Le schéma de document CORPUS-Contacts

Pascal Vaillant

25 novembre 2010

1 Objectifs généraux

Le schéma de document *CORPUS-Contacts* a été défini dans le cadre du programme *Contacts de langues*[1] de l'unité de recherche CELIA[2].

Son but est de fournir un schéma de structure de documents contenant des corpus linguistiquement hétérogènes.

L'utilisation de ce schéma de documents a des objectifs pratiques :

1. Le schéma de documents normalise la représentation des corpus recueillis par les différents chercheurs intéressés aux phénomènes de contacts de langues. Cette normalisation permet à chacun d'entre eux de profiter de l'ensemble du corpus commun.
2. Les corpus enregistrés contiennent une représentation de la structure et non plus de la forme. Le fichier contient directement, par exemple, une indication que telle unité lexicale est un emprunt à telle langue — et non plus une mise en forme comme la mise du mot en italique, qui oblige à avoir recours à une convention extérieure pour savoir ce qu'elle représente, et qui est moins aisément généralisable.

   Ceci n'empêche pas la mise en forme de ces indications de structure pour les rendre plus lisibles, et conformes aux convention habituelles de représentation des chercheurs. Mais cette mise en forme ne nécessite pas de travail supplémentaire de la part de l'utilisateur : elle est le résultat d'une conversion automatique réalisée par l'utilisation d'une feuille de style XSLT[3].
3. La représentation structurée des corpus permet ensuite de définir des fonctions de recherche d'information et de classification sur des documents structurés.

La figure 1 illustre ce découplage entre structure et présentation, qui permet la normalisation de plusieurs corpus et de leurs schémas d'annotation.

Dans la conception d'un schéma de documents répondant aux besoins ainsi définis, on est soumis à deux contraintes :

– une contrainte de **normalisation** d'une part, nécessaire dans la perspective de maximiser la réutilisabilité et le partage des corpus (transposition aisée dans un autre format, portage aisé sur un autre site, ouverture de certaines parties du corpus à des communautés d'utilisateurs plus vastes, utilisation d'outils standards ...) ;

[1] URL : `http://celia.cnrs.fr/Fr/AxeDeRecherche.htm`.

[2] *Centre d'Études des Langues Indigènes d'Amérique*, UMR CNRS 8133, IRD 135, INaLCO, Université Paris-6. URL : `http://celia.cnrs.fr/`.

[3] Une feuille de style a également été définie dans le cadre du programme *Contacts* : `corpus.xsl`. Description disponible à l'URL : `http://clapoty.vjf.cnrs.fr/contacts/doc/corpus_xsl.pdf`.

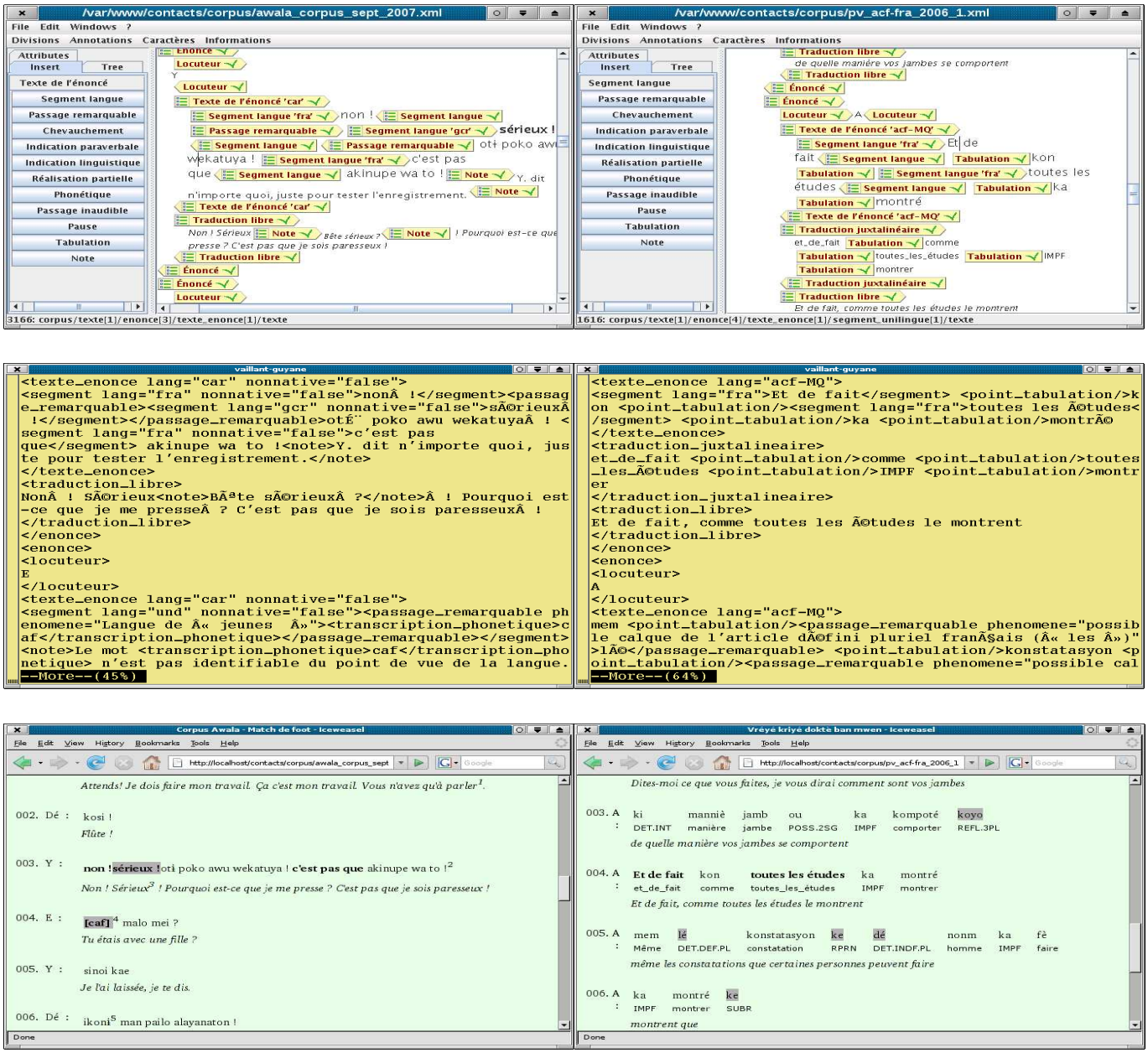


FIG. 1 – Codage de corpus : (a) saisie dans un éditeur XML, JAXE (en haut) ; (b) représentation informatique interne (« code source ») (au milieu) ; (c) présentation visuelle associée (en bas). À gauche, un exemple de corpus multilingue kali'na / français ; à droite, un exemple créole martiniquais / français.

On constate d'une part que le découplage contenu / présentation permet d'inclure des annotations explicites dans le corpus, en se concentrant sur leur contenu, sans avoir à se préoccuper de la manière dont elles seront visualisées à l'écran ; et d'autre part que le formalisme de représentation interne des données est le même à gauche et à droite, ce qui rend les corpus facilement mutualisables, et accessibles aux mêmes méthodes de recherche et d'analyse.

– une contrainte de **simplicité** d'autre part, absolument vitale pour l'acceptation de la ressource par les utilisateurs (si un utilisateur doit investir un temps important pour se former à un langage documentaire complexe, puis à nouveau du temps, lors de chaque saisie ou modification de corpus, pour remplir des dizaines de champs d'information obligatoires et la plupart du temps inutiles, il risque tout simplement de renoncer aux avantages de la normalisation).

Le choix qui a été fait ici a été de donner la priorité à la simplicité, sans sacrifier la normalisation. En l'occurrence, ceci signifie que :

– Seules les informations nécessaires, au stade actuel du programme de recherche, ont été incluses dans le schéma de documents CORPUS-Contacts. Des pans entiers de normes existantes (notamment TEI, cf. ci-dessous) n'y ont pas été intégrées car elles n'ont pas d'utilité immédiate pour les chercheurs impliqués dans le programme de recherche sur les contacts de langue. Le fichier contenant la description du schéma (`corpus.xsd`) est donc « aussi petit que possible ».
– Le minimum concevable d'informations a été défini comme étant obligatoire. L'utilisateur peut presque commencer à utiliser le schéma de document en tapant du texte au kilomètre dans un éditeur XML, et ne commencer à utiliser les possibilités d'enrichissement de l'information que lorsque le besoin s'en fait sentir pour lui.
– Pour autant, le sous-ensemble des informations utiles représentées dans CORPUS-Contacts respecte l'organisation et la nomenclature de la norme TEI[4], qui s'est imposée dans l'usage international comme la norme générale de référence pour la représentation des textes.

2 Structure globale

Le schéma de documents CORPUS-Contacts est un schéma de documents XML[5] au sens donné à ce terme par le W3C[6].

Un document XML est un document structuré contenant des informations stockées avec le texte sous forme de « balises ». Il se compose d'une hiérarchie d' *éléments* (par exemple : un livre se définit comme un ensemble de chapitres eux-mêmes constitués d'un ensemble de paragraphes), eux-mêmes caractérisables par des *attributs* (par exemple : tel paragraphe est en français, tel paragraphe est en créole).

XML[7] n'est pas une norme qui définit à l'avance, et dans le détail, tous les éléments et tous les attributs utilisables dans tous les types de documents. C'est une norme paramétrable, qui offre la possibilité de définir des types de documents en fonction des applications. Ainsi, on peut créer une famille de documents XML correspondant à des fiches bibliographiques, une autre correspondant à des documents destinés à la publication Web, une autre correspondant à des entités de base de données, une autre encore contenant des textes littéraires, etc.

Le concept qui permet cette polyvalence est le *schéma de documents*. Un schéma de documents est la définition de la structure commune que doivent avoir plusieurs documents XML de la même famille et destinés aux mêmes usages (ex. des livres définis comme des ensembles de chapitres ...)

[4] *Text Encoding Initiative*, ensemble de recommandations définies par un consortium international de chercheurs et d'universitaires pour la représentation électronique des textes. URL : `http://www.tei-c.org`. Le détail des recommandations est contenu dans le document `http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf`.

[5] `http://www.w3.org/XML/Schema`

[6] *World Wide Web Consortium*; URL : `http://w3.org`.

[7] *eXtensible Markup Language*. URL : `http://www.w3.org/XML/`.

Le schéma de documents CORPUS-Contacts est donc la description générique de ce en quoi consiste un corpus de données linguistiques utile pour la recherche sur le contact des langues. Il contient un squelette de document minimal (description des informations obligatoires), et définit des types d'éléments et d'attributs à utiliser pour tout un ensemble d'informations supplémentaires possibles.

3 Description détaillée

3.1 Squelette de document

Le schéma définit un type de document appelé *corpus*.

Un *corpus* est constitué d'un *en-tête* global, et d'un ou plusieurs *textes* (fig. 2).

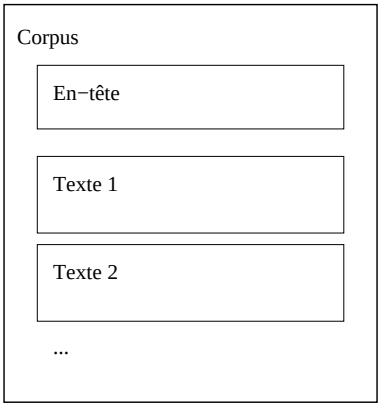


FIG. 2 – Structure d'un document de type *corpus*

3.2 L'en-tête du corpus

L'*en-tête* du corpus (fig. 3) contient les informations suivantes :

- un titre global pour tout le corpus (obligatoire) ;
- un ou plusieurs noms d'éditeurs (facultatif) ;
- une date d'édition (facultative), qui peut préciser soit la date complète (format : 2002-12-28), soit seulement le mois (format : 2002-12), soit seulement l'année (format : 2002) ;
- une catégorisation du corpus selon trois typologies de situations de contact proposées respectivement par Winford, Lüdi, et Auer (la catégorisation est facultative, et l'on peut y ajouter des commentaires facultatifs) ;
- un texte de présentation du corpus, formé d'une suite de paragraphes (facultatif) ;
- un inventaire des différentes langues représentées dans le corpus, qu'elles le soient dans des énoncés entiers, ou seulement par petits segments empruntés (obligatoire).
- un inventaire des différents locuteurs intervenant à un moment ou à un autre dans le corpus (obligatoire).

Précisions :

La grille de classement selon les trois typologies est présentée dans le tableau 1. Les commentaires sur la catégorisation (présentation des situations, justification des choix, doutes, possibilités alternatives, comparaisons avec d'autres situations ...) peuvent être ajoutés au format texte libre.

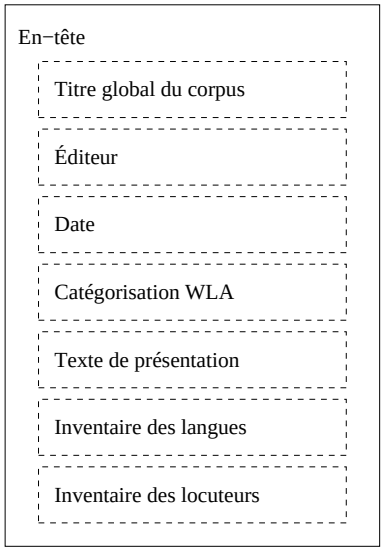


FIG. 3 – Structure d'un *en-tête de corpus*

Le texte de présentation du corpus, ainsi que les éventuels commentaires annexés à la grille de catégorisation WLA, sont au format *texte libre*, c'est-à-dire qu'ils consistent en une séquence d'un nombre indéterminé de paragraphes. Dans chaque paragraphe, on peut insérer en outre des sauts de ligne afin d'aller à la ligne (par exemple pour une énumération de plusieurs points).

3.2.1 Inventaire des langues

L'inventaire des langues est obligatoire. Si toutes les langues du corpus n'y sont pas mentionnées, le document n'est pas « valide ».

L'inventaire des langues se compose simplement d'une énumération de langues, identifiées par leur code (pour une explication sur les codes des langues, voir plus bas, § 5).

L'ordre de cette énumération est important : la première langue mentionnée est en effet considérée comme étant la langue « de base » du corpus ; les suivantes doivent être énumérées autant que possible par ordre d' « importance » décroissante (ou par ordre d'apparition dans le corpus, s'il n'y a pas de critère pour distinguer leurs degrés d'importance relative).

Il faut également mentionner dans cet inventaire, s'ils sont utilisés, les codes correspondant à *« rattachement de langue multiple »* (`'mul'`), *« langue inconnue »* (`'und'`), ou à *« contenu non-linguistique »* (`'zxx'`) — cf. le détail concernant les codes de langue en § 5.

3.2.2 Inventaire des locuteurs

De la même manière que l'inventaire des langues, l'inventaire des locuteurs est obligatoire. Tous les locuteurs figurant dans le corpus doivent y être mentionnés pour que le document soit « valide »[8].

La description du locuteur comprend un seul élément obligatoire : le **nom abrégé** ; c'est l'identifiant qui sera repris devant chaque énoncé dans le corpus. Il doit être court, et consister

[8]Le document est considéré comme « valide » si toutes les contraintes de bonne formation définies dans le schéma sont respectées (structure hiérarchique des éléments et des attributs, présence de tous les éléments obligatoires, contraintes d'unicité et de co-référence). Cela étant, on peut travailler temporairement avec un document non-valide ; il vaut mieux simplement s'assurer qu'il est valide avant de le rendre disponible.

Typologie de Winford

(W1) **situations de bilinguisme** dites **stables et anciennes** où le bilinguisme apparaît comme la norme pour les locuteurs (Suisse, Belgique, etc.)

(W2) **situations où la colonisation a introduit des langues européennes qui sont en contact avec des langues autochtones** (Afrique, Asie du Sud-Est, Caraïbe, Amérique du Sud, etc.)

(W3) **situations liées aux migrations vers des pays industrialisés**, qui ont conduit à la création de minorités linguistiques devenant bilingues dans la langue du pays d'accueil et qui, dans certains cas, aboutissent à une situation de monolinguisme en L2 (Europe, Amérique du Nord, etc.)

(W4) **situations où des locuteurs de variétés non-standard doivent apprendre la variété standard de leur langue** pour des raisons socio-économiques, et dont la conséquence est le bidialectalisme et donc des alternances entre variétés de langues

Typologie de Lüdi

Échange	bilingue	monolingue
Exolingue	(L1) interactions entre des locuteurs de langues différentes	(L3) interactions entre des locuteurs natifs et des locuteurs non-natifs
Endolingue	(L2) interactions entre bilingues	(L4) interactions entre monolingues

Typologie de Auer

(A1) **Alternance conversationnelle** (AC) : certains passages de la conversation se déroulent dans une langue, d'autres dans une autre. Ce qui est significatif dans ce type de discours, c'est le fait que les participants considèrent qu'à un moment donné l'alternance de code signale un changement contextuel ou encore un problème de compétence. Les locuteurs sont capables d'identifier les langues ou variétés de langues.

(A2) **Mélange de langues** (ML) : juxtaposition d'éléments appartenant à des langues ou variétés de langues différentes. Ce qui est significatif ici, c'est l'usage alternatif des langues en soi : le « mélange » est en soi la « langue » de l'interaction. On ne peut expliciter la fonction des alternances (ou mélanges). Par contre les locuteurs ont conscience d'échanger sur un mode bilingue-plurilingue qui ne peut être employé qu'avec des bilingues-plurilingues (compétents ?) comme eux. Les unités alternées/mélangées sont de taille variable, elles peuvent se situer aux frontières des propositions, constituants ou à l'intérieur de ceux-ci. Cette catégorie se divise facultativement en deux types de mélange : mélange « alternationnel » (A2a) et mélange « insertionnel » (A2b).

- (A2a) **Mélange alternationnel** : les alternances sont symétriques et il est difficile de déterminer s'il y a une langue matrice
- (A2b) **Mélange insertionnel** : présence d'une langue matrice au sens de Myers-Scotton (1993) dans laquelle sont insérés des éléments d'une autre langue (items, constituants, îlots)

(A3) **Fusion de langues** (FL) : il n'y a plus conscience, de la part des locuteurs, que ce qu'ils produisent est du « mélange » de deux ou plusieurs langues ; pour eux il s'agit d'une langue ou d'une variété de langue à part entière : on peut donc la décrire comme on le ferait pour n'importe quelle langue.

TAB. 1 – Typologies des situations de contact selon Winford, Lüdi, Auer

par exemple en une initiale seulement (A pour Antoine), si celle-ci permet à elle seule d'identifier sans ambiguïté le locuteur. À défaut, il peut consister en une courte suite de lettres (par exemple, Lé pou Léon et Lo pour Louis), ou en une lettre et un chiffre (L1 et L2).

Deux locuteurs ne doivent pas avoir le même nom abrégé, car cet élément sert à identifier de manière unique le locuteur, dans les énoncés et dans la table des langues parlées par les locuteurs.

Les autres éléments de description sont facultatifs et peuvent être ajoutés en fonction des besoins. Ils comprennent :

- un **nom complet** : par exemple : Barbara Dennerlein ;
- une **description** en texte libre.

3.3 Le texte inclus dans un corpus

Chaque *texte* (fig. 4) faisant partie d'un corpus se définit essentiellement comme une série d'énoncés (ou le cas échéant d'indications extra-linguistiques), additionnée de quelques informations supplémentaires. Plus précisément, chaque *texte* possède la structure suivante :

- une date de recueil (obligatoire, au moins avec la précision de l'année — les formats de date possible sont les mêmes que pour la date d'édition du corpus, cf. plus haut, § 3.2) ;
- un titre spécifique à ce texte particulier (facultatif) ;
- un texte de présentation de ce texte particulier, au format texte libre (facultatif) ;
- une série d'un ou plusieurs événements.

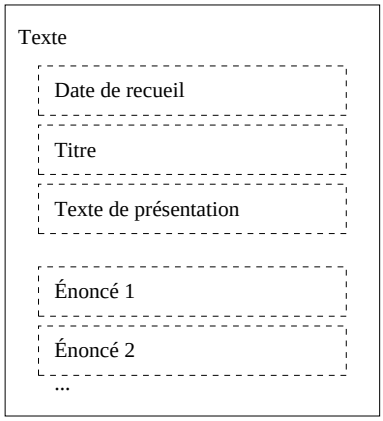


FIG. 4 – Structure d'un *texte*

L'ensemble des constituants internes d'un texte peut concrètement être stocké dans le même fichier XML que le corpus et l'en-tête du corpus ; mais il peut aussi être stocké dans un autre fichier, indépendant, dont *texte* constitue l'élément racine. Dans ce dernier cas, dans le fichier XML servant à stocker le corpus, on remplace le contenu concret du texte par une simple référence au fichier externe qui le contient : l'élément *texte* ne contient plus alors qu'un seul élément, *réf. texte*, qui ne donne qu'un chemin d'accès au fichier contenant l'ensemble des énoncés[9]. Cela ne change rien au contenu ni (fondamentalement) au mode de codage du contenu, si ce n'est que le corpus, au lieu d'être stocké dans un seul grand fichier XML, est alors stocké dans plusieurs fichiers : un donnant des indications sur le corpus de manière générale (et contenant l'en-tête du corpus), plus un fichier XML supplémentaire pour chaque texte référencé dans le

[9]Par exemple : `https://clapoty.vjf.cnrs.fr/contacts/corpus/corpus-001_texte-02.xml`.

corpus. Cette possibilité peut être exploitée pour constituer plusieurs corpus à partir du même ensemble de textes, mais en les regroupant en sous-ensembles différents.

3.4 Les événements

Chaque texte est constitué d'une série d'*événements*. Un événement peut être soit un énoncé (cas le plus fréquent), soit une indication sur la situation extra-linguistique. Dans le premier cas, on utilise l'élément de description *énoncé* (cf. § 3.5) ; dans le second, l'élément de description *indication paraverbale* (cf. § 4.2.1).

3.5 Les énoncés

L'énoncé est l'unité de base des corpus linguistiques. Il est associé à un locuteur. Il doit pouvoir être accompagné d'une traduction interlinéaire (pour décrire morphème par morphème la structure exacte de l'énoncé dans la langue utilisée), ainsi que d'une traduction libre (fig. 5).

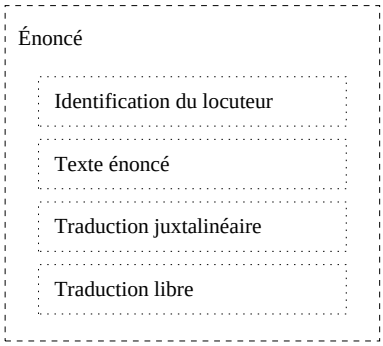


FIG. 5 – Structure d'un *énoncé*

Éléments inclus dans un énoncé :

- **Identification du locuteur** : on indique ici le nom abrégé du locuteur de l'énoncé (qui peut être par exemple son initiale — voir plus haut, § 3.2.2) (obligatoire).
- **Texte énoncé** : la transcription exacte de l'énoncé (obligatoire).
  Le texte de l'énoncé est accompagné de deux attributs : l'attribut obligatoire *lang*, qui indique la langue de base de l'énoncé (cf. plus bas, § 5.1), et l'attribut *nonnative* qui précise, le cas échéant, que le locuteur n'est pas un locuteur natif de cette langue (cf. plus bas, § 5.3).
  Certains énoncés contiennent plusieurs langues ; dans ce cas, il est possible de le préciser en ajoutant un élément *langues*, qui dresse la liste des langues présentes dans ces énoncés (cf. § 5.2 pour une description précise de ce mécanisme). Cependant, cette représentation de la liste des langues de l'énoncé n'est pas nécessaire lorsque la même information peut être donnée à un niveau plus spécifique, c'est-à-dire en identifiant précisément des *segments* de telle ou telle langue au sein de l'énoncé (cf. § 4.1.1). Dans ce cas, dresser en outre la liste des langues de l'énoncé serait redondant.
  On ne devrait donc utiliser cette possibilité de donner la liste de langues au niveau de l'énoncé que dans les cas où il est difficile de segmenter l'énoncé en segments linguistiquement homogènes, et où tout l'énoncé ne forme qu'un seul segment au rattachement incertain (ce qui arrive pour des énoncés courts).
- **Traduction juxtalinéaire** : la traduction juxtalinéaire (ou interlinéaire) de l'énoncé, transcrit morphème par morphème l'énoncé d'origine, en donnant des équivalents des mor-

phèmes dans la langue de traduction, ou, à défaut, des abréviations métalinguistiques conventionnelles des morphèmes grammaticaux (facultative)[10].
La traduction juxtalinéaire est accompagnée d'un attribut *lang*, qui indique la langue utilisée pour la traduction (c'est la langue utilisée par le linguiste pour son travail, par exemple le français ou l'anglais).

– **Traduction libre** : la traduction de l'énoncé par un énoncé équivalent dans la langue de traduction (facultative).
La traduction libre, comme la traduction juxtalinéaire, est accompagnée d'un attribut *lang*, qui indique la langue utilisée pour la traduction.

Le texte de l'énoncé — et dans une certaine mesure également la traduction juxtalinéaire et la traduction libre — peuvent, outre du texte brut, faire figurer différentes annotations linguistiques : indication de changement de langue, indications prosodiques, paralinguistiques, appel de notes ... Ces éléments particuliers de description linguistique sont décrits plus en détail plus bas, dans la section 4.

3.5.1 Correspondance entre texte de l'énoncé et traduction juxtalinéaire

Il est important, lorsque l'on propose une traduction juxtalinéaire de l'énoncé d'origine, de bien faire correspondre chaque unité de cette traduction juxtalinéaire avec le morphème correspondant dans le texte de l'énoncé. Pour cela, le schéma CORPUS-Contacts propose un mécanisme d'alignement : le *point de tabulation*.

Il faut placer un point de tabulation à chaque frontière de morphèmes du texte de l'énoncé d'origine, et placer le même nombre de points de tabulation dans la traduction juxtalinéaire.

Il est ainsi possible à un programme d'analyse du corpus balisé XML de retrouver sans ambiguïté la correspondance biunivoque entre morphèmes d'origine et traductions de morphèmes. C'est par exemple ce que fait automatiquement le logiciel de navigation, lorsqu'il charge la feuille de style associée à ce schéma (`'corpus.xsl'`), afin de présenter une visualisation tabulée à l'utilisateur (ex. fig. 6).

3.6 Contraintes d'unicité et de co-référence

3.6.1 Identification des langues et des locuteurs

Comme mentionné ci-dessus (§ 3.2), l'en-tête de chaque corpus doit contenir un inventaire des langues utilisées dans le corpus, ainsi qu'un inventaire des locuteurs qui y apparaissent.

L'intérêt est de pouvoir ensuite attribuer à chaque énoncé un locuteur qui, étant présent dans la liste des locuteurs référencés, soit associé à des informations personnelles (âge, niveau d'études, etc.)

[10]Les principes et abréviations les plus courantes des traductions juxtalinéaires utilisées en linguistique (et plus particulièrement en typologie) sont documentées d'une part par un article de Christian Lehmann : « Interlinear Morphemic Gloss », in G. Booij, C. Lehmann, J. Mugdan & S. Skopeteas (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband*. Berlin, Walter de Gruyter (Handbücher der Sprach- und Kommunikationswissenschaft, 17.2).
Disponible à l'URL : `http://www.christianlehmann.eu/CL_Publ/IMG.PDF`.
d'autre part par un guide synthétique édité par les linguistes de Leipzig : *The Leipzip Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses* (collectif, Max-Planck-Gesellschaft et Université de Leipzig).
Disponible à l'URL : `http://www.eva.mpg.de/lingua/pdf/LGR09_02_23.pdf`.

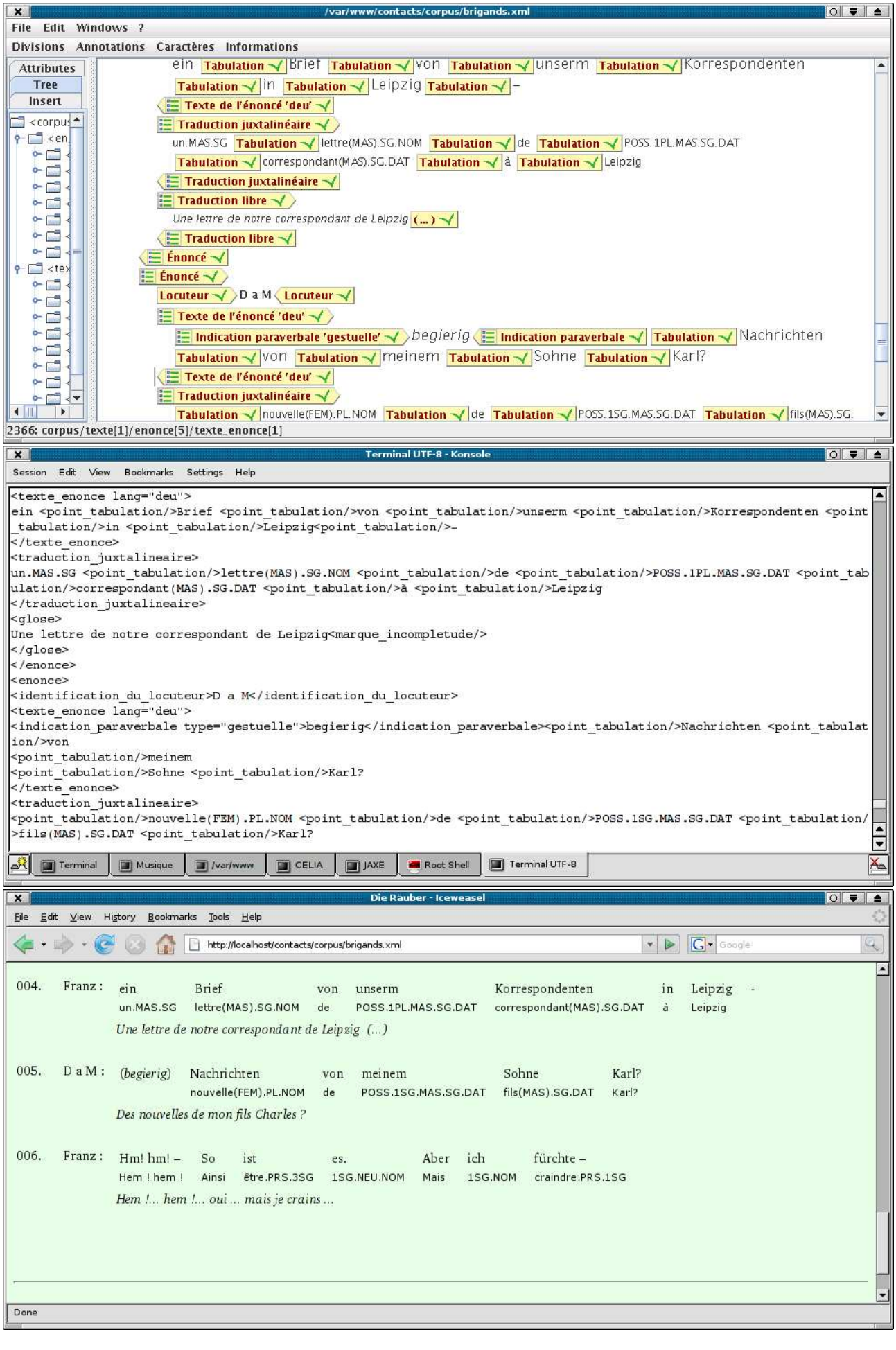


FIG. 6 – Alignement point par point de la traduction juxtalinéaire sur le texte de l'énoncé.

De même, on peut attribuer à chaque énoncé (et à chaque segment d'énoncé, dans le cas d'alternances) une langue (voire plusieurs), qui soit décrite par ailleurs dans l'inventaire des langues du corpus.

Ce principe est représenté dans la fig. 7.

Un document peut ne pas comporter d'inventaire complet des langues et des locuteurs, mais il n'est pas alors considéré comme « valide » (c'est-à-dire satisfaisant pleinement aux contraintes de bonne structuration). Les conséquences sont alors :

- **au niveau de la feuille de style :** Si les langues ne sont pas référencées, et que l'utilisateur visualise le document à travers la feuille de style XSLT définie pour ce type de documents dans le cadre du programme CELIA Contacts (`corpus.xsl`), alors le logiciel de navigation ne pourra pas utiliser la fonction permettant de visualiser les différentes langues du corpus à l'aide de différentes conventions typographiques, par ordre d'importance (première langue en romain maigre, deuxième langue en romain gras, troisième langue en romain souligné, etc.) Il se rabattra sur une solution par défaut, qui fera afficher les passages des langues non-inventoriées comme si elles étaient toutes dans la catégorie « langue de rang 5 ou au-delà » (à savoir : en italique souligné) ;
- **au niveau des fonctions de recherche :** Si les langues et/ou les locuteurs ne sont pas référencés, il ne sera pas possible de faire des recherches par langue et/ou par locuteur.

En pratique, il est possible d'alléger le travail consistant à créer cet inventaire de langues et de locuteurs, en vérifiant, en cours de saisie ou en fin de saisie, que toutes les langues représentées dans le corpus, et que tous les locuteurs mentionnés dans le corpus, sont bien inclus dans l'inventaire. Cette possibilité est liée à la fonction de *validation* de l'éditeur XML utilisé.

3.6.2 Identification des énoncés et des notes

Les énoncés peuvent (facultativement) être « étiquetés », afin de pouvoir y faire référence par la suite au moyen de cette étiquette, par exemple dans une note ultérieure (à titre de comparaison, ou au cours d'un commentaire), ou afin de pouvoir les retrouver rapidement lors d'une recherche dans le corpus.

Afin de permettre cette référence, l'élément *énoncé* est donc doté d'un attribut qui joue le rôle une étiquette générique destinée à pouvoir identifier de manière unique et non-ambiguë l'énoncé : l'attribut *id* (identifiant).

L'attribut *id* est à valeur *facultative* : on peut le laisser vide, et ne le remplir que lorsqu'il est nécessaire. Un énoncé auquel on n'a pas besoin de faire référence n'a pas besoin d'identifiant.

Lorsqu'on doit faire référence à un énoncé, en revanche, il faut remplir son attribut *id*. La valeur de l'attribut *id* est une série de caractères :

- dont le premier caractère est obligatoirement une lettre (qui peut être une lettre accentuée) ou un caractère *blanc souligné* (`"_"`) ;
- et dont les caractères suivants peuvent être soit des lettres (y compris des lettres accentuées), soit des chiffres, soit l'un des signes typographiques « autorisés » suivants : *blanc souligné* (`"_"`), *trait d'union* (`"-"`), ou *point* (`"."`)[11].

L'attribut *id* doit permettre l'identification unique et non-ambiguë de l'énoncé sur lequel il porte ; ceci implique que deux énoncés ne peuvent pas avoir le même identifiant. Cette contrainte d'unicité vaut pour tout le fichier corpus (pas seulement pour un seul texte du corpus). Il faut donc trouver un schéma personnel pour étiqueter chaque énoncé de manière unique. On

[11] Spécifications de XML 1.1 : http://www.w3.org/TR/2006/REC-xml11-20060816/#NT-Names.

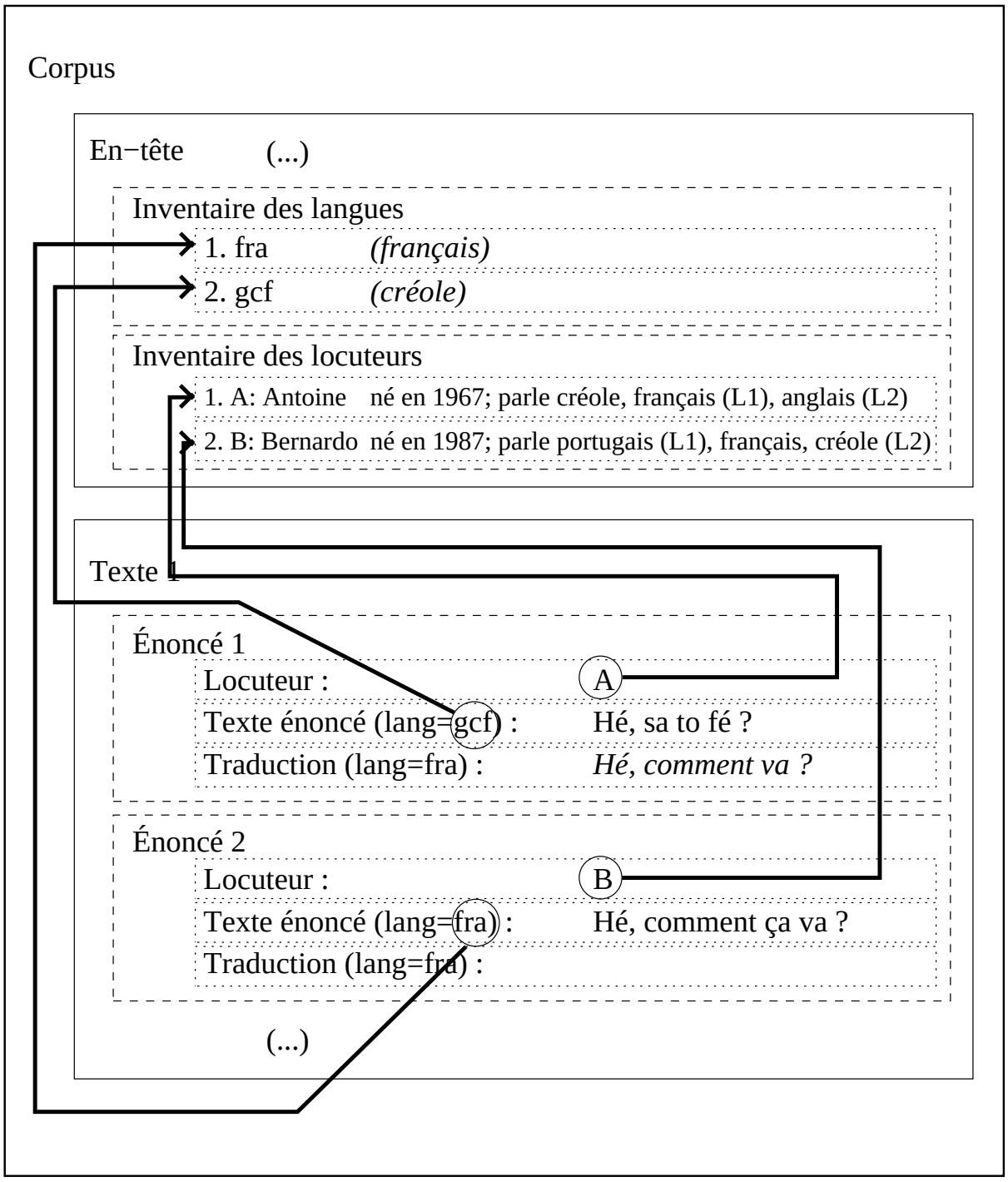


FIG. 7 – Indication des langues et des locuteurs par référence à une entrée dans un inventaire complet

peut prendre par exemple le principe consistant à donner à *id* une valeur composée de lettres rappelant le nom du locuteur, et de chiffres rappelant le numéro de l'énoncé dans la série des énoncés étiquetés (par exemple, `"e2"` désigne le deuxième énoncé étiqueté du locuteur E). On peut aussi donner aux énoncés des identifiants rappelant le phénomène qui les rend dignes d' « intérêt » (par exemple : `"remarque_idiote_de_frank"`). En fait, n'importe quelle séquence de lettres et de chiffres est possible, à partir du moment où la contrainte d'unicité est respectée.

De la même manière que les énoncés, les *notes* (cf. § 4.10) peuvent (facultativement) être étiquetées par un attribut *id* afin qu'il puisse y être fait référence par ailleurs. La contrainte d'unicité des valeurs de l'attribut *id* s'applique sur tout le domaine d'un document de type corpus, et non pas seulement sur l'ensemble des énoncés : cela signifie concrètement qu'une note ne peut pas avoir la même valeur *id* qu'un énoncé, même si c'est une autre espèce d'élément. Chaque élément (énoncé ou note) qui comporte un attribut *id* rempli, doit avoir une valeur qui lui soit absolument propre, à travers tout le document.

4 Éléments de la description linguistique

Cette section décrit les différents types d'éléments spécifiques qui peuvent enrichir les corpus, en permettant d'y porter des annotations.

4.1 Caractérisation des langues manifestées dans l'énoncé

4.1.1 Segments dans une autre langue

Au sein d'un énoncé dans une langue A, peut s'insérer un segment, plus court, dans une langue B. L'élément *segment* permet de noter ce phénomène d'alternance. Il encadre (par une balise de début, et une balise de fin) un segment qui est homogène dans la langue B.

rus	eng	rus	eng
Sobral sebe ogromnuju kollekcuju vorovannyx	*piece*	ov	*of art*

Exemple de segments en anglais dans un énoncé en russe (d'après Babyonyshev, 2004).

Les segments sont des atomes en ce qui concerne l'homogénéité linguistique. Ceci ne signifie pas qu'ils ne peuvent pas relever, dans certains cas, de plusieurs langues à la fois : en effet, dans des contextes de communication bilingue, il arrive que certains segments soient à rattachement ambigü. En revanche, cela signifie qu'ils ne peuvent pas être décomposés en deux sous-segments *consécutifs* (sur la dimension syntagmatique) relevant de façon clairement identifiable de deux langues différentes.

En conséquence, les segments ne peuvent donc pas être imbriqués : un segment ne peut être inclus dans un autre segment.

Par ailleurs, un élément de type *segment* peut être inséré dans le texte de l'énoncé, mais pas dans la traduction (ni juxtalinéaire, ni libre) — qui est entièrement dans la méta-langue choisie par le linguiste.

Le segment est caractérisé, outre par ses positions de début et de fin, par les mêmes attributs d'identification de la langue qui servent à caractériser les énoncés complets, à savoir :

– d'une part, l'attribut *lang* — cf. § 5.1 ;
– d'autre part, l'attribut *nonnative* — cf. § 5.3 ;

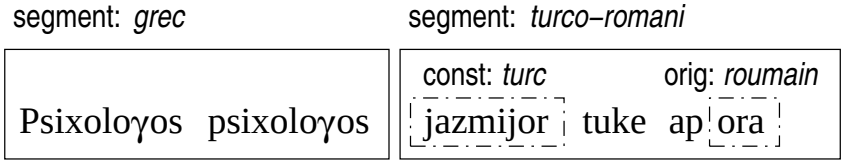
– enfin (facultativement), une liste de langues (*langues*), en cas de segment à rattachement ambigü — cf. § 5.2.

4.1.2 Constituants d'un sous-système linguistique

Dans le cas particulier de systèmes linguistiques mixtes de façon stable (langues mixtes), l'annotateur peut parfois souhaiter annoter de manière séparée les langues constitutives du mélange d'une part, et les décrochages dans d'autres langues d'autre part. À cette fin, ce schéma de documents comporte également un élément *constituant*.

Cet élément s'utilise à peu près de la même manière que l'élément *segment*, défini ci-dessus, à une différence près : le *constituant* ne peut en aucun cas relever de plusieurs langues à la fois, puisqu'il est déjà dans un contexte où deux langues sont potentiellement indiquées, l'une étant dans une relation de composante par rapport à l'autre.

Cet élément peut être utilisé dans le type de contextes illustré ci-dessous.



Exemple de constituant turc et de morphème d'origine roumaine en turco-romani (d'après Adamou, 2009).

Ici, on voit un segment grec (à gauche) dans un énoncé turco-romani. Par ailleurs, au sein du turco-romani, l'annotateur marque de façon spécifique un élément constituant turc (cadre pointillé gauche) et un morphème d'origine roumaine (cadre pointillé droit).

4.1.3 Indications d'origine

Par ailleurs, dans tous types de corpus, on peut souhaiter noter des indications sur les origines de certains éléments de la langue. Cette notation n'est pas obligatoire (sinon elle devrait être exhaustive, ce qui est impossible) ; mais la possibilité existe, sous la forme d'un élément *origine*, dont le fonctionnement est également illustré par la figure ci-dessus.

4.2 Indications paraverbales ou non-verbales, indications linguistiques

Le schéma de documents offre la possibilité de noter des indications paraverbales ou non-verbales. Conformément aux recommandations de la norme TEI (et à la suggestion de Sophie Alby), il distingue en outre :

– les *indications paraverbales* (non-linguistiques), qui explicitent des éléments de corpus liés à des événements situationnels, sans que ces éléments soient des fragments de langue (p.ex. *« (rires) »*, *« (A siffle) »*, ou *« (X se tourne vers Y) »*) ;
– les *indications linguistiques*, dont le rôle est de préciser certaines caractéristiques (notamment prosodiques) de la parole transcrite (p.ex. *« (en parlant plus fort) »*, *« (en accélérant le débit) »*, ou *« (sur un ton monotone) »*.

N.B. Ces deux types d'annotation recouvrent les éléments dénommés *incident*, *kinesic*, *vocal* d'une part, et *shift* d'autre part, dans les propositions du guide TEI, chapitre 8[12].

Les indications paraverbales ou linguistiques peuvent figurer dans le texte de l'énoncé lui-même, ainsi que dans la traduction juxtalinéaire. Elles ne doivent pas en revanche être insérées dans la traduction libre, qui n'a pas forcément la même structure syntagmatique que l'énoncé.

4.2.1 Indications paraverbales

Les indications paraverbales sont subdivisées à titre indicatif en trois sous-catégories : *incident*, *gestuelle* (angl. *kinesic*), et *vocal*. La distinction entre ces trois catégories est entendue comme suit :

- Un événement *vocal* est un événement audible — bien que non-linguistique — produit avec la bouche. Exemples : rire ; soupir ; tchîp ; grognement.
  Bien que non-linguistique, l'événement « vocal » a un contenu sémiotique.
  N.B. Si le bruit émis avec la bouche est articulé et qu'il est possible de le transcrire (ex. « tchip tchip tchip tchip tchibedouwaa »), on doit choisir d'utiliser plutôt l'élément *segment* (§ 4.1.1), en utilisant le code `zxx` (contenu non-linguistique) comme valeur de l'attribut de langue. L'élément *événement vocal* est plutôt fait pour les bruits qu'il est impossible de *transcrire* (comme une suite de phonèmes), et qu'il vaut mieux *décrire* (cf. exemples ci-dessus).
- un événement *gestuel* a également un contenu sémiotique, mais passe par une autre modalité-substrat que la langue. Exemples : haussement d'épaule ; regard en direction d'un interlocuteur ; regard vers le haut ; geste de la main.
- enfin, un *incident* est un événement qui mérite d'être noté dans la transcription du corpus car il peut avoir une incidence sur la conversation, ou être mentionné dans le contexte, mais qui n'a pas *a priori* de contenu sémiotique. Exemples : un nouvel interlocuteur entre ; une porte claque ; un coup de tonnerre éclate.

Les catégories ne sont bien sûr pas forcément étanches et exhaustives, mais elles fournissent une première catégorisation grossière.

Par ailleurs, la nature exacte de chaque événement paraverbal est ensuite indiquée avec précision dans le champ de l'élément : « A entre » ; « B hausse les épaules », « C siffle », etc.

4.2.2 Indications linguistiques

Au contraire, les indications linguistiques (qui commentent, dans la transcription, certains paramètres du flot de parole), ne contiennent pas de description supplémentaire, en « texte libre », du phénomène qu'ils doivent transcrire. Ils sont étiquetés suivant une typologie pré-établie de quelques paramètres de parole de base, et des valeurs que ceux-ci peuvent prendre. Cette typologie est présentée dans le tableau 2.

Dans la transcription d'un énoncé, un élément d'indication linguistique est inséré à l'endroit où le phénomène commence à se produire (par exemple à l'endroit où le locuteur commence à parler plus fort). À l'endroit où le phénomène s'arrête (où la variable prosodique modifiée retourne à une valeur normale), on insère un autre élément du même type, mais vide.

Par exemple, une intonation montante en fin de phrase pourra être repérée en encadrant le dernier segment de la phrase (celui où le phénomène d'intonation montante est constaté) d'une

[12] TEI, *Guidelines for Electronic Text Encoding and Interchange*, chap. 8 : « Transcriptions of speech ». (`http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf`)

Phénomène linguistique			
Variable modifiée	Nouvelle valeur	Code utilisé en interne (XML)	Explicitation
débit	rapide *(allegro)*	tempo=a	plus rapidement
	très rapide	tempo=aa	beaucoup plus rapidement
	accélérant	tempo=acc	de plus en plus vite
	lent *(lento)*	tempo=l	plus lentement
	très lent	tempo=ll	beaucoup plus lentement
	ralentissant	tempo=acc	de plus en plus lentement
volume	fort *(forte)*	loud=f	plus fort
	très fort	loud=ff	beaucoup plus fort
	crescendo	loud=cresc	de plus en plus fort
	bas *(piano)*	loud=p	plus bas
	très bas	loud=pp	beaucoup plus bas
	diminuendo	loud=dimin	de plus en plus bas
hauteur	haut	pitch=high	plus aigü
	bas	pitch=low	plus grave
	étendu	pitch=wide	grandes variations de hauteur de voix
	étroit	pitch=narrow	faibles variations de hauteur de voix
	ascendant	pitch=asc	de plus en plus aigü
	descendant	pitch=desc	de plus en plus grave
	monotone	pitch=monot	voix monotone
articulation	relâchée	tension=sl	relâché
	peu tendue	tension=lax	un peu relâché / peu tendu
	tendue	tension=ten	tendu
	très tendue	tension=pr	précis / très tendu
	staccato	tension=st	accent renforcé sur les syllabes toniques
	legato	tension=leg	accent dilué sur l'ensemble des syllabes
rythme	régulier	rhythm=rh	en marquant un rythme régulier
	irrégulier	rhythm=arrh	avec un rythme particulièrement irrégulier
qualité de voix	chuchotement	voice=whisp	en chuchotant
	halètement	voice=breath	en haletant / en respirant fort
	rauque	voice=husky	voix rauque
	éraillée	voice=creaky	voix éraillée
	fausset	voice=fals	voix de fausset
	résonnante	voice=reson	voix résonnante
	gloussement	voice=giggle	en gloussant
	rire	voice=laugh	en riant
	trémolos	voice=trem	trémolos dans la voix
	sanglot	voice=sob	en sanglotant
	bâillement	voice=yawn	en bâillant
	soupirs	voice=sigh	en soupirant

TAB. 2 – Typologie complète des indications linguistiques

balise portant l'attribut `pitch=asc`, et d'une balise de retour à la normale (sans attribut). Pour noter une intonation descendante, il suffirait de remplacer l'attribut `pitch=asc` par l'attribut `pitch=desc`.

4.3 Passages remarquables

Il est possible d'insérer, à certains points d'un corpus, une annotation générique destinée à signaler un passage remarquable, c'est à dire intéressant d'un point de vue linguistique. Ce type d'annotation est destiné à ce que le linguiste (qu'il soit l'éditeur du corpus ou un autre) puisse retrouver rapidement ces passages lors d'une recherche sur le corpus, par exemple dans le but d'y consacrer une étude plus approfondie.

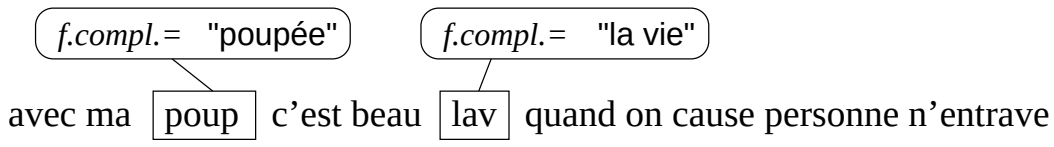
À cette fin, le schéma de documents CORPUS-Contacts définit un élément *passage remarquable*. Celui-ci est accompagné d'un seul attribut : *phénomène*, destiné à décrire en quelques mots quel phénomène linguistique particulier est ici considéré comme remarquable. La diversité de ce que les uns et les autres peuvent être amenés à trouver intéressant étant infinie, cet attribut n'est pas normalisé, et consiste en une chaîne de caractères libre : on peut y mettre des indications de toute nature, comme : « langage de jeunes », « lexème anglais avec désinence nominale russe », « registre de vocabulaire inhabituel dans ce contexte » ...

Une annotation de passage remarquable peut être insérée dans le texte de l'énoncé et dans la traduction juxtalinéaire. Elle ne peut figurer en revanche dans la traduction libre.

Les *passages remarquables* ne peuvent être imbriqués les uns dans les autres.

4.4 Réalisations partielles

Lorsqu'un locuteur ne réalise pas de manière complète la forme linguistique considérée comme normale dans la langue utilisée, on peut le noter dans le corpus à l'aide d'un élément *réalisation partielle*. Cet élément encadre (par une balise de début et une balise de fin) la forme considérée comme tronquée, et doit être caractérisé par un attribut, *forme complète*, qui précise quelle est la forme standard, non-tronquée, de l'unité réalisée.



Exemple de réalisations partielles.

L'indication de réalisation partielle ne peut être insérée que dans le texte original de l'énoncé (pas dans une traduction).

4.5 Transcription phonétique

Dans de nombreux corpus, la langue orale est transcrite en suivant les conventions orthographiques normalisées de la langue écrite. Si, au milieu d'une transcription de ce type, on trouve une forme qui est bien une séquence de phonèmes, mais que l'on est incapable de transcrire selon les conventions orthographiques habituelles de la langue écrite (soit que cette forme ne soit pas répertoriée dans un dictionnaire de mots connus, soit que sa transcription

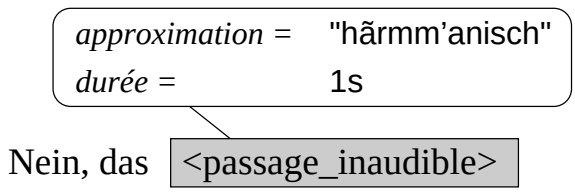
orthographique soit incertaine), on peut l'indiquer en utilisant l'élément *transcription phonétique*. Cet élément (qui encadre la chaîne identifiée par une balise de début et une balise de fin) signale simplement que l'on a transcrit ce que l'on a entendu « comme ça se prononce ».

L'élément *transciption phonétique* peut être utilisé dans le texte de l'énoncé, mais aussi dans la traduction : en effet, si l'on transcrit phonétiquement un segment parce qu'on ne l'a pas identifié, on peut souhaiter remettre la transcription phonétique telle quelle dans la traduction ; alors que si on l'a identifié et qu'on ne sait juste pas comment il faut l'écrire, on n'a pas besoin de le transcrire phonétiquement dans la traduction, puisqu'on peut le traduire.

4.6 Passage inaudible

Un passage inaudible dans l'enregistrement peut être rendu dans le document par un élément *passage inaudible*. Cet élément a deux attributs facultatifs :

- *approximation* : on peut éventuellement, si l'on souhaite donner une idée approximative de ce que l'on entend (vaguement) dans le passage en question, insérer quelque chose dans le champ « approximation » (ex. ci-dessous).
- *durée* : peut être utilisé pour indiquer la durée du passage inaudible.



Le passage inaudible peut être signalé aussi bien dans le texte que dans les traductions.

4.7 Pause

On peut noter une pause dans le flot de parole grâce à l'élément *pause*. Celui-ci est doté de deux attributs possibles : *type de pause* et *durée*. Ces deux attributs ont plus ou moins la même fonction (indiquer la durée de la pause), mais le font de manière différente : *type de pause* est un attribut à trois valeurs possibles permettant d'indiquer — à la manière classique des corpus oraux transcrits — s'il s'agit d'une pause courte, moyenne, ou longue ; *durée* permet d'indiquer une durée en secondes.

La pause peut être notée dans le texte et dans sa traduction juxtalinéaire.

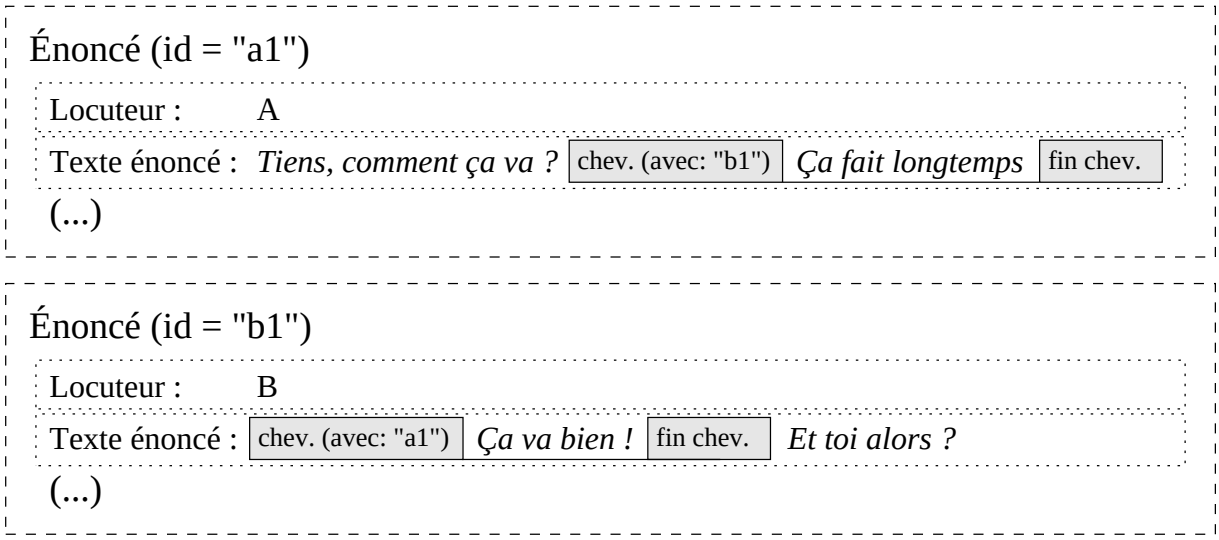
4.8 Chevauchement

Il est possible de faire figurer dans un fichier corpus l'indication que deux énoncés se chevauchent. Cette indication est donnée par l'élément *chevauchement*, qui doit délimiter (par une balise de début d'un côté, et une balise de fin de l'autre) le segment concerné par ce chevauchement.

Par définition, un chevauchement concerne au moins deux énoncés. Afin de pouvoir retrouver les autres morceaux d'énoncés concernés par ce chevauchement, on doit indiquer un point de référence temporelle. Les autres segments d'énoncé qui ont été prononcés en même temps sont eux aussi délimités par un élément *chevauchement*, qui pointe vers la même référence temporelle.

Un chevauchement concerne par définition au minimum *deux* énoncés ; il est donc également prévu de pouvoir indiquer *avec quel autre* énoncé l'énoncé en cours est en situation de

chevauchement. À cette fin, l'élément *chevauchement* possède un attribut *avec*, dont la valeur doit être l'identifiant de l'autre énoncé concerné (son attribut *id* — cf. plus haut, § 3.6.2).



Exemple d'indication de chevauchement.

Le chevauchement ne concerne bien entendu que le texte de l'énoncé lui-même, pas la traduction. L'utilisation de l'élément *chevauchement* est donc restreinte au contexte du texte original de l'énoncé. Par ailleurs, les chevauchements ne peuvent pas être imbriqués les uns dans les autres.

4.9 Marque d'incomplétude

Dans la traduction libre, on peut indiquer que l'énoncé en cours de traduction est incomplet. Il faut pour cela insérer un élément *marque incomplétude*.

4.10 Notes

On peut insérer une note dans le texte d'un corpus, en tout point du texte d'un énoncé, de sa traduction juxtalinéaire, ou de sa traduction libre (tout dépend du contexte et de ce que l'on souhaire commenter).

Le schéma de documents CORPUS-Contacts met pour cela à disposition un élément *note*. Cet élément délimite, par une balise de début et une balise de fin, le texte de la note.

La note est insérée, dans le fichier corpus formaté en XML, au point exact du texte auquel elle fait référence, ainsi que l'illustre l'exemple ci-dessous :

```
<enonce>
<identification_du_locuteur>kuliyaman</identification_du_locuteur>
<texte_enonce lang="way">Malonme, tupaphe ihpotïnpï amohawinpï imukutpëhtak,
kuliputpë pupokatpë</texte_enonce>
<traduction_libre lang="fra">Puis ils ont jeté ses écailles et ses griffes au
dépotoir, les restes de la tortue<note>Rappelons rapidement les faits : la mère
des enfants, Salumakani, cheminait vers Kuyuli - qui l'avait magiquement fécondée
- sous forme de tortue. Elle se trompe de chemin, et les Gens des fauves la capturent
et la dévorent. Seuls ses enfants seront sauvés (ils sont encore dans ses œufs)
```

```
grâce à la grand-mère
crapaud.</note>.</traduction_libre>
</enonce>
```

Il appartient éventuellement au logiciel utilisé pour la visualisation du corpus de présenter la note de manière plus lisible pour un utilisateur humain, en la séparant d'une façon ou d'une autre du flot du texte. Par exemple, la feuille de transformations XSLT "`corpus.xsl`" (qui permet à un logiciel navigateur récent de visualiser le XML sous forme de HTML engendré automatiquement) prévoit de n'insérer au niveau de l'appel de notes qu'un petit chiffre en exposant (déterminé par numérotation automatique), et de regrouper le texte de toutes les notes, numérotées, en bas de chaque texte du corpus, comme illustré dans la figure 8.

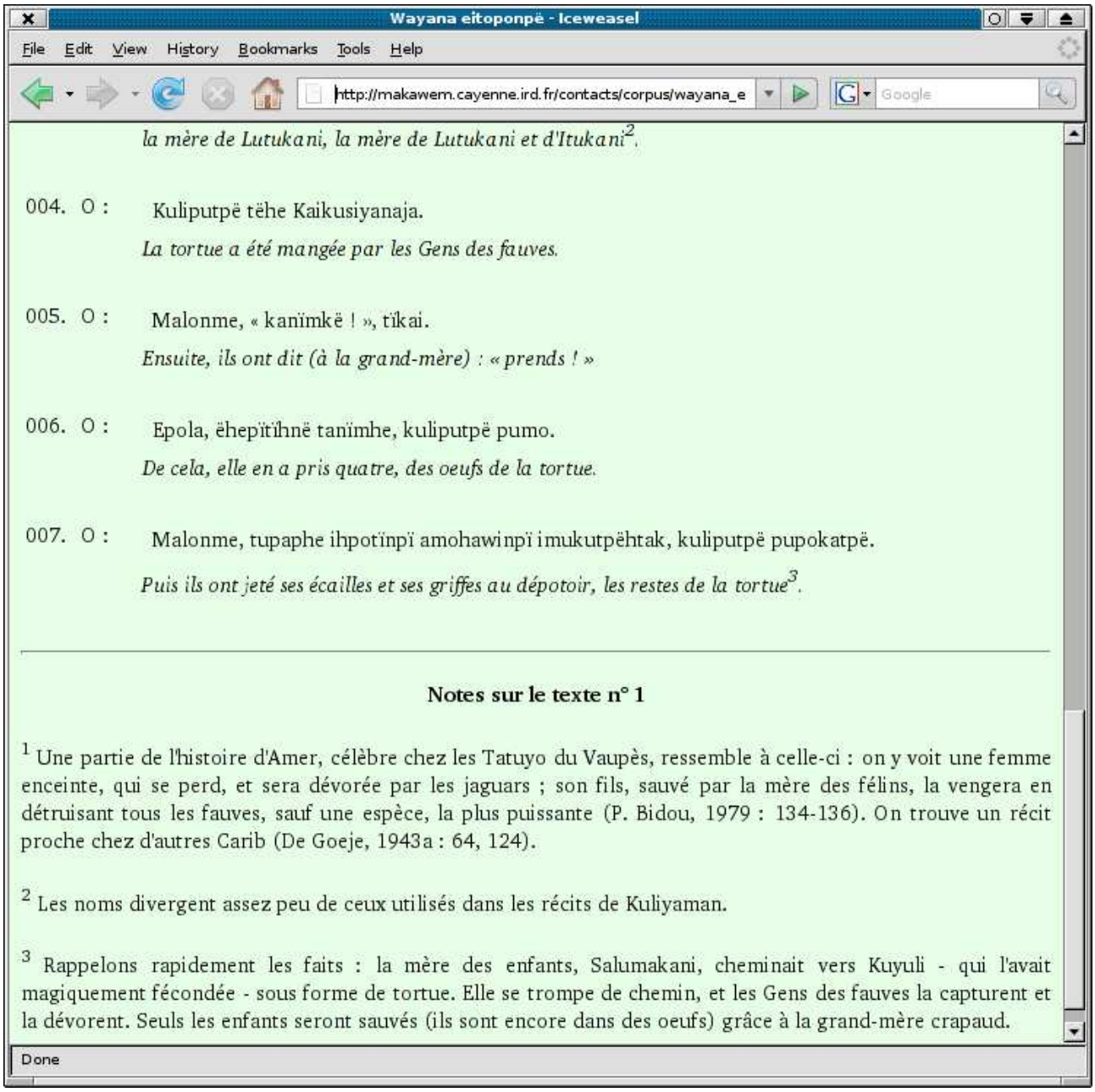


FIG. 8 – Présentation des notes

Une note ne peut pas être imbriquée dans une autre note ; en revanche, une note peut *faire référence* à une autre note (cf. ci-dessous, § 4.10.1).

Une note n'a pas vocation à contenir du texte de corpus : on ne peut donc pas y insérer des éléments de description linguistique comme ceux décrits jusqu'ici dans la section 4 de ce document. En revanche, elle est là pour fournir des éléments de description, d'analyse, ou de comparaison, et il est donc pertinent qu'elle puisse faire référence à d'autres points du texte (énoncés) ou à d'autres notes.

Si l'on souhaite faire référence à une note, il est nécessaire de donner une valeur à son attribut *id* (cf. plus haut, § 3.6.2).

4.10.1 Référence à une note

Pour faire référence à une note au sein d'une autre note, il suffit alors d'insérer un élément *référence à une note*, en précisant l'identifiant (*id*) de la note vers laquelle on veut pointer, dans l'attribut *réf. note*.

L'éditeur du corpus n'a pas à se préoccuper du « numéro » de la note. Les numéros sont calculés automatiquement au moment de la présentation visuelle finale : ils appartiennent à cette présentation visuelle, mais pas au fichier corpus lui-même. Dans le fichier XML au format de document CORPUS-Contacts, les notes n'ont pas de numéros, elles n'ont que des identifiants (facultatifs) contenus dans l'attribut *id*.

Ainsi, si la troisième note fait référence à la deuxième, mais que par la suite, en poursuivant le travail d'édition du corpus, on décide d'insérer une nouvelle note entre ces deux notes-là, puis encore une nouvelle note au début, l'éditeur n'a pas à se préoccuper de re-numérotation. La note anciennement troisième, qui est devenue cinquième, fait toujours référence à la note anciennement deuxième, qui est devenue troisième. Elle y fait référence, au sein du fichier corpus, par le truchement de son identifiant, qui, lui, n'a pas changé. Dans la présentation visuelle obtenue par l'intermédiaire d'un agent de visualisation (navigateur) et de la feuille de transformations `corpus.xsl`, le numéro sera automatiquement recalculé, et l'utilisateur verra apparaître un chiffre 3 au lieu d'un chiffre 2 en exposant.

4.10.2 Référence à un énoncé

De la même manière, au sein d'une note, on peut faire référence à un énoncé du corpus en insérant un élément *référence à un énoncé*, et en précisant l'identifiant (*id*) de cet énoncé dans l'attribut *réf. énoncé*.

N.B. Les remarques qui ont été faites ci-dessus sur la numérotation des notes valent également pour les énoncés. Les énoncés n'ont pas de numéro au sein d'un fichier XML CORPUS-Contacts. La numérotation est faite automatiquement au moment de l'affichage (et donc, recalculée systématiquement, ce qui permet de ne pas se soucier de l'ordre d'édition de différentes parties du corpus).

4.11 Répartition des éléments de description linguistique

Le tableau 3 résume les contextes dans lesquels les différents éléments de description, énumérés dans la section 4 ci-dessus, peuvent apparaître. Il donne leur répartition dans les quatre contextes possibles (qui apparaissent en colonnes) : dans le texte de l'énoncé transcrit ; dans la traduction juxtalinéaire ; dans la traduction libre ; et dans le texte des notes (la note étant elle-même comptée comme un élément).

Dans ce tableau (comme dans le suivant), un signe *plus* ("+") indique que l'élément mentionné sur la ligne considérée peut apparaître dans le contexte de la colonne où figure ce signe. Le signe *moins* ("-") indique le contraire.

	TEXTE	TRAD JUXTALINÉAIRE	TRAD LIBRE	NOTE
segment	+	−	−	−
passage remarquable	+	+	−	−
chevauchement	+	−	−	−
indication paraverbale	+	+	−	−
indication linguistique	+	+	−	−
réalisation partielle	+	−	−	−
transcription phonétique	+	+	+	+
passage inaudible	+	+	+	−
pause	+	+	−	−
point tabulation	+	+	−	−
note	+	+	+	−
marque incomplétude	−	−	+	−
reference à un énoncé	−	−	−	+
reference à une note	−	−	−	+

TAB. 3 – Répartition des éléments de description linguistique dans les différents contextes possibles

Le tableau 4 résume les enchâssements possibles entre les différents éléments.

5 Attributs d'identification de la langue

Dans les situations de mélanges de langues, il n'est pas toujours possible d'identifier de manière univoque *la* langue d'un énoncé ou d'un segment d'énoncé. Nous proposons donc ici, dans le schéma de documents CORPUS-Contacts, un système qui permet d'attribuer *plusieurs* langues à un passage (avec un ordre de dominance/vraisemblance).

Le système proposé, avec cette extension, est rétro-compatible avec la méthode « traditionnelle » d'identification de la langue des textes (méthode de l'attribut *lang*, recommandée par l'IEFC dans le monde XML/HTML, ainsi que par la TEI). Il permet de préciser plusieurs langues lorsque cela est nécessaire, et de s'en tenir à l'attribution à une seule langue dans le cas général.

Par ailleurs, dans le cas d'énoncés à rattachement multiple, un agent logiciel qui ne reconnaît pas cette extension CORPUS-Contacts pourra se rabattre sur la première langue, qui est celle qui sera indiquée dans l'attribut *lang* (méthode classique).

Nous allons commencer ci-dessous par exposer la méthode générale d'identification de la langue d'un énoncé (méthode utilisée lorsqu'il y a une seule langue : § 5.1), ainsi que les normes de représentation utilisées. Nous donnerons l'exemple des différentes langues de Guyane et du Mexique pour illustrer ces principes.

Ensuite, nous expliquons la structure de l'élément d'information utilisé pour coder le rattachement simultané d'un passsage à plusieurs langues (§ 5.2).

	segment	constituant	origine	passage remarquable	chevauchement	indication paraverbale	réalisation partielle	transcription phonétique	note
segment	−	−	−	+	+	−	−	−	−
constituant	+	+	−	+	+	−	−	+	−
origine	+	+	−	+	+	−	−	+	−
passage remarquable	+	+	+	−	+	−	−	+	−
chevauchement	+	+	+	+	−	−	−	+	−
indication paraverbale	+	+	+	+	+	−	−	+	−
indication linguistique*	+	+	+	+	+	−	−	+	−
réalisation partielle	+	+	+	+	+	−	−	+	−
transcription phonétique	+	+	+	+	+	−	−	+	+
passage inaudible*	+	+	+	+	+	−	−	+	−
pause*	+	+	+	+	+	−	−	+	−
point tabulation*	+	+	+	+	+	−	−	+	−
note	+	+	+	+	+	−	−	+	−
marque incomplétude*	−	−	−	−	−	−	−	−	−
référence à un énoncé*	−	−	−	−	−	−	−	−	+
référence à une note*	−	−	−	−	−	−	−	−	+

** Les éléments marqués d'une astérisque sont « vides » : ils ne sont donc pas susceptibles de contenir d'autres éléments enchâssés (ils sont constitués d'une balise unique, pas d'une balise de début et d'une balise de fin avec du texte entre les deux). Ils ne figurent donc pas en colonne.*

TAB. 4 – Répartition des éléments de description linguistique dans les différents contextes possibles

5.1 L'attribut *xml:lang*

Un attribut, *lang*, sert à identifier la langue d'un segment linguistique, soit au niveau d'un énoncé entier, soit au niveau d'un fragment d'énoncé (élément *segment*).

Suivant les recommandations de la TEI, la valeur de l'attribut *lang* est déterminée selon la norme en usage sur internet[13], et codifiée par l' *"Internet Society"* sous la référence RFC-5646[14].

La norme RFC-5646 prévoit que la langue proprement dite est codée par une étiquette tirée de l'une des versions de la norme ISO-639. La version 1 de cette norme (ISO-639-1) comprend des codes à deux lettres utilisés pour les langues les plus courantes (`'fr'` pour le français, `'en'` pour l'anglais, `'de'` pour l'allemand ...) La version 2 (ISO-639-2[15]) comprend des codes à deux et trois lettres, mais son répertoire est assez limité[16]. Nous utilisons donc concrètement la version la plus étendue, centralisée par le SIL, l'ISO 639-3[17], qui a pour vocation d'attribuer un code à trois lettres à toutes les langues connues.

La plupart des langues ont donc un codage sous la forme d'une étiquette à trois lettres, comme `'fra'` pour le français ou `'eng'` pour l'anglais.

Il existe par ailleurs trois étiquettes spéciales :

- `'mul'` (« *multiple languages* ») : pour les passages contenant plusieurs langues à la fois ;
- `'und'` (« *undetermined* ») : pour les passages dont on n'a pas réussi à identifier la langue ;
- `'zxx'` : pour les passages de contenu articulé mais non-linguistique (ex. « chouba douba douwa »).

Les recommandations du RFC-5646 prévoient en outre la possibilité d'ajouter des précisions à l'étiquette de langue. Les précisions (facultatives) peuvent concerner :

1. une indication de variante ou de dialecte, codée par une étiquette de 5 à 8 lettres (par exemple, *hye*[18] est utilisé pour l'arménien en général, et *hye-arevela* pour l'arménien oriental ;
2. une indication de système d'écriture[19] (par exemble, *srp-Cyrl* est utilisé pour désigner du serbe écrit en alphabet cyrillique, et *srp-Latn* pour du serbe écrit en alphabet latin) ;
3. une indication d'aire géographique, servant à préciser qu'on souhaite identifier une variante régionale d'une langue de grande extension (par exemple, *en-US* pour l'anglais des États-Unis, et *en-GB* pour l'anglais de Grande-Bretagne[20].

[13] Cette norme est employée par exemple pour caractériser la langue utilisée par un site web.

[14] *Internet Engineering Task Force,* RFC *(Request For Comments)* 5646 : *Tags for Identifying Languages.* URL : `http://tools.ietf.org/html/rfc5646`.

[15] ISO *(International Standards Organization)* standard 639-2 : *Codes for the representation of names of languages– Part 2.* URL : http ://www.loc.gov/standards/iso639-2/.

[16] Sa communauté d'utilisateurs est constituée essentiellement de documentalistes, et son usage est donc principalement orienté vers les langues de l'édition.

[17] ISO *(International Standards Organization)* standard 639-3 : *Codes for the representation of names of languages– Part 3.* C'est le SIL *(Summer Institute of Linguistics)* qui a été désigné comme organisme centralisateur de cette norme. URL : http ://www.sil.org/iso639-3/.

[18] De *hayerēn* (qui signifie « arménien » en arménien).

[19] Les étiquettes utilisables pour désigner les systèmes d'écriture font également l'objet d'une norme, qui représente chacun des systèmes d'écriture les plus répandus par une chaîne de quatre caractères. Cette norme est la norme ISO 15924 — URL : `http://unicode.org/iso15924/iso15924-codes.html`.

[20] La norme RFC 4646 prévoit que l'étiquette utilisée pour dénoter une extension géographique soit, dans le cas typique, tirée de la liste des codes de noms de pays établie par la norme ISO 3166 (`http://www.iso.ch/iso/fr/country_codes/iso_3166_code_lists/french_country_names_and_code_elements.htm`). Cependant, on peut également souhaiter dénoter une aire géographique ne correspondant pas à un pays, et il est alors possible d'utiliser les codes numériques désignant des zones géographiques du monde utilisés par la division des statistiques de l'ONU (M 49 : *Codage statistique normalisé des pays et zones* — URL :

L'institution chargée de coordonner les conventions techniques régissant le fonctionnement d'internet, l'IANA[21], a également pour objectif de maintenir à jour une table normalisée des étiquettes utilisables pour déterminer l'attribut *lang*. Cette table[22] comprend aussi bien des étiquettes de langue proprement dit (ISO 639), que des étiquettes de variante dialectale (l'exemple de l'arménien oriental, cité plus haut, en est tiré). Cependant, assez peu de ces étiquettes sont normalisées jusqu'à maintenant.

5.1.1 Application : langues de Guyane

La nomenclature qu'il est recommandé d'utiliser pour identifier les langues de Guyane[23] suit les différentes recommandations et normes, en y ajoutant le cas échéant des étiquettes non-encore normalisées pour certaines variantes. Dans le tableau suivant, la colonne de gauche présente une taxonomie (hiérarchique) des langues de Guyane, tandis que la colonne de droite donne le code utilisé en interne à l'attribut *lang*, dans les documents XML, pour représenter chaque cas.

Remarques concernant le tableau 5 :

1. « Kali'na » (car) : parfois également *Galibi* (dans l'usage francophone), mot de même origine mais déformé par l'emprunt.
2. « Arawak » (arw) : parfois *Lokono*, pour éviter les ambiguïtés sur la portée du terme (langue vs. famille de langues).
3. « Émerillon » (eme) : parfois *Teko*.
4. La caractérisation « Français de Guyane » (au lieu de « Français » tout court) peut éventuellement servir dans les contextes où il est pertinent de préciser qu'il s'agit d'une variété régionale de français.
5. « Espagnol d'Amérique latine ou de la Caraïbe » (spa-419) : le code 419 désigne conventionnellement toute l'Amérique latine, et pas seulement l'Amérique du Sud ; ceci inclut les îles hispanophones de la Caraïbe (comme le nom l'indique explicitement) mais également la Méso-Amérique, Mexique inclus. Cette étiquette désigne donc toute variété d'Espagnol américain.
6. « Espagnol de la Caraïbe » (spa-029) : espagnol des îles hispanophones de la Caraïbe (Cuba, République Dominicaine ...).
7. « Créole français » (cpf-019) : `cpf` est une étiquette générique pour tous les créoles ou pidgins à base française ; elle appartient à la nomenclature d'ISO 639-2 (et pas d'ISO 639-3, qui a ensuite précisé ce domaine). Utilisée avec le code géographique pour tout l'hémisphère américain (019), elle sert à désigner ici, sans plus d'autre précision, une variété de créole à base française de l'aire Atlantique-Caraïbe (aire créole de l'arc Antillais, de la Louisiane à la Guyane en passant par Haïti et les Petites Antilles).
8. « Créole anglais » (cpe-019) : `cpe` est une étiquette générique pour tous les créoles ou pidgins à base anglaise ; on en fait donc un usage analogue à `cpf` (ci-dessus), en en restreignant l'extension géographique par un suffixe désignant la zone américaine.

`http://unstats.un.org/unsd/methods/m49/m49regnf.htm`). Par exemple, *spa-419* peut être utilisé pour désigner globalement l'espagnol d'Amérique latine et des Caraïbes

[21] IANA *(Internet Assigned Numbers Authority)*.

[22] IANA *Language subtags regitry*. URL : `http://www.iana.org/assignments/language-subtag-registry`.

[23] Nous recensons comme « langues de Guyane » les langues parlées par une communauté significative établie de façon stable en Guyane — l'interprétation du premier critère dépendant bien sûr du second. Ainsi, l'arawak est comptabilisé même s'il y a vraisemblablement aujourd'hui moins de locuteurs arawak en Guyane que d'ingénieurs et de techniciens russes ; mais la présence de russophones en Guyane est attestée depuis moins longtemps.

Langue			
Généricité	Nom	Code	Mnémonique
	Langues amérindiennes		
Langue	Kali'na	car	(CARib)
Langue	Wayana	way	(WAYana)
Langue	Apalaí	apy	(APalaY)
Langue	Arawak	arw	(ARaWak)
Langue	Palikur	plu	(PaLikUr)
Langue	Wayampi	oym	(OYaMpi)
Langue	Émerillon	eme	(EMErillon)
	Langues européennes		
Langue	Français	fra	(FRAnçais)
Variante	*Français de Guyane*	fra-GF	
Langue	Anglais	eng	(ENGlish)
Variante	*Anglais du Guyana ou de la Caraïbe*	eng-419	
Variante	*Anglais du Guyana*	eng-GY	
Langue	Portugais	por	(PORtuguês)
Variante	*Portugais du Brésil*	por-BR	
Langue	Espagnol	spa	(eSPAñol)
Variante	*Espagnol d'Amérique Latine ou de la Caraïbe*	spa-419	
Variante	*Espagnol de la Caraïbe*	spa-029	
Variante	*Espagnol d'Amérique du Sud*	spa-005	
Langue	Néerlandais	nld	(NederLanDs)
Variante	*Néerlandais du Suriname*	nld-SR	
	Langues créoles		
Macro-langue	**Créole français**	cpf-019	(Creoles and Pidgins, French)
Langue	Haïtien	hat	(créole HAïTien)
Langue	Créole des Antilles Françaises	acf	(Antillean Creole French)
Variante	*Créole guadeloupéen*	acf-GP	
Variante	*Créole martiniquais*	acf-MQ	
Variante	*Créole saint-lucien*	acf-LC	
Langue	Créole guyanais	gcr	(french Guiana CReole)
Macro-langue	**Créole anglais**	cpe-019	(Creoles et Pidgins, English)
Langue	Créole anglais du Guyana	gyn	(GuYaNese creole)
Langue	Sranan	srn	(SRaNan)
Langue	Saamaka	srm	(SaRaMaca)
Langue	Nenge	djk	(DJuKa)
Variante	*Aluku*	djk-aluku	
Variante	*Ndyuka*	djk-ndyuka	
Variante	*Pamaka*	djk-pamaka	
	Langues d'Asie		
Langue	Hindustani	hns	(HiNduStani))
Langue	Javanais	jvn	(JaVaNese))
Macro-langue	**Hmong**	hmn	(HMoNg)
Langue	Hmong Daw	mww	(Miao White / hmong daW)
Langue	Hmong Njua	blu	(miao BLUe)
Macro-langue	**Chinois**	zho	(ZHOng wen)
Langue	Mandarin	cmn	(Chinois MaNdarin)
Langue	Cantonais	yue	(YUE yu)
Langue	Hakka	hak	(HAKka)

TAB. 5 – Liste des codes normalisés pour identifier les langues de Guyane

9. « Saamaka » (srm) : parfois *Saramaca*, dans l'usage courant en Guyane Française.
10. « Nenge » (djk) : désigne une variété de *businenge tongo*, sans plus de précision : ndyuka, aluku, pamaka, voire toute autre variété mixte, urbanisée, koinèïsée, approximative ... ou tout simplement non-identifiable de manière plus spécifique. L'étiquette `djk` peut induire en erreur : jusqu'à présent, l'ISO 639-3 (qui hérite cela du catalogue *Ethnologue*) désigne la langue dans son ensemble en référence à sa variété la plus parlée, le ndyuka[24]. Mais cette étiquette `djk` n'est qu'un code interne (une interface de saisie peut faire figurer le mot « nenge »).
11. « Aluku » (djk-aluku) : parfois *Boni*, dans l'usage courant en Guyane Française[25].
12. « Ndyuka » (djk-ndyuka) : parfois également *Aukan*, au Suriname.
13. « Pamaka » (djk-pamaka) : parfois *Paramaca*, dans l'usage courant en Guyane Française.
14. « Hindustani » (hns) : ce code est spécifiquement conçu pour désigner la langue de la communauté d'origine hindoustanie (Inde du Nord-Ouest) sur la côte des Guyanes et à Trinidad et Tobago. Les évolutions récentes dans le sous-continent indien (ourdou vs. hindi, arabisation vs. « sanscritisation » du lexique ...) ne sont pas pertinentes dans le contexte de la diaspora américano-caribéenne. Appellation familière : « la langue de Bollywood ».
15. « Javanais » (jvn) : ce code est spécifiquement conçu pour désigner la langue de la communauté d'origine indonésienne installée au Suriname. Là encore, les évolutions d'après l'indépendance de l'indonésie (création d'une langue construite — le *bahasa indonesia* — comme langue nationale ...) ne sont pas pertinentes.
16. « Hmong » (hmn), « Hmong Daw » (mww), et « Hmong Njua » (blu) : la grande majorité des Hmongs de Guyane (comme d'ailleurs plus généralement ceux de la diaspora) parlent le dialecte Daw — également appelé *hmong blanc* — qui a une situation méridionale dans la zone d'extension des langues Hmong en Asie du Sud (plus méridional — Laos et Vietnam — signifiant en l'occurrence plus éloigné du foyer d'origine). Par ailleurs, moins d'une dizaine de familles parlent le dialecte Njua[26] — également appelé *hmong vert*, ou parfois *hmong bleu* —, qui, parmi les dialectes plus septentrionaux (c'est-à-dire plus proches du foyer d'expansion des Hmongs en Chine), est celui qui est le plus proche du dialecte Daw. Les deux langues sont intercompréhensibles. Au cas où la distinction ne soit pas possible (fragments trop brefs) ou pas pertinente, on peut étiqueter au niveau de la macro-langue, le *Hmong*.
17. « Chinois mandarin » (cmn) : il ne s'agit pas de la langue lettrée, mais de la langue nationale de RPC *(Hànyǔ*, ou *Pǔtōnghuà)*, répandue chez les dernières générations de Chinois arrivés en Guyane.
18. « Chinois cantonais » (yue) : *Yuèyǔ*, langue du Sud (de la province de Canton, *Yuè*).

Les codes à trois lettres utilisés (norme ISO 639-3) ont généralement été donnés par le SIL. Parfois, ils sont des mnémoniques d'un nom qui n'est pas celui qui est généralement utilisé dans

[24]La norme ISO-639-3 donne également le nom d' « Aukan » comme étant le nom correspondant à ce code, ce qui laisse imaginer à tort qu'il ne s'agit que du Ndyuka, à l'exclusion des autres variétés. Cette norme appelée à évoluer sur ce point, suite à une requête que nous avons introduite le 31/08/2009 auprès de l'autorité d'enregistrement (SIL), pour que le nom inscrit pour la langue dans son ensemble soit « Nenge » et non plus « Aukan » (cf. URL : `http://www.alvestrand.no/pipermail/ietf-languages/2009-August/009297.html`).

[25]Par ailleurs, les étiquettes codant les trois variétés « ethniques » du Nenge (`'aluku'`, `'ndyuka'`, et `'pamaka'`) ont été enregistrées dans le répertoire normalisé de l'IANA suite à une requête que nous avons introduite le 23/01/2009 (cf. URL : `http://www.alvestrand.no/pipermail/ietf-languages/2009-January/008780.html`). Elles font maintenant partie du répertoire public d'étiquettes de langues.

[26]Chô Ly, communication personnelle.

les travaux francophones, ni usité en Guyane ou dans la communauté elle-même : par exemple, `'car'` pour *CARib* (et non *Galibi* ou *Kali'na*) ; ou `'djk'` pour *DJuKa* (utilisé pour désigner le *Nenge* dans son ensemble, et pas seulement sa variante *Ndyuka*). Nous les gardons cependant dans l'encodage informatique pour la compatibilité avec la norme et avec les ressources antérieures[27] ; cela n'empêche pas d'utiliser le nom « en clair » — celui de la deuxième colonne — dans l'interface de saisie et de visualisation.

Il est possible de spécifier la valeur de l'attribut *lang* à différents niveaux de généricité. Par exemple, un énoncé en français prononcé par un francophone de Guyane peut également être étiqueté `'fr'` (*Français*) ou `'fr-GF'` (*Français de Guyane*). De même, un énoncé en Hmong peut également être étiqueté `'mww'` (*Hmong Daw*) ou `'hmn'` (*Hmong*, au niveau de la macro-langue) ; et un énoncé en Pamaka peut également être étiqueté `'djk-pamaka'` (*Nenge*, variante *Pamaka*) ou `'djk'` (*Nenge*). Dans chacun des exemples cités, les deux usages sont conformes à la norme.

La décision de fixer le niveau approprié de généricité appartient entièrement à l'éditeur du corpus. Il n'y a pas de règle générale, si ce n'est celle de la pertinence. Au moment de choisir le niveau de généricité, il faut simplement se rappeler que l'information concernant les niveaux supérieurs est automatiquement incluse dans l'information concernant les niveaux les plus spécifiques (par exemple, *Nenge-Ndyuka* implique *Nenge*), alors que l'inverse n'est pas vrai. Il convient donc simplement de songer au degré de précision qui mérite d'être conservé.

Ce principe de base étant énoncé, ensuite, le choix de l'étiquetage dépend entièrement du cas de figure. On pourrait considérer, d'un point de vue privilégiant l'exactitude, qu'il est conseillé de donner toujours la précision maximale (les niveaux plus génériques pouvant de toute façon en être déduits). On peut cependant imaginer des cas de figures où il est justifié d'utiliser un niveau plus générique :

- à moins de vouloir dénoter explicitement l'usage de tournures régionales en français (« il ne faut pas faire les enfants crier, c'est ça j'ai dit à la personne »), il est probablement inutile — voire potentiellement erroné — de spécifier systématiquement *Français de Guyane* pour tout énoncé en Français enregistré en Guyane ;
- dans le cas de variantes intermédiaires, mélangées, koinèisées, approximatives ... il peut être utile de ne pas être obligé de choisir entre des variantes très spécifiquement caractérisées (par exemple, entre *Ndyuka* et *Pamaka*), mais de pouvoir laisser seulement l'indication de langue (*Nenge*) ;
- de même dans le cas de l'étiquetage de segments courts, où la forme manifestée est une forme générique qui, sans contexte supplémentaire, ne permet pas de distinguer entre deux variétés ;
- enfin, la personne qui édite un échantillon de corpus peut elle-même avoir des compétences linguistiques qui lui permettent d'identifier une langue ou une macro-langue survenant dans son échantillon à un niveau générique (par exemple du *Chinois*, ou du *Créole français des Petites Antilles*) mais pas de distinguer entre différentes langues ou variantes à un niveau plus spécifique (par exemple, entre *Mandarin* et *Cantonais*, ou entre *Créole guadeloupéen* et *Créole martiniquais*) ; elle peut alors choisir un niveau générique, qui apporte malgré tout une information pertinente dans l'usage qu'elle fait de son échantillon (par exemple : ici le locuteur s'exprime en chinois, et pas en français, en portugais ou en créole ; il s'adresse donc plus probablement à un membre de sa famille qu'à un client), même s'il n'est pas d'une rigueur suffisante pour la détermination exacte de la langue.

[27] Par exemple le CRDO (Centre de Ressources pour la Description de l'Oral) du LACITO — URL : `http://crdo.risc.cnrs.fr/exist/crdo/`.

Il importe de noter que le schéma CORPUS-Contacts ne *contraint pas* l'attribut *lang*, destiné à l'identification de la langue, à avoir l'une des valeurs mentionnées dans le tableau 5 et seulement l'une de ces valeurs-là. Le tableau n'est donné que pour indiquer les choix préconisés pour les langues qui seront le plus souvent rencontrées dans les corpus du programme Contacts en Guyane. Ceci n'empêche pas par ailleurs le corpus de contenir l'indication d'un énoncé en russe (*lang* = `'rus'`), pour reprendre l'exemple précédemment mentionné, si des locuteurs s'expriment en russe dans un enregistrement.

La table des langues de Guyane peut par ailleurs être utilisée dans l'*interface de saisie* des corpus[28] pour aider l'éditeur à entrer les étiquettes de langues normalisées facilement, sans se référer à chaque fois à la documentation, en faisant le choix, par exemple, dans un menu déroulant proposant l'ensemble des lignes du tableau 5 ; mais sans empêcher pour autant la saisie d'une autre valeur[29].

5.1.2 Langues du Mexique

Le fichier de configuration n'a pas inclus des abréviations pour toutes les langues du Mexique dans la liste des suggestions pour remplir le champ « langue » : cela n'est pas utile dans la configuration actuelle du programme Contacts. Les codes les plus utiles pour l'instant (suggestion de Claudine Chamoreau) sont ceux concernant l'espagnol et le purépecha. L'espagnol (code ISO 639-3 : `'spa'`) peut éventuellement (si la précision semble nécessaire, cf. discussion ci-dessus) être spécifié par une indication géographique. Par ailleurs, pour le purépecha, deux codes sont disponibles, l'un pour les variétés de l'est (bassin du lac de Pátzcuaro), et l'autre pour celles de l'ouest (région des montagnes de la « *Meseta Tarasca* », au Nord-Ouest du Michoacan) (tableau 6).

Cette liste peut être étendue si les besoins de couverture s'étendent. Par ailleurs, là encore, il convient de rappeler que les codes du tableau 6 ne sont que des *suggestions* destinées à être présentées en priorité dans des menus d'interface, mais que ces suggestions ne sont pas contraignantes.

5.2 L'attribution d'un passage à plusieurs langues

Afin de représenter le fait qu'un énoncé ou segment d'énoncé peut être rattaché à plusieurs langues à la fois, on a introduit dans le schéma de documents CORPUS-Contacts un élément (facultatif), appelé *langues*. Cet élément contient une liste de langues, exactement au même format que l'inventaire des langues déclaré au début du corpus (§ 3.2.1) : à savoir une liste d'éléments de type *langue*, contenant chacun simplement dans un attribut *lang* la description standardisée d'une langue (cf. § 5.1).

Cet élément *langues* est facultatif, au contraire de l'attribut principal *lang*, rattaché directement au niveau supérieur (au niveau de l'élément *texte énoncé* ou *segment*). Dans le cas où l'attribution de l'énoncé ou du segment d'énoncé à une langue est univoque, seul l'attribut *lang* est requis.

[28] C'est le choix que nous avons fait dans le fichier de configuration pour l'éditeur XML JAXE, développé spécifiquement pour CORPUS-Contacts : `CORPUS_config.xml`.

[29] Si l'on utilise JAXE avec le fichier de configuration `CORPUS_config.xml`, un menu déroulant propose l'ensemble des langues de Guyane et du Mexique utiles pour le programme Contacts. Si l'utilisateur veut saisir une autre langue que celles proposées, il lui suffit de ne pas utiliser le menu déroulant, et de saisir directement un code de langue conforme à la norme expliquée plus haut — typiquement, dans le cas le plus simple, un code à trois lettres de l'ISO 639 (`http://www.sil.org/iso639-3/codes.asp?order=639_3&letter=%25`), comme `'deu'` pour l'allemand ou `'rus'` pour le russe.

Langue			
Généricité	Nom	Code	Mnémonique
	Langues amérindiennes		
Langue	Purépecha central (est)	tsz	(TaraSco - Z (?))
Langue	Purépecha des montagnes (ouest)	pua	(PUrépechA)
	Langues européennes		
Langue	Anglais	eng	(ENGlish)
Variante	*Anglais des États-Unis d'Amérique*	eng-US	
Langue	Espagnol	spa	(eSPAñol)
Variante	*Espagnol du Mexique*	spa-MX	

TAB. 6 – Liste des codes normalisés pour identifier les langues du Mexique utilisées dans le programme Contacts

Dans le cas, en revanche, où l'on souhaite rattacher un énoncé ou segment d'énoncé à plusieurs langues, on doit utiliser les deux :

- la liste de langues, contenue dans un élément *langues* immédiatement subordonné à l'élément *texte énoncé* ou *segment* concerné ;
- **et** l'attribut *lang*, qui reste obligatoire, en tant qu'attribut, rattaché à l'élément *texte énoncé* ou *segment*.

La liste de langues ne contient qu'une liste ordonnée d'étiquettes de codes de langues, conformes à la norme décrite ci-dessus (§ 5.1). Il n'y a pas d'attribut supplémentaire pour indiquer, par exemple, une probabilité de rattachement (qui semble trop difficile à quantifier). En revanche, on doit considérer que l'ordre dans lequel les langues sont mentionnées peut être signifiant : il reflète un ordre de dominance, ou de vraisemblance, du rattachement. Si l'on a utilisé cette possibilité de rattachement multiple parce qu'on *hésite* entre deux langues, alors la première langue est plus « probable » que la deuxième ; et si on l'a utilisé parce qu'on pense que le segment relève *à la fois* de deux langues, alors la première langue citée est plus « dominante », dans le mélange, que la seconde.

Dans des cas limites où l'on a aucun critère d'ordonnancement, l'ordre peut être déterminé arbitrairement.

Voici une règle pour déterminer la valeur qu'il faut donner à l'attribut *lang* d'un segment dans le cas où l'on utilise également la liste de langues pour indiquer un rattachement multiple :

- dans le cas où l'on souhaite indiquer un rattachement « plus probable », ou « principal », à une langue donnée, alors celle-ci doit obligatoirement figurer en première position dans la liste de langues. Dans ce cas, l'attribut *lang* doit avoir la même valeur que la première langue de la liste ;
- dans le cas où l'on n'a aucun critère permettant de préférer attribuer à l'une des langues en balance un rôle prédominant, alors on peut donner à l'attribut *lang* la valeur `'mul'`, qui dénote un contenu multilingue (cf. plus haut, § 5.1).

Dans le cas d'un énoncé, la valeur `'mul'` peut être attribuée à l'attribut *lang* dans deux cas de figure :

- soit lorsque l'énoncé, bref, est lui-même un unique segment à rattachement incertain, exactement comme dans le cas exposé plus haut pour l'utilisation de `'mul'` au niveau du segment ;

– soit — dans le cas où l'énoncé est décomposable en segments — lorsqu'il est trop fragmenté en segments de langues différentes, ou à rattachement multiple, pour qu'on puisse sans hésitation lui attribuer une langue principale.

Exemple d'utilisation :

```
<enonce>
<identification_du_locuteur>Jo</identification_du_locuteur>
<texte_enonce lang="mul">
<segment lang="gcr">Vini </segment>
<segment lang="gcr">
<langues><langue lang="gcr"/><langue lang="fra"/></langues>
non </segment>
<segment lang="fra">bande de putes</segment>
</texte_enonce>
<traduction_libre>Venez ici, bande de putes !</traduction_libre>
</enonce>
```

Cet énoncé est composé de trois segments :

– « Vini » : impératif (*venez*), en créole guyanais (`gcr`) ;
– « non » : particule énonciative (« ponctuant ») renforçant l'impératif, d'usage courant en Guyane, et dont on ne sait pas très bien ici si elle doit être considérée comme un mot français, un mot créole, un mot de français régional de Guyane ... ;
– « bande de putes » : groupe nominal en fonction d'apostrophe, en français (`fra`).

Admettons pour l'exemple que l'annotateur a choisi de catégoriser le segment du milieu (« non ») comme étant *plutôt* du créole (*lang*=`"gcr"`), tout en notant l'incertitude par l'utilisation de la liste *langues* (`gcr`, `fra`).

L'annotateur a choisi, dans l'impossibilité d'établir une hiérarchisation des langues au niveau de cet énoncé, de noter celui-ci comme étant à rattachement linguistique multiple (*lang*=`"mul"`).

5.3 Variétés non-natives

Enfin, pour compléter les possibilités d'identification de la langue dans un contexte de contacts de langues, il a été ajouté un attribut supplémentaire, *nonnative*, servant à caractériser les énoncés produits dans une langue par les non-natifs de cette langue[30]. L'attribut *nonnative* prend trois valeurs possibles : *vrai* (il s'agit d'un locuteur non-natif), *faux* (il s'agit d'un locuteur natif), ou *incertain* (on n'est pas en mesure de préciser s'il s'agit d'un locuteur natif ou non). L'indication de la valeur de cet attribut est facultative : sa valeur par défaut, c'est-à-dire la valeur non-marquée, dans le cas où aucune précision n'est donnée, est *faux* : à défaut d'indication contraire explicite, on considère donc que l'énoncé étiqueté comme *Palikur*, par exemple, est produit par un locuteur ayant une compétence native ou quasi-native du Palikur.

[30]La possibilité de coder cette information n'est pas prévue dans les recommandations de la TEI, ni dans la norme ISO 639 d'identification des langues. Le document RFC 4646 prévoit une possibilité générique d'ajouter à l'attribut d'identification de langue des étiquettes non-normalisées, à usage privatif, définies par convention entre utilisateurs ; mais l'usage d'un attribut à part pour dénoter une variété non-native nous a semblé préférable dans notre système, dans le but d'améliorer la lisibilité de cette information tant au niveau de la saisie qu'au niveau de l'exploration.