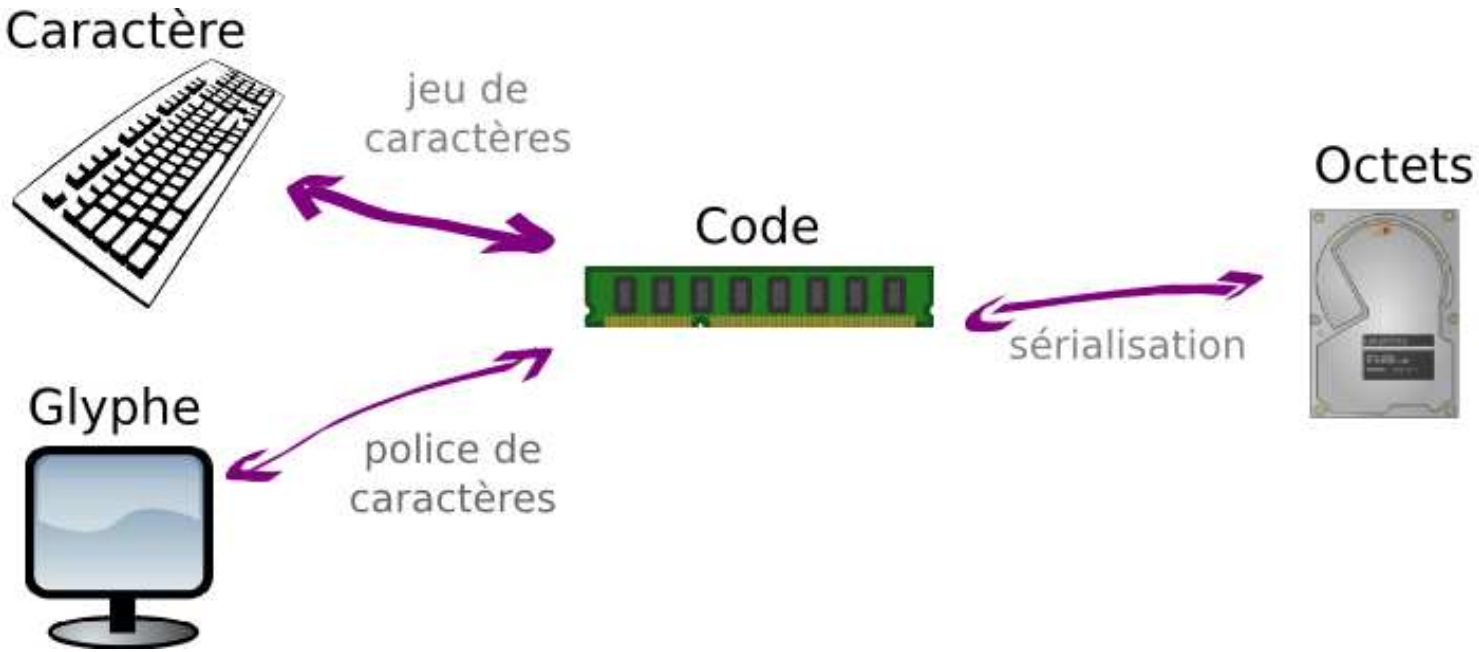
Python et Unicode

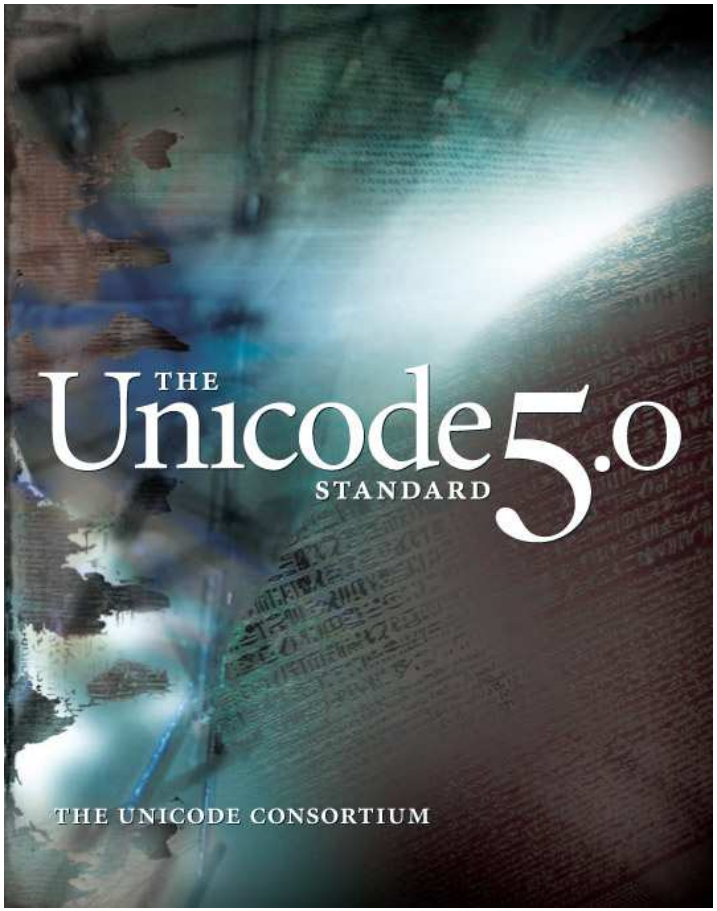
Victor Stinner

Pycon FR, Paris, mai 2009

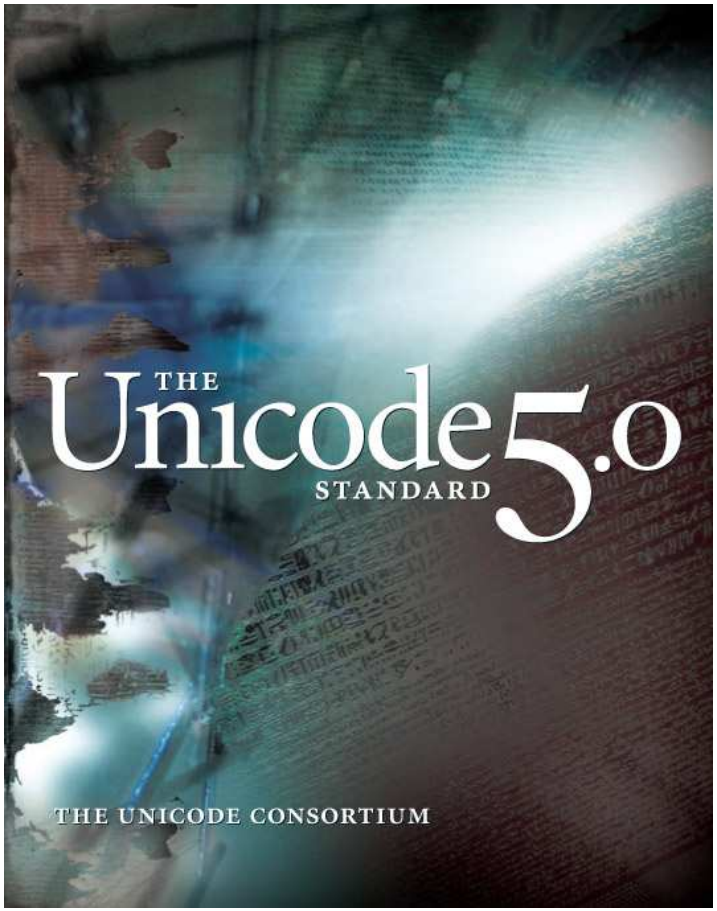

Caractère
Greek capital letter phi
jeu de caractères
Code
0x03A6
943
Octets
'\x00\x00\x03\xa6'
'\xce\xa6'
sérialisation
Glyphe
police de caractères

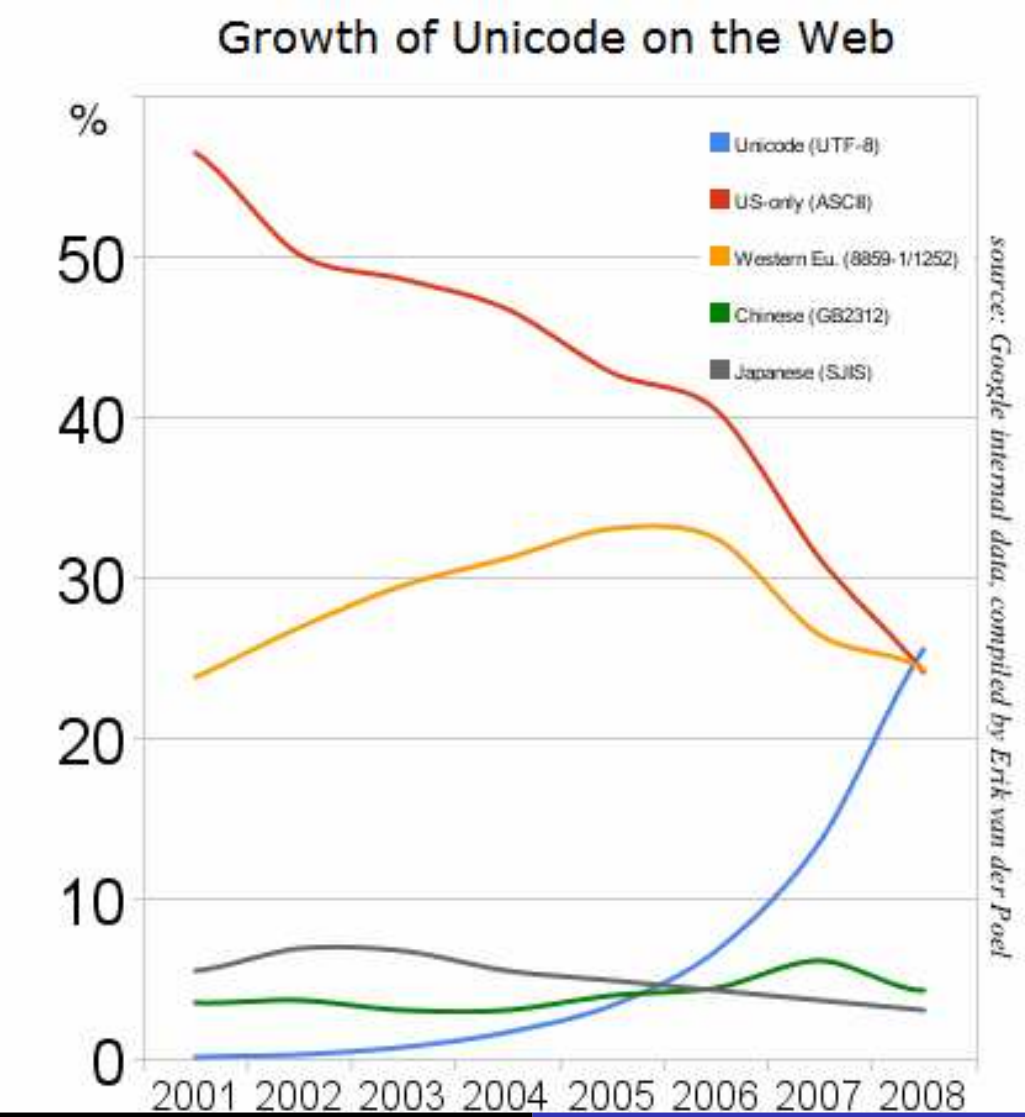

Caractère
jeu de caractères
Code
Octets
sérialisation
Glyphe
police de caractères

- 1961 : ASCII, 128 codes, Américain
- 1987 : ISO-8859-1, 256 codes, Europe Ouest
- 1990 : Unicode, millions de codes, Universel



- Surensemble de tous les jeux de caractères existants
- Indépendant de la langue et du matériel
- 1.114.112 codes (0x00..0x10FFFF)




Growth of Unicode on the Web
%
50
40
30
20
10
0
2001
2002
2003
2004
2005
2006
2007
2008
Unicode (UTF-8)
US-only (ASCII)
Western Eu. (8859-1/1252)
Chinese (GB2312)
Japanese (SJIS)
source: Google internal data, compiled by Erik van der Poel

- Types str (octets) et unicode (caractères)
- "abc" : chaîne d'**octets**
- u"def" : chaîne de **caractères**

- Conversion **implicite** dans les deux sens
- "abc" + u"def" $\rightarrow$ "abcdef"
- "ab\xe9" + u"def" $\rightarrow$ UnicodeDecodeError
- Maux de tête douloureux

- Types bytes (octets) et str (caractères)
- "abc" : chaîne de **caractères**
- b"def" : chaîne d'**octets**
- b"abc" + "def" → TypeError
- Nouvelle bibliothèque io

- Python 2.x : ASCII par défaut
- Python 3.x : UTF-8 par défaut
- PEP 263 : "#coding:utf8"

```
print u"E accent aigu : é" (PEP 263)
# 0x00..0xff
print u"E accent aigu : \xe9"
# 0x00..0xffff
print u"Phi : \u03a6"
# 0x00..0x10ffff
print u"A (cunéiforme) : \U00120000"
print u"Phi : \N{GREEK CAPITAL LETTER PHI}"
```

```
>>> unicodedata.name(u"é")
'LATIN SMALL LETTER E WITH ACUTE'
>>> unicodedata.decomposition(u"é")
'0065 0301'
>>> unicodedata.normalize("NFC", u"u\u0308")
u'ü'
```

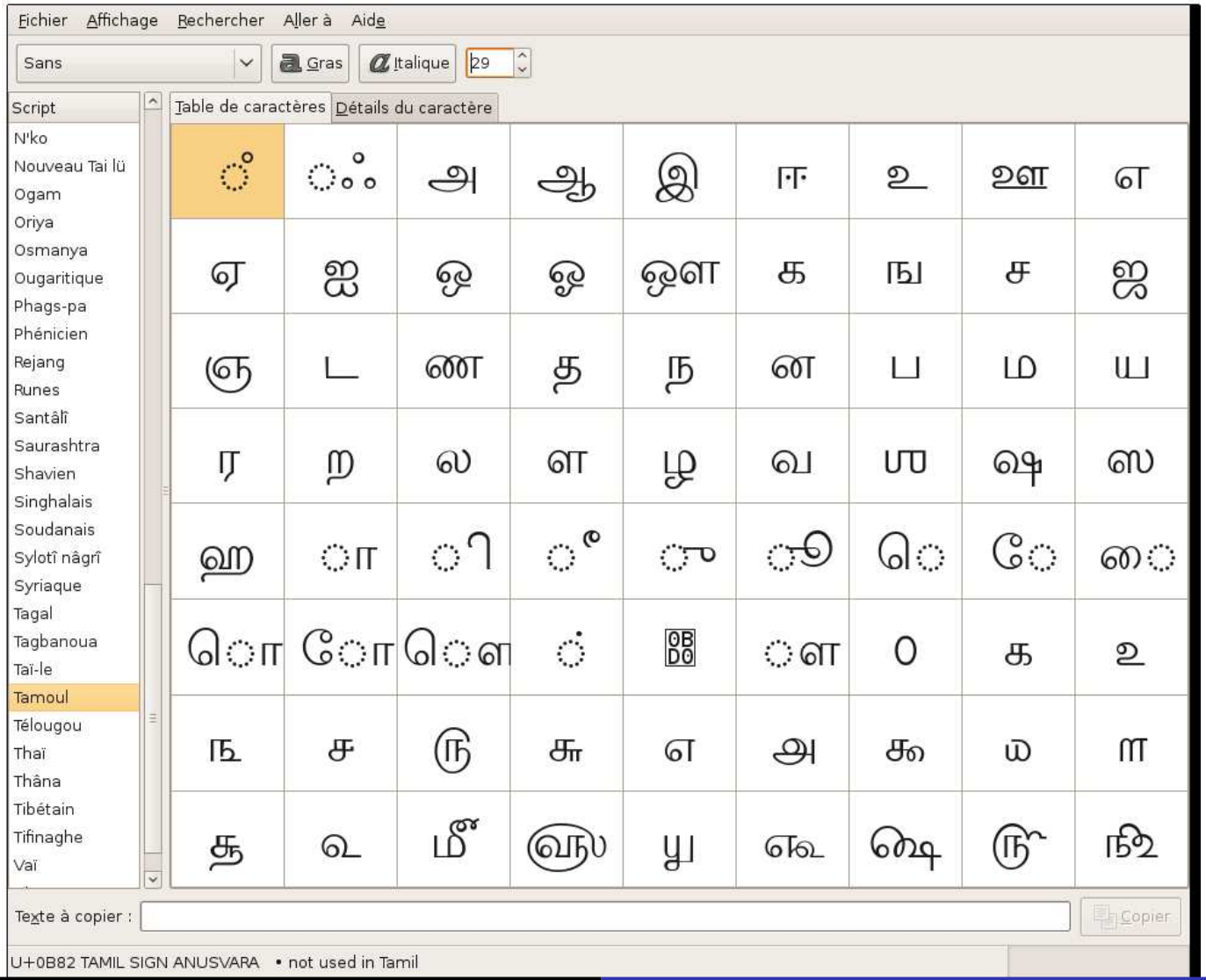
- codecs.open("fichier", "r", encoding="utf-8")
- texte = octets.decode("utf-8")

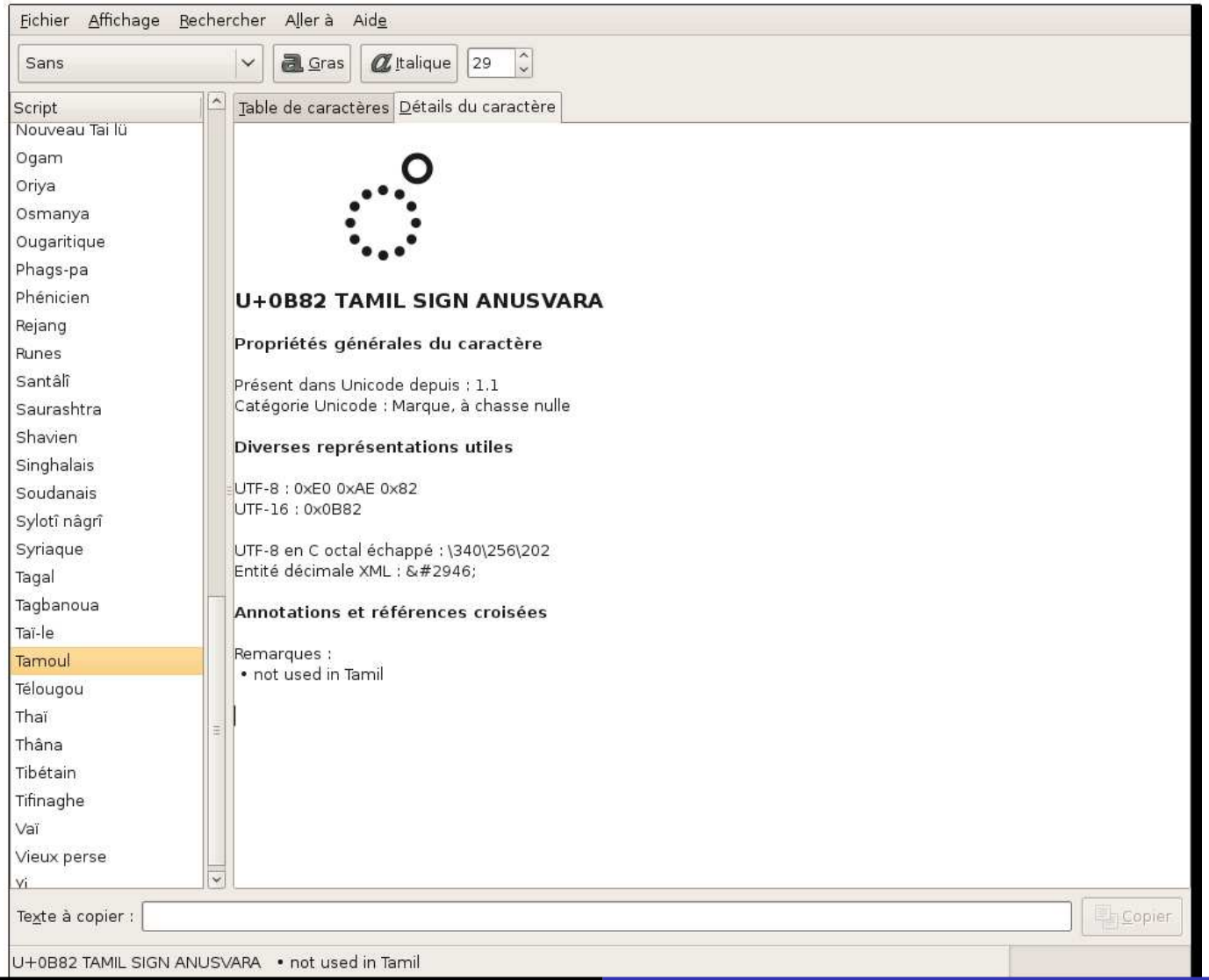
- Fiable pour ASCII et UTF-8
- Heuristique pour les autres
- `chardet` (Mozilla)

- "caractères" → u"caractères"
- from __future__ import unicode_literals
- str(var) → unicode(var)
- def __str__(self) → def __unicode__(self)

- codecs.open("fichier", "w", encoding="utf-8")
- octets = texte.encode("utf-8")

- gucharmap / KCharSelect
- Livre Unicode
- BeautifulSoup
- Django 1.0 parle unicode


Fichier Affichage Rechercher Aller à Aide
Sans
Gras
Italique
29
Script
Table de caractères
Détails du caractère
N'ko
Nouveau Tai lü
Ogam
Oriya
Osmanya
Ougaritique
Phags-pa
Phénicien
Rejang
Runes
Santâlî
Saurashtra
Shavien
Singhalais
Soudanais
Sylotî nâgrî
Syriaque
Tagal
Tagbanoua
Taï-le
Tamoul
Télougou
Thaï
Thâna
Tibétain
Tifinaghe
Vaï
Texte à copier :
Copier
U+0B82 TAMIL SIGN ANUSVARA • not used in Tamil


Fichier Affichage Rechercher Aller à Aide
Sans
Gras
Italique
29
Script
Nouveau Taï lü
Ogam
Oriya
Osmanya
Ougaritique
Phags-pa
Phénicien
Rejang
Runes
Santâlî
Saurashtra
Shavien
Singhalais
Soudanais
Sylotî nâgrî
Syriaque
Tagal
Tagbanoua
Taï-le
Tamoul
Télougou
Thaï
Thâna
Tibétain
Tifinaghe
Vaï
Vieux perse
Yi
Table de caractères
Détails du caractère
U+0B82 TAMIL SIGN ANUSVARA
Propriétés générales du caractère
Présent dans Unicode depuis : 1.1
Catégorie Unicode : Marque, à chasse nulle
Diverses représentations utiles
UTF-8 : 0xE0 0xAE 0x82
UTF-16 : 0x0B82
UTF-8 en C octal échappé : \340\256\202
Entité décimale XML : ஂ
Annotations et références croisées
Remarques :
• not used in Tamil
Texte à copier :
Copier
U+0B82 TAMIL SIGN ANUSVARA • not used in Tamil

- http://www.haypocalc.com/wiki/Python_Unicode
- *Unicode e Python 3* par Ezio Melotti
  http://wolf.netsons.org/pycon/unicode-python3.html

- http://www.flickr.com/photos/thorhakonsen/2349601149/
- http://www.flickr.com/photos/serendipitypeace/2267477928/
- http://www.flickr.com/photos/ecosnake/477735690/