement ordonnés par niveau

5 Symboles et abréviations

Selon l'ISO/CEI 10646, les caractères se représentent à l'aide de U*X*, où *X* correspond à une série d'un à huit chiffres hexadécimaux (où toutes les lettres de la série de chiffres hexadécimaux sont en

majuscules) et où *X* est le numéro du caractère dans l'ISO/CEI 10646. Cette convention est reprise dans la présente Norme internationale.

Dans la table-modèle commune, des symboles arbitraires représentent des poids selon la notation BNF décrite en 6.3.1.

6 Comparaison de chaînes

6.1 Prétraitement des chaînes de caractères avant comparaison

Il peut s'avérer nécessaire de transformer les chaînes de caractères avant de leur appliquer la méthode de comparaison de référence (l'Annexe C fournit un exemple d'une telle préparation). Bien que n'étant pas l'objet de la présente Norme internationale, le prétraitement peut être une partie importante du processus de tri. On consultera l'Annexe C pour des exemples de prétraitement.

S'il y a lieu, une partie importante de la phase préparatoire consiste à transformer les caractères d'un codage non-JUC à des caractères du JUC fournis en entrée à la méthode de comparaison. Cette tâche peut comprendre notamment le traitement correct de séquences d'échappement dans le codage original, la transformation de caractères sans attribution dans le JUC à des positions de code dans la zone privée et la transposition de caractères dans le cas de chaînes qui ne seraient pas stockées en ordre logique. Par exemple, dans le cas de codages arabes en ordre visuel, les caractères doivent être mis en ordre logique; dans le cas de certains codages à usage bibliographique, les accents combinatoires stockés avant leur caractère de base doivent être déplacés après le caractère de base. La suite résultante peut devoir être re-transformée dans le codage original.

La table-modèle commune est conçue de telle sorte que les séquences combinatoires et les caractères simples (précomposés) correspondants aient exactement le même ordre. Pour éviter de violer par mégarde cet invariant (et au passage la conformité à Unicode), l'adaptation devrait changer le classement des séquences combinatoires quand le classement des caractères précomposés correspondant est changé. Par exemple, si Ä est déplacé après Z, alors le classement de la séquence combinatoire <A>+<tréma combinatoire> devrait aussi être changé. Pour éviter de révéler des différences de codage invisible à l'utilisateur, on recommande de normaliser les chaînes selon la forme FND de l'algorithme de normalisation Unicode – voir le Unicode Technical Report n° 15 dans la bibliographie.

Les séquences d'échappement et les caractères de commande sont très délicats à interpréter; il est fortement recommandé de les filtrer ou de les transformer.

NOTE Puisque la méthode de comparaison de référence est une description logique du procédé de comparaison de chaînes, rien n'empêche une mise en œuvre de cette méthode d'utiliser exclusivement un codage autre qu'un codage du JUC, pour autant que les résultats obtenus soient les mêmes que si la méthode de référence était utilisée.

6.2 Construction des clés et comparaison

6.2.1 Préliminaires

6.2.1.1 Hypothèses

La table de tri est une transformation des éléments de tri en éléments de poids. Pour chaque élément de poids, la table-modèle commune décrit quatre niveaux. L'adaptation peut augmenter ou réduire ce nombre de niveaux, mais pas à moins de trois.

NOTE Dans la table-modèle commune, les niveaux ont généralement les significations suivantes, bien que cet usage ne soit pas absolu:

Niveau 1: ce niveau correspond généralement au jeu de lettres de base pour une écriture alphabétique, au jeu de caractères courants pour une écriture idéographique ou syllabique.

Niveau 2: ce niveau correspond généralement aux diacritiques pouvant accompagner les caractères de base de chaque écriture. En certaines langues, les lettres accentuées sont considérées comme des lettres de base de l'alphabet et ne sont pas affectées par ce niveau, mais seulement par le premier niveau. En espagnol par exemple, le N TILDE est considéré comme une lettre de base de l'alphabet latin; par conséquent, une adaptation pour l'espagnol changera la définition de N TILDE de «le poids d'un N au premier niveau et le poids d'un TILDE au second niveau» à «le poids d'un N TILDE (entre N et O) au premier niveau et une indication de l'absence de diacritique au second niveau». Pour certains caractères, on prend également en compte des variantes de forme au second niveau, par exemple ß (la LETTRE MINUSCULE LATINE S DUR), qui est traitée comme un équivalent de ss au premier niveau mais s'en distingue traditionnellement au second niveau.

Niveau 3: ce niveau est généralement associé aux distinctions de casse (majuscules-minuscules) ou aux variantes de formes (comme la distinction entre hiragana et katakana).

Niveau 4: ce niveau est généralement consacré aux distinctions pondérales plus fines que celles des autres niveaux. Le dernier niveau (le quatrième dans la table-modèle commune) est souvent utilisé pour donner des poids additionnels à des caractères «spéciaux», c'est à dire des caractères qui ne sont pas normalement utilisés dans l'orthographe des mots d'une langue (ponctuation, vignettes, etc.), souvent appelés «ignorables» dans le contexte du tri informatique.

6.2.1.2 Propriétés de traitement

Une table de tri adaptée donnée possède des propriétés spécifiques de balayage et de classement. Ces propriétés peuvent avoir été changées par l'adaptation.

Une direction de balayage (vers l'avant ou vers l'arrière) pour chaque niveau est utilisée pour indiquer comment traiter la chaîne. La direction de balayage est une propriété globale de chaque niveau défini dans la table adaptée.

Si le dernier niveau est supérieur à trois, il existe une propriété optionnelle de ce niveau appelée l'option «position»: lorsque active, une comparaison des positions numériques de chaque caractère «ignorable» dans les deux chaînes est effectuée, avant de comparer leurs poids. En d'autres mots, si deux chaînes sont équivalentes à tous les niveaux sauf le dernier, la chaîne contenant un caractère ignorable en position la plus basse est classée avant l'autre. Si les caractères ignorables ont les mêmes positions, alors leurs poids sont considérés jusqu'à ce qu'une différence soit trouvée. Le traitement correct de cette propriété optionnelle n'est pas nécessaire à la conformité à la présente Norme internationale.

NOTE La direction de balayage (vers l'avant ou vers l'arrière) n'est normalement pas reliée à la direction naturelle d'écriture. La direction de balayage s'applique à la suite logique de la chaîne de caractères codés.

Dans le cas d'écritures de droite à gauche comme l'arabe, l'ISO/CEI 10646 prescrit que les premiers caractères en ordre logique sont ceux apparaissant à droite en ordre de présentation. En écriture latine au contraire, les premiers caractères en ordre logique apparaissent à gauche en ordre de présentation.

Le balayage vers l'avant commence au début de la séquence en ordre logique, alors que le balayage vers l'arrière commence à la fin, sans égard à la direction de présentation. La direction de balayage pour fins de tri est une propriété globale de chaque niveau décrit dans la table.

Dans l'ISO/CEI 10646, l'écriture arabe est artificiellement séparée en deux pseudo-écritures: 1) l'écriture arabe logique, intrinsèque, codée indépendamment des formes contextuelles et 2) les formes de présentations arabes. Les deux permettent le codage complet de l'arabe, mais le codage intrinsèque est normalement privilégié pour sa meilleure capacité de traitement, alors que certaines applications de présentation préfèrent les formes de présentation. L'ISO/CEI 10646 ne prescrit pas l'ordre de stockage des formes de présentation; dans certaines réalisations, elles sont stockées en ordre inverse de celui utilisé pour le codage intrinsèque. Par conséquent, la phase de préparation devrait assurer que les formes de présentation arabes et les autres caractères arabes soient fournis en ordre logique à la méthode de comparaison.

Une table de tri adaptée peut être séparée en sections pour faciliter l'adaptation. On donne alors à chaque section un nom, conformément aux dispositions de 6.3.1. Une des possibilités d'adaptation est

de donner un certain ordre à chaque section et de changer l'ordre relatif d'une section par rapport à d'autres.

6.2.2 Méthode de référence de construction des clés

Lorsque deux chaînes doivent être comparées pour déterminer leur ordre relatif, elles sont d'abord analysées en séquences d'éléments de tri, en tenant compte des déclarations «collating-element» à caractères multiples présents dans la table de tri (*si* la syntaxe de 6.3.1 est utilisée). Dans la syntaxe utilisée pour exprimer la table-modèle commune, le nom d'un élément de tri associé à un seul caractère est formé de la lettre «U» suivie du numéro du caractère dans le JUC, en notation hexadécimale. Les noms et caractères associés aux éléments de tri multi-caractères sont définis par les déclarations d'éléments de tri.

NOTE Les éléments de tri comportant plus de caractères sont préférés à ceux qui sont plus courts. Par exemple, si un élément de tri comportant plusieurs caractères est défini pour «abc» et qu'un autre est defini pour «ab» ou qu'un autre l'est pour «bc», alors, si «abc» se présente, l'élément de tri pour «abc» s'appliquera et non celui pour «ab» ou «bc».

Une suite de *m* sous-clés intermédiaires est alors formée de chaque chaîne, *m* étant le nombre de niveaux décrits dans une table de poids de tri adaptée.

Chaque clé de tri est une suite de sous-clés. Chaque sous-clé est une liste de poids numériques. Une sous-clé est construite en ajoutant successivement la liste des poids attribués à chaque élément de tri de la chaîne au niveau de la sous-clé en construction. Dans la table-modèle commune, le mot-clé «`IGNORE`» trouvé en place d'une suite de poids à un niveau indique que la suite de poids à ce niveau pour cet élément de tri est vide.

Il y a trois façons de former des sous-clés: vers l'avant (paramètre de traitement «`forward`»), vers l'arrière (paramètre de traitement «`backward`») et de façon positionnelle (paramètre de traitement «`forward,position`»). Les sous-clés formées de façon positionnelle ne peuvent apparaître qu'au dernier niveau et seulement si ce niveau est supérieur à trois. La conformité n'exige pas la formation de sous-clés de façon positionnelle; une réalisation incapable de formation positionnelle interprétera «`forward,position`» comme s'il s'agissait de «`forward`».

Si la table adaptée ne contient pas d'entrée pour un caractère de la chaîne d'entrée, les poids de ce caractère ne sont pas définis. Les caractères de poids indéfinis devraient être triés, par rapport aux caractères ayant des poids définis, comme s'ils avaient le poids nommé «UNDEFINED» au premier niveau. Si le symbole «UNDEFINED» n'a pas de poids attribué avant l'attribution d'un poids à symbole <SFFFF> dans la table adaptée, on considèrera que le poids de «UNDEFINED» est tout juste inférieur à celui de <SFFFF>. Le classement relatif des caractères de poids indéfinis entre eux n'est pas précisé par la présente Norme internationale.

NOTE 1 Une manière possible de classer les caractères de poids indéfinis entre eux est de supposer l'existence de lignes d'adaptation comme celles qui suivent, rangées en ordre de numéro de caractères dans le JUC (<PLAIN> représente ici le poids maximal de niveau 4):

<UXXXX> «<UNDEFINED><UXXXX>»;<BASE>;<MIN>;<PLAIN>

NOTE 2 <SFFFF> est le plus grand poids de premier niveau dans table-modèle commune.

6.2.2.1 Formation de sous-clé vers l'avant (paramètre de traitement «`forward`»)

En présence du paramètre de traitement «`forward`» à un niveau donné, on construit la sous-clé de la façon suivante:

On balaie vers l'avant un à un les éléments de tri de la chaîne de caractères d'entrée pour leur attribuer un poids. On obtient les poids en recherchant les éléments de tri dans la table de poids de tri adaptée donnée et en extrayant la liste de poids pour le niveau considéré. Cette liste de poids s'ajoute à la fin de la sous-clé.

6.2.2.2 Formation de sous-clé vers l'arrière (paramètre de traitement «`backward`»)

En présence du paramètre de traitement «`backward`» à un niveau donné, on construit la sous-clé vers l'avant et on la renverse, poids par poids.

6.2.2.3 Formation de sous-clé positionnelle (paramètre de traitement «`forward,position`»)

En présence du paramètre de traitement «`forward,position`» au dernier niveau, on construit la sous-clé de la même façon que vers l'avant, si ce n'est que les poids des éléments de tri qui sont pris en compte à tous les niveaux sauf le dernier sont remplacés par un seul poids (appelé <PLAIN> ici) supérieur à tous les poids du dernier niveau dans la table adaptée donnée. Les éléments de tri qui sont ignorés à tous les niveaux sauf le dernier conservent leurs poids tels que donnés dans la table adaptée donnée. Enfin, toute séquence de queue de la valeur maximale (<PLAIN>) est retirée de la sous-clé (ce qui en pratique remplace chaque <PLAIN> par un poids nul).

NOTE Il est permis, à chaque niveau, d'appliquer une réduction de toutes les sous-clés de ce niveau, en autant que cette réduction préserve l'ordre. Une telle réduction est utile pour les niveaux 2, 3 et 4. Les sous-clés de niveau 2 contiennent souvent de longues suites du poids appelé <BASE> dans la table donnée à l'Annexe A. Les sous-clés de niveau 3 contiennent souvent de longues suites du poids appelé <MIN> à l'Annexe A. Les sous-clés de niveau 4 contiennent souvent de longues suites du poids appelé <PLAIN> ici. Une telle technique de réduction préservant l'ordre consiste à coder, dans la sous-clé de dernier niveau, la position relative de chaque caractère autrement ignoré; c'est là l'origine du nom de l'option «*`position`*».

6.2.3 Méthode de comparaison de référence pour le tri des chaînes de caractères

La méthode de comparaison de référence pour le classement de deux chaînes de caractères (*après* le prétraitement, qui ne fait pas partie de cette méthode de comparaison) consiste à comparer les clés de tri construites selon la méthode de référence décrite en 6.2.2 de la présente Norme internationale:

— En utilisant une table de poids de tri adaptée donnée, construire une clé de tri pour chacune des chaînes à comparer.

— Comparer ensuite les clés selon la définition de l'ordre des clés donnée ci-dessous dans cet article. Les clés peuvent être comparées jusqu'à un niveau donné ou jusqu'au dernier niveau de la table de poids de tri adaptée donnée.

NOTE 1 La comparaison peut être effectuée *pendant* la construction des clés, en arrêtant cette construction dès que l'ordre des chaînes peut être déterminé. Cette technique est parfois appelée *évaluation paresseuse* et certains systèmes l'utilisent implicitement. Elle permet d'éviter la construction complète des clés quand une différence peut être trouvée tôt pendant la construction. Quand un ensemble important de chaînes doit être trié, il est recommandé de construire et de stocker les clés – ou tout au moins un segment initial – avant de les comparer.

Les poids associés à des niveaux différents ne doivent pas être comparés, on ne doit donc pas comparer les sous-clés de différents niveaux. Les clés construites à partir de tables adaptées différentes ne doivent pas être comparées.

NOTE 2 Ceci permet aux mises en œuvre d'attribuer les poids à chaque niveau indépendamment des autres niveaux et sans égard à d'autres tables adaptées.

m est le plus grand niveau d'une table adaptée donnée. Rappelons qu'une clé est une liste, de longueur m, de sous-clés; une sous-clé est une liste de poids; un poids est un entier positif. D'autres notations utilisées ci-dessous sont:

— L_z est la longueur de la sous-clé z, c'est-à-dire le nombre de poids dans cette sous-clé.

— $z_{\mathbf{pd}(a)}$, où $1 \le a \le L_z$, est le poids à la position a (un entier > 0) de la sous-clé z.

— $u_{\mathbf{sc}(b)}$, où $1 \le b \le m$, est la sous-clé de niveau b (un entier > 0) de la clé u.

Les ordres des poids, des sous-clés et des clés de tri (jusqu'à un certain niveau ou jusqu'au dernier niveau) sont des relations d'ordre total, définies pour une table de tri adaptée donnée comme suit:

a) Les poids sont des valeurs entières positives (dans la méthode de référence) et sont comparés comme tels aux fins du classement.

b) Une sous-clé v est *plus petite* qu'une sous-clé w (on notera $v < w$) **si et seulement s'il** existe un entier i, où $1 \le i \le L_v+1$ et $i \le L_w$, tel que

   — $i = 1$ et $v_{\mathbf{pd}(i)} < w_{\mathbf{pd}(i)}$, ou

   — pour tous les entiers j, $1 \le j < i$, l'égalité $v_{\mathbf{pd}(j)} = w_{\mathbf{pd}(j)}$ est maintenue, et soit

      — $i \le L_v$ et $v_{\mathbf{pd}(i)} < w_{\mathbf{pd}(i)}$, soit

      — $i = L_v+1$ et $0 < w_{\mathbf{pd}(i)}$.

Une sous-clé v est *plus grande* qu'une sous-clé w (on notera $v > w$) **si et seulement si** w est plus petite que v. Une sous-clé v est *égale* à une sous-clé w (on notera $v = w$) **si et seulement si** v n'est pas plus petite que w et w n'est pas plus petite que v.

c) Une clé de tri x est *plus petite* qu'une clé de tri y au niveau s (on notera $x <_s y$) **si et seulement s'il** existe un entier i, où $1 \le i \le s$ et $i \le m$, tel que

   — $i = 1$ et $x_{\mathbf{sc}(i)} < y_{\mathbf{sc}(i)}$, ou

   — pour tous les entiers j, $1 \le j < i$, l'égalité $x_{\mathbf{sc}(j)} = y_{\mathbf{sc}(j)}$ est maintenue, et $x_{\mathbf{sc}(i)} < y_{\mathbf{sc}(i)}$.

Une clé de tri x est *plus grande* qu'une clé de tri y au niveau s (on notera $x >_s y$) **si et seulement si** y est plus petite que x au niveau s. Une clé de tri x est *égale* à une clé de tri y au niveau s (on notera $x =_s y$) **si et seulement si** x n'est pas plus petite que y au niveau s et y n'est pas plus petite que x au niveau s.

d) Pour les clés de tri, $<$, $>$ et $=$ sont définis comme $<_m$, $>_m$ et $=_m$ respectivement.

NOTE 3 Pour les clés de tri, $x <_t y$ implique que $x <_{t+1} y$, $x >_t y$ implique que $x >_{t+1} y$, $x =_t y$ implique que $x =_{t-1} y$, $x <_0 y$ est faux, $x >_0 y$ est faux et $x =_0 y$ est vrai. Au-delà du niveau m, pour une table adaptée donnée, il n'y a plus de distinctions d'ordre. On notera que cette définition implique que si une clé est «plus petite» qu'une autre au niveau 1, elle est aussi «plus petite» aux niveaux 2, 3, 4, etc. En général, lorsqu'une clé est plus petite qu'une autre à un certain niveau, elle l'est aussi à tous les niveaux subséquents. *A contrario*, lorsque deux clés sont égales à un certain niveau, elles sont aussi égales à tous les niveaux inférieurs.

6.3 Table-modèle commune: composition et interprétation

Ce paragraphe précise:

— La syntaxe utilisée par la table-modèle commune donnée à l'Annexe A ou par une table adaptée basée sur la table-modèle commune telle que donnée à l'Annexe A.

— Les contraintes de forme d'une table utilisant cette syntaxe.

— L'interprétation à donner aux énoncés d'adaptation dans les deltas pour des tables utilisant cette syntaxe.

— L'évaluation des symboles en poids dans les tables adaptées utilisant cette syntaxe.

— Les conditions d'équivalence de deux tables.

— Les conditions d'équivalence des résultats de comparaison.

6.3.1 Règles de syntaxe BNF pour la table-modèle commune de l'Annexe A

Les définitions <entre crochets> utilisent des termes qui ne sont pas définis dans cette syntaxe BNF et suivent l'usage général en français.

Autres conventions:

* * indique une répétition (0 fois ou plus) d'un atome ou d'un groupe d'atomes;
* + indique une répétition (1 fois ou plus) d'un atome ou d'un groupe d'atomes;
* ? indique l'apparition optionnelle (0 ou 1 fois) d'un atome ou d'un groupe d'atomes;
* des parenthèses servent à grouper des atomes;
* les productions se terminent par un point-virgule.

Définition des tables de tri comme des suites de lignes:

```
table_de_poids =  table-modèle_commune | table_adaptée;
table-modèle_commune =
                  ligne_simple+;
table_adaptée =       ligne_de_table+;
```

Définition des types de ligne:

```
ligne_simple =    (définition_de_symbole | élément_de_tri |
                  attribution_de_poids | fin_ordre)? achèvement_de_ligne;
ligne_de_table =     ligne_simple | ligne_adaptation;  ligne_adaptation
=  (reclasser_après | début_ordre | fin_de_reclassement |
                  définition_de_section | reclasser_section_après)
                  achèvement_de_ligne;
```

Définition de la syntaxe de base pour les poids de tri:

```
définition_de_symbole =
                  'collating-symbol' espace+ élément_symbole;
élément_symbole =  symbole | symbole_intervalle;
symbole_intervalle =  symbole '..' symbole;
symbole =          symbole_simple | symbole_juc;
symbole_juc =      ('<U'  chaîne_de_un_à_huit_hexa '>') |
                   ('<U-' chaîne_de_un_à_huit_hexa '>');
symbole_simple =   '<' identifiant '>';
élément_de_tri =
                   'collating-element' espace+ symbole espace+
                   'from' espace+ suite_de_symboles_cités;
suite_de_symboles_cités =
                   '«' poids_simple+ '»';
attribution_de_poids =
poids_simple | poids_symbolique;
poids_simple =     élément_symbole | 'UNDEFINED';
poids_symbolique =
élément_symbole espace+ liste_de_poids;
liste_de_poids =   atome_de_niveau (point-virgule atome_de_niveau)*;
atome_de_niveau =  groupe_de_symboles | 'IGNORE';
groupe_de_symboles = élément_symbole | suite_de_symboles_citée;
fin_ordre =     'order_end';
```

Définition de la syntaxe d'adaptation:

```
reclasser_après = 'reorder-after' espace+ symbole_cible;
symbole_cible =   symbole;
début_ordre =     'order_start' espace+ direction_multi_niveaux;
direction_multi_niveaux =
                  (direction point-virgule)* direction (',position')?;
direction =          'forward' | 'backward';
fin_de_reclassement = 'reorder-end';
définition_de_section =
                      définition_de_section_simple |
                      définition_de_section_liste;
définition_de_section_simple =
                     'section' espace+ identifiant_de_section;
identifiant_de_section =
                     identifiant;
définition_de_section_liste =
                     'section' espace+ identifiant_de_section espace+ liste_de_
symboles;
liste_de_symboles =
élément_symbole (point-virgule élément_symbole)*;
```