# TD 1 : Codes et Codage de caractères

## 1 Codage de caractères

**Exercice 1** Convertissez
- les nombres $(17)_{10}$, $(42)_{10}$, $(555)_{10}$ en base 16 et 2
- les nombres $(3A)_{16}$ et $(DEC)_{16}$ en base 10 et 2

BEGIN SOLUTION

- les nombres $(17)_{10} = (11)_{16}$, $(42)_{10} = (2A)_{16}$, $(555)_{10} = (22B)_{16}$ en base 16 et 2
- les nombres $(3A)_{16} = 58$ et $(DEC)_{16} = 3564$

END SOLUTION

**Exercice 2** Quelle partie de l'espace de code est utilisée par UTF-32 ?

BEGIN SOLUTION

L'espace de alphabet représentable en Unicode est U+0 ...U+$(10FFFF)_{16}$, il y a donc $(11.0000)_{16}$ caractères dans l'alphabet. Un mot 32 bit permet de représenter $(1.0000.0000)_{16}$ valeurs. La fraction utilisée est donc $(\frac{11}{10000})_{16} \approx (\frac{1}{3855})_{10}$.

END SOLUTION

**Exercice 3** Quelle est la taille (en octets) d'un texte avec $n$ caractères ASCII codé en format

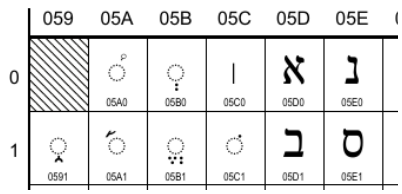
1. UTF-8
2. UTF-16
3. UTF-32

BEGIN SOLUTION

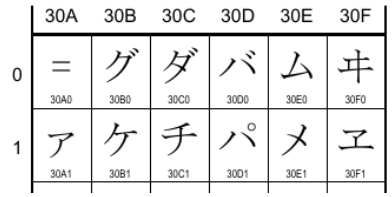
1. $n$
2. $2n$
3. $4n$

END SOLUTION

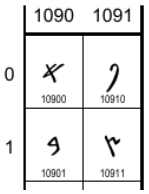
**Exercice 4** Voici des extraits de la table de codage Unicode pour l'alphabet hébreu, japonais et phénicien.




(a) Hebrew




(b) Katakana




(c) Phoenician

Codez les caractères Alef, Bet et Nun (05D0, 05D1, 05E0), les caractères Gu et We (30B0, 30F1) et Alf, Bet (10900, 10901) en UTF-8, UTF-16, UTF-32.

BEGIN SOLUTION

- `unicode 0x5D0` : U+05D0 HEBREW LETTER ALEF, UTF-8 : d7 90 UTF-16BE : 05d0
- `unicode 0x5D1` : U+05D1 HEBREW LETTER BET, UTF-8 : d7 91 UTF-16BE : 05d1
- `unicode 0x5E0` : U+05E0 HEBREW LETTER NUN, UTF-8 : d7 a0 UTF-16BE : 05e0
- `unicode 0x30B0` : U+30B0 KATAKANA LETTER GU, UTF-8 : e3 82 b0 UTF-16BE : 30b0
- `unicode 0x30F1` : U+30F1 KATAKANA LETTER WE, UTF-8 : e3 83 b1 UTF-16BE : 30f1
- `unicode 0x10900` : U+10900 PHOENICIAN LETTER ALF, UTF-8 : f0 90 a4 80 UTF-16BE : d802dd00
- `unicode 0x10901` : U+10901 PHOENICIAN LETTER BET, UTF-8 : f0 90 a4 81 UTF-16BE : d802dd01

END SOLUTION

## 2 Codes et codages

**Exercice 5** Cochez les cases où $m_1 \preceq m_2$ :

| $m_1$ / $m_2$ | 01 | 010 | 110 |
|---|---|---|---|
| 01 | | | |
| 101 | | | |
| 10 | | | |
| 11 | | | |

**Exercice 6** On dit qu'une relation $R$ est un ordre si elle est

- réflexive : $\forall x.R(x,x)$
- antisymétrique : $\forall xy.R(x,y) \wedge R(y,x) \rightarrow x = y$
- transitive : $\forall xy.R(x,y) \wedge R(y,z) \rightarrow R(x,z)$

Pour la relation $\preceq$, on utilise une notation *infixe*, c.à.d. on écrit $x \preceq y$ au lieu de $\preceq (x,y)$.
Montrez que la relation $\preceq$ définie par $m \preceq m' =_{def} \exists r.m' = m \cdot r$ est un ordre.

**Exercice 7** On définit trois codes $c_1, c_2, c_3$ pour un alphabet $A = \{a,b,c,d\}$ selon le tableau suivant :

| $x$ | $c_1(x)$ | $c_2(x)$ | $c_3(x)$ |
|---|---|---|---|
| a | 0 | 10 | 0 |
| b | 010 | 00 | 10 |
| c | 01 | 11 | 110 |
| d | 10 | 110 | 111 |

Montrez que :

- $c_1$ est injectif, mais son extension homomorphe $c_1^*$ n'est pas unique
- $c_2$ n'est pas un code préfixe, mais son extension homomorphe $c_2^*$ est unique
- $c_3$ est un code préfixe

BEGIN SOLUTION

- Non-unicité de $c_1$ : $c_1(ca) = c_1(b)$
- Unicité de $c_2$ : Uniquement $c_2(c)$ et $c_2(d)$ n'ont pas la propriété préfixe. Pour toute chaîne 110..., on a besoin d'un lookahead arbitrairement long. Si 11 est suivi d'un nombre pair de 0, on a la séquence $cb$.... Si 11 est suivi d'un nombre impair de 0, on a la séquence $db$....

END SOLUTION

**Exercice 8** Montrez formellement que tout codage unique est un codage injectif.
*Remarque :* L'exercice 7 montre que cette inclusion est stricte.

BEGIN SOLUTION

Soit $c$ un codage unique. Il existe donc un $d$ tel que pour tout message $m$, $d(c(m)) = m$. Soient $m_1, m_2$ messages avec $m_1 \neq m_2$. Supposons $c(m_1) = c(m_2)$. Par unicité de $c$, on a $m_1 = d(c(m_1)) = d(c(m_2)) = m_2$, une contradiction.

END SOLUTION

**Exercice 9** Montrez formellement que tout codage $c^*$ qui est l'extension homomorphe d'un code préfixe $c$ est injectif.

BEGIN SOLUTION

Soit $c^*$ un code préfixe homomorphe. On montre qu'il est injectif : Si $c^*(m_1) = c^*(m_2)$, alors $m_1 = m_2$. Par induction sur la taille (= nombre de caractères) de $m_1$.

— Si $|m_1| = 0$, donc $m_1 = []$, alors $c^*(m_1) = c^*(m_2) = []$, donc forcément $m_2 = []$.

— Soit $|m_1| = n+1$, donc $m_1 = a \cdot m_1'$. On voit que $m_2$ ne peut pas être vide. Il est donc de la forme $b \cdot m_2'$. Alors, $c^*(m_1) = c^*(a \cdot m_1') = c(a) \cdot c^*(m_1') = c^*(m_2) = c(b) \cdot c^*(m_2')$. On voit $c(a) \preceq c(b)$, donc $a = b$ à cause de $c$ code préfixe, donc $c(a) = c(b)$, donc $c(m_1') = c(m_2')$, donc $m_1' = m_2'$ par hypothèse d'induction.

END SOLUTION

**Exercice 10** Est-ce que le code Morse est unique / injectif / un code préfixe ?

BEGIN SOLUTION

Réponse : il n'est pas injectif (par example code(ATT) = code(J)) et par conséquent pas unique.

END SOLUTION

**Exercice 11** Appliquez l'algorithme `arbre_dec` aux codages $c_1$ et $c_2$ de la table suivante.

| $x$ | $c_1(x)$ | $c_2(x)$ | $c_3(x)$ |
|---|---|---|---|
| a | 00 | 01 | 0 |
| b | 01 | 11 | 10 |
| c | 10 | 00 | 110 |
| d | 11 | 001 | 1110 |

Si la construction de l'arbre échoue, identifiez les causes. Est-ce que vous pouvez proposer des codages qui évitent le problème ?

BEGIN SOLUTION

1. $c_2$ est "l'inverse" du code $c_2$ de l'exercice 7, donc pas un code préfixe. Irréparable.
2. $c_3$ ne permet pas la construction d'un arbre binaire complet (voir exercice 14). On peut trouver le code plus court `0, 10, 110, 111`

END SOLUTION

**Exercice 12** Appliquez l'algorithme `tab_cod` aux arbres de décodage obtenus dans l'exercice 11 et vérifiez que vous obtenez bien les tables d'origine.

**Exercice 13** Pourquoi est-ce que l'algorithme `tab_cod` termine ?

BEGIN SOLUTION

Récursion structurelle sur l'arbre.

END SOLUTION

**Exercice 14** ***(Devoir maison)***

Analyse de l'algorithme `arbre_dec` :

1. Quels problèmes se poseraient pour un algorithme de décodage si l'arbre n'était pas un arbre binaire (mais si un noeud intérieur pouvait avoir un seul successeur) ?
2. Un invariant de `arbre_dec` est qu'il prend la représentation d'une table `tab` d'un codage préfixe. Démontrez que cet invariant est maintenu par les appels récursifs, donc, que

$$\{(c, m) | (c, 0 \cdot m) \in tab\}$$

   représente bien une table d'un codage préfixe (et pareil pour $1 \cdot m$).
3. Démontrez que si $(c, []) \in tab$, alors il n'est pas possible d'avoir un $(d, m) \in tab$, pour un $c \neq d$.
4. Démontrez que l'algorithme termine.

BEGIN SOLUTION

1. Arbre binaire incomplet : certains codes ne seraient pas un codage correct d'un code source, par exemple `1111`. Mais c'est aussi un problème pour des arbres complets, si le message est tronqué.
2. Supposons qu'il existent $m_1, m_2$ avec $(c, 0 \cdot m_1) \in tab$ et $(c, 0 \cdot m_2) \in tab$ et $m_1 \preceq m_2$. Alors, aussi $0 \cdot m_1 \preceq 0 \cdot m_2$. Contradiction avec codage préfixe de $tab$.
3. Si $(c, []) \in tab$ et $(d, m) \in tab$, puisque $[] \preceq m$, on a une contradiction avec codage préfixe de $tab$.
4. Terminaison : dans chaque appel récursiv, pour chaque $(c, m)$, la taille de $m$ décroit, jusqu'à ce que $tab$ contient uniquement des éléments avec $m = []$. Selon (3), il peut seulement y avoir un seul $(c, []) \in tab$, on termine donc avec la clause (2).

END SOLUTION

**Exercice 15**

1. Utilisez l'inégalité de Kraft pour déterminer s'il est possible de construire un code préfixe pour les caractères $a \ldots d$ avec les longueurs de code suivants :

| caractère | a | b | c | d |
|---|---|---|---|---|
| longueur | 1 | 2 | 2 | 3 |

2. Quelle serait votre réponse si on admet un code de longueur 3 pour le caractère $b$ ? Proposez effectivement un code.

BEGIN SOLUTION

1. $2^{-1} + 2 * 2^{-2} + 2^{-3} = \frac{9}{8} > 1$, un code n'est pas possible.
2. $2^{-1} + 2^{-2} + 2 * 2^{-3} = 1$, un code est possible, par exemple $a \mapsto 0, c \mapsto 10, b \mapsto 110, d \mapsto 111$

END SOLUTION

**Exercice 16**

1. Une entreprise veut installer un système téléphonique interne où les 5 membres du directoire ont un numéro à un seul chiffre (de 0 à 9) et les 80 autres employés un nombre à deux chiffres. Est-ce possible ?
2. Serait-il possible d'avoir des numéros à deux chiffres pour les membres du directoire et de trois chiffres pour les autres employés ? Faites une proposition concrète.

*A noter :* L'inégalité de Kraft se généralise d'un code binaire à un code $n$-aire (alphabet à $n$ chiffres) comme suit : Il existe un code préfixe $n$-aire avec $k$ codes $u_1 \dots u_k$ si et seulement si

$$\sum_{i=1}^{k} n^{-|u_i|} \leq 1$$

où $|u_i|$ est la longeur du code $u_i$.

BEGIN SOLUTION

1. $5 * 10^{-1} + 80 * 10^{-2} = \frac{50}{100} + \frac{80}{100} = 1.3 > 1$, un tel système ne peut donc pas exister.
2. $5 * 10^{-2} + 80 * 10^{-3} \leq 1$. On peut attribuer des codes `01` ...`05` au directeurs et `101` ...`180` aux autres employés.

END SOLUTION