Gestion des caractères API dans le projet AMPER

Albert Rilliard – LIMSI CNRS
Albert.Rilliard@limsi.fr
28 juin 2007

1- Introduction

Afin de pouvoir accéder à une transcription en Alphabet Phonétique International des phrases des enquêtes réalisées dans le cadre du projet AMPER, une procédure générique est nécessaire, afin que chacun puisse lire les phrases transcrites pour tous les comités, et qu'il n'y ait pas de problèmes de lectures des caractères API dans la base de données.

La proposition actuelle consiste donc à utiliser des caractères codés au format standard **UNICODE** (cf. http://www.unicode.org/standard/WhatIsUnicode.html pour des détails). Ce codage de caractères offre l'avantage de permettre de travailler avec les caractères API directement, et non pas avec un recodage de ces caractères, comme c'est le cas en SAMPA par exemple. Il oblige par contre l'utilisation de polices de caractères compatibles avec UNICODE (des détails seront donnés ci-dessous).

Afin de faciliter la transcription des phrases et l'utilisation des caractères API en UNICODE, une **interface** est disponible sur Internet, avec des symboles IPA disposés selon les tableaux IPA classiques. Elle est disponible à l'adresse suivante :
http://people.w3.org/rishida/scripts/pickers/ipa/

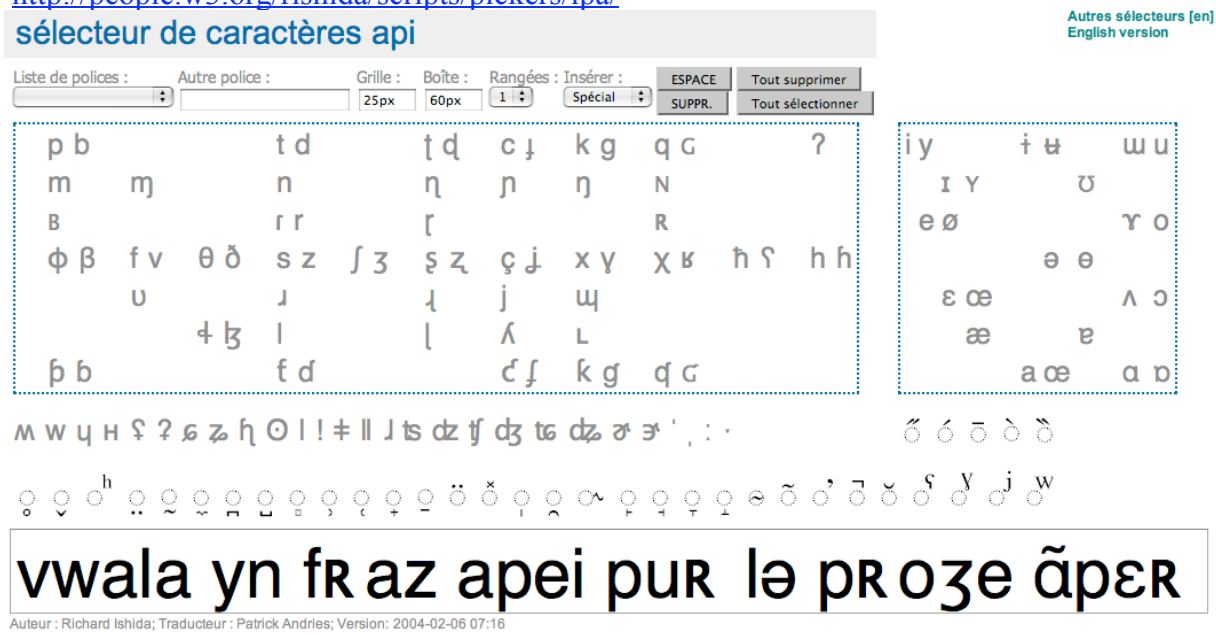


Image 1 : interface Web permettant de sélectionner les caractères phonétique API directement dans des tableaux conçus selon des critères phonétiques.

L'affichage des caractères codés en UNICODE dépend de la police de caractère utilisée. Afin de pourvoir visualiser les caractères API, il faut installer sur son système d'exploitation au moins une police de caractère Unicode qui ait les glyphes spécifiques à l'API. Plusieurs polices existent : par défaut sur les systèmes Windows XP les polices « *Microsoft Sans Serif* » et « *Arial Unicode MS* » sont installées et fournissent la plupart des caractères phonétiques. Cependant, **la police de caractère « Doulos SIL »**[1] les fournit tous, y compris les diacritiques. Pour l'installer, téléchargez-la à l'adresse suivante :
http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=DoulosSIL_download

[1] Version Unicode de la police « SIL Doulos IPA » - attention, ce n'est pas la même police de caractères. « Doulos SIL » a cependant le même aspect que « SIL Doulos IPA » et permet donc de ne pas changer modifier les documents pour ceux qui l'utilisaient.

Des informations supplémentaires à propos d'Unicode et de l'API ainsi que des liens vers d'autres polices de caractères peuvent être consultées à l'adresse suivante : http://www.phon.ucl.ac.uk/home/wells/ipa-unicode.htm

La suite de ce document présente de manière rapide comment il est possible de transcrire les phrases des enquêtes en alphabet phonétique, à l'aide de l'interface web et de la police de caractères « Doulos SIL ». Ce n'est pas la seule manière et les informations minimales requises pour garantir la cohérence des enquêtes et de la base de données seront données en fin de document.

2- Exemple d'utilisation des programmes pour transcrire en API

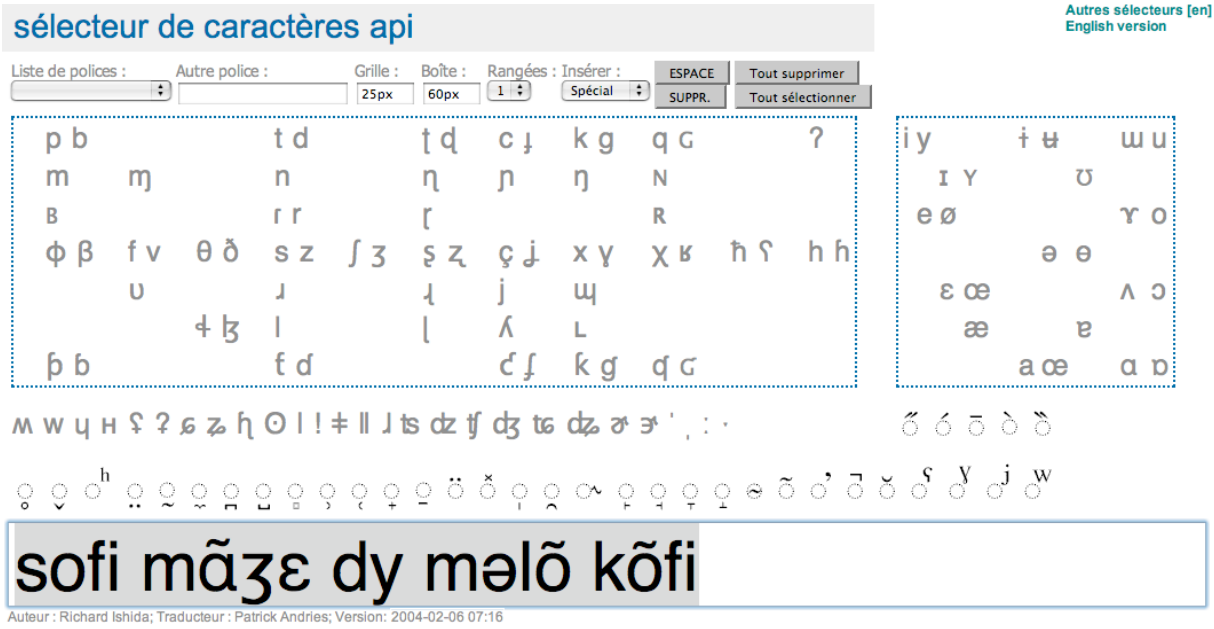
L'interface permet de transcrire une phrase en API, en tapant les caractères latins au clavier ou en sélectionnant les caractères spéciaux de l'API dans l'interface (cf. image 1). La chaîne de caractères ainsi transcrite se trouve en bas de la fenêtre. Elle peut aisément être copiée / collée dans un logiciel de traitement de texte. Il faudra cependant **prendre garde au format du fichier et à la police de caractères** : voici deux formats de fichiers qui permettent de conserver les informations :

- Les fichiers au format **Microsoft Word** (*.doc)
- Les fichiers au format **Texte** AVEC UN CODAGE UNICODE – par exemple l'un des codages suivants : UTF8, UTF16, UNICODE ; mais surtout pas avec les codages ANSI, 8BITS, ASCII qui ne gèrent pas Unicode

Nous allons donner en exemple la transcription de la phrase « *Sophie mangeait du melon confit* » en API, en utilisant les principes données ci-dessus :

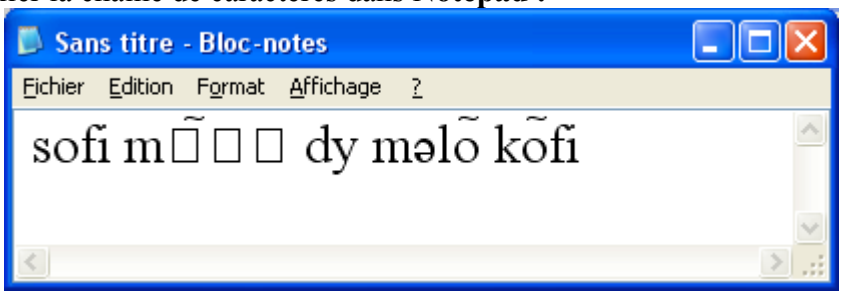
1- Ouvrir l'interface Web

2- Sélectionner les caractères API voulus. On obtient la fenêtre suivante :

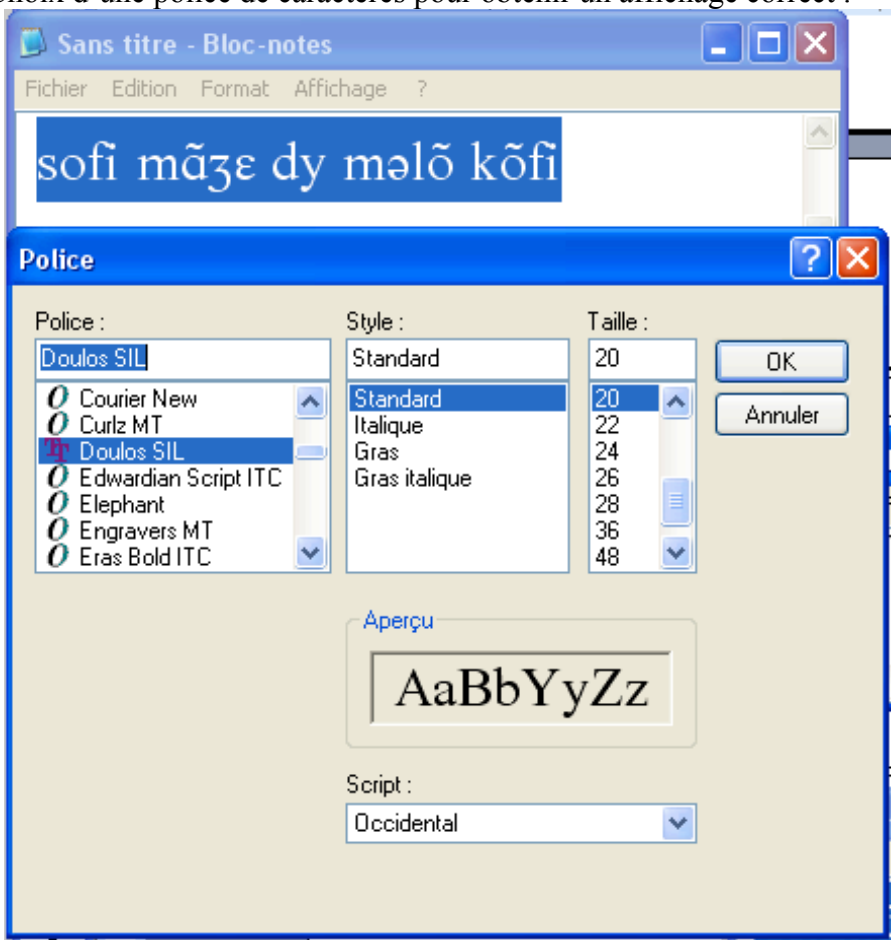


3- Copier la chaîne de caractères

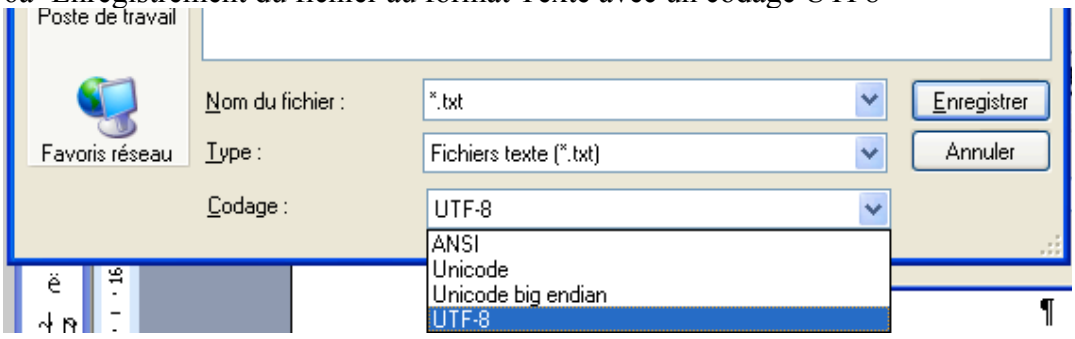
4- Coller la chaîne de caractères dans **Notepad :**



5a Choix d'une police de caractères pour obtenir un affichage correct :



6a- Enregistrement du fichier au format Texte avec un codage UTF8



3- Recommandations pour les transcriptions API dans le projet AMPER

La solution retenue pour le stockage de la transcription en alphabet phonétique internationale des phrases enregistrées dans le cadre du projet AMPER est d'utiliser le codage Unicode de l'API.

Afin que tous les membres du projet puissent fournir des données conformes à ce choix, une solution logicielle possible a été décrite ci-dessus. Ce n'est pas la seule manière d'obtenir une transcription Unicode d'une chaîne phonétique et chacun est libre d'utiliser la solution qu'il trouve la plus simple.

En cas de difficulté, merci de nous contacter afin de trouver ensemble une solution.

Afin que chacun puisse envoyer des données homogènes, les spécifications techniques suivantes devront être respectées au minimum :

- Les transcriptions phonétiques seront réalisées selon le codage de l'API : http://www.arts.gla.ac.uk/ipa/
- Les transcriptions seront transmises dans des fichiers informatiques encodés selon le standard UNICODE : http://www.unicode.org/standard/standard.html

Les recommandations suivantes sont facultatives, mais peuvent être utiles afin d'obtenir des données les plus homogènes possibles :

- Il est recommandé d'utiliser la police de caractère Unicode « Doulos SIL » http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=DoulosSIL_download
- Il est préférable d'enregistrer les fichiers contenant les transcriptions phonétiques au format TEXTE, avec un encodage des caractères UTF8 (c'est la norme qui est utilisée dans la base de données, et cela correspond à l'exemple ci-dessus)

Bon travail à tous.