Élaborer des corpus XML en langues partenaires : quelles technologies appropriées ?

Une expérience en Mauritanie et au Sénégal

Thierno Cisse
Département de linguistique
Université Cheikh Anta Diop de Dakar

Paul Muraille – Marc Van Campenhoudt
Centre de recherche Termisti
Institut supérieur de traducteurs et interprètes
Haute École de Bruxelles

Action de recherche du réseau LTT

www.termisti.refer.org/ltt/ltt.htm

Le paysage d'une recherche appliquée

- L'absence de grands corpus textuels disponibles *in situ*
- Le poids des travaux des « partenaires » du Nord
- L'exigence du développement durable
- La réalité du terrain
- Les obstacles à la graphisation
- L'anglais, *lingua franca*
- Quelle chaîne de traitement du corpus ?

Le déficit en matériaux linguistiques

Pas d'appropriation :

- ❖ dynamisme et moyens aux États-Unis et en Europe
- ❖ terrain occupé par les chercheurs anglophones

Absence de corpus informatisés disponibles sur place :

- ❖ manque de ressources (humaines, matérielles, financières)
- ❖ absence de maintenance informatique
- ❖ absence de projets de grande ampleur

En amont : nécessité de la collecte de données textuelles, audio et vidéo

En aval : quelle exploitation et par qui ?

Développement durable et langues partenaires

Une histoire récente…

1987 : Rapport Brundtland (Nations unies)

1992 : Conférence de Rio (« sommet de la terre »)

➔ 27 principes

➔ Agenda 21

…Qui néglige trop souvent les langues partenaires

Absence de référence dans les textes

Exception : Ouagadougou 2004

« les participants au colloque réaffirment le caractère inaliénable de la diversité culturelle et linguistique comme fondement du développement durable »

Le mouvement des technologies appropriées

Constat

Échec des grands programmes de coopération (époque post-coloniale)

Principe

« Toute société dispose de technologies qui assurent son développement ou du moins sa survie. Ces technologies sont le résultat de la capacité d'invention et d'adaptation de cette société. Leur degré de sophistication et de complexité varie considérablement d'une société à une autre pour des raisons multiples. »

(Crombrugghe 1984 : 65)

Un ensemble non exhaustif de critères :

- économie de devises
- économie d'investissement
- intensité en main-d'œuvre
- économie d'énergie
- usage d'énergies renouvelables
- préservation de l'écologie
- autonomie technique et financière (maintenance)
- acceptation par les populations
- reproductibilité locale
- potentiel de diffusion
- utilisation de matériaux locaux
- utilisation de l'expérience et des savoir-faire locaux

(Crombrugghe 1984 : 65, Darrow & Saxenian 1993)

Un souci de la communication

- Usage privilégié du français fondamental (*cf.* manuel de l'Inades)
- Diffusion de fiches pratiques
- Centres de documentation
- Mouvement d'ONG :
  - Nord : Gret, Cota, Skat…
  - Sud : Enda-Graf, Inades…

Principales applications

- Besoins fondamentaux : eau, terre, énergie, santé
- Peu de travaux portant sur la graphisation, la bureautique, la microédition…

Technologies appropriées : quels critères pour les outils de l'ingénierie linguistique ?

À la recherche du logiciel à cinq pattes :

- Gratuité ou faible coût
- Interface disponible en langue commune (français)
- Documentation disponible en langue commune (français)
- Installable depuis un support ancien
- Connexion Internet superflue
- Multiplateformes, même sur des systèmes vieillis
- Peu gourmand en mémoire vive
- Orienté utilisateur final (linguiste non informaticien)
- Compatible UNICODE
- Sortie en formats non propriétaires (p.ex. XML)
- …

Autonomie du chercheur

- Absence de recours à un informaticien, sans devoir s'improviser informaticien
- Absence de manipulations complexes sur les données
- Ne pas devoir s'initier à une succession de logiciels distincts

La réalité du terrain

Une linguistique de corpus est-elle possible au Sud ?

- ❖ Longue tradition de description des traditions orales et de constitutions de corpus oraux
- ❖ Problèmes structurels liés aux différentes fractures
- ❖ Conditions d'exercices de la profession de chercheur
- ❖ Problèmes logistiques
  - coupures de courant
  - parc informatique réduit, sinon vieillissant
  - antivirus déficients
  - connexions rares et lentes
  - pénurie de personnel informatique

La graphisation : un enjeu fondamental

Un tableau difficile…

- Pas de claviers spécifiques
- UNICODE peu répandu
- Faible représentation dans les grands consortiums (ISO, UNICODE, W3C…)
- Pas ou peu de tables d'écriture spécifiques
- Rendu des polices UNICODE pas toujours satisfaisant (diacritiques)

…avec des conséquences évidentes

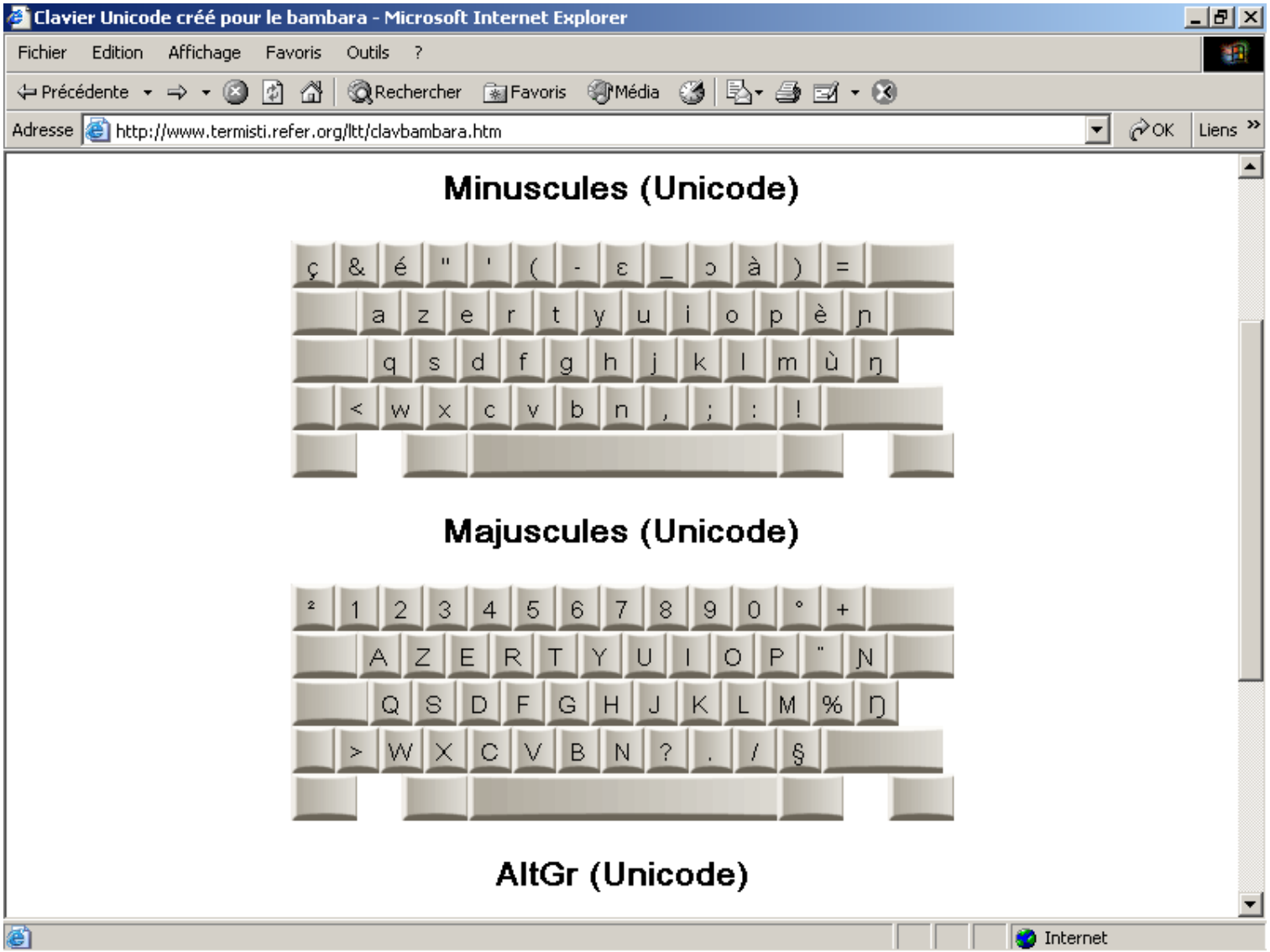
- Difficulté d'utiliser la langue partenaire sous un format numérique
- Normes d'écriture peu utilisées
- Langue partenaire cantonnée à l'oral
- Difficulté à rassembler des corpus écrits d'envergure

+ Proposition d'une simplification des écritures ?

mais technologie appropriée = respectueuse de l'expression locale

Une proposition concrète : le clavier virtuel

- Logiciels :
  - *Keyman Developper*
  - *Microsoft Keyboard Layout Creator (MKLC)*
- Permet de réaffecter les touches d'un clavier
- « Trans-applications » (sauf tel ou tel logiciel…)
- Certitude d'utiliser le bon caractère UNICODE (universalité)
- La création suppose une connaissance d'UNICODE
- Interface non localisée en français
- Gratuit
- Aisé à installer
- Aisé à diffuser


Clavier Unicode créé pour le bambara - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Favoris Média
Adresse http://www.termisti.refer.org/ltt/clavbambara.htm
OK Liens
Minuscules (Unicode)
ç & é " ' (- ɛ _ ɔ à) =
a z e r t y u i o p è ɲ
q s d f g h j k l m ù ŋ
< w x c v b n , ; : !
Majuscules (Unicode)
² 1 2 3 4 5 6 7 8 9 0 ° +
A Z E R T Y U I O P " Ɲ
Q S D F G H J K L M % Ŋ
> W X C V B N ? . / §
AltGr (Unicode)
Internet

Il serait aisé et peu coûteux de lancer un programme international de diffusion systématique de tels claviers dans les endroits *ad hoc* :

- écoles, lycées, universités
- ministères, mairies, bureaux de police…
- ONG
- fournisseurs d'accès Internet, banques, assurances…

Permettre l'écriture est une étape fondamentale pour la numérisation des langues partenaires et donc la présence sur Internet.

On navigue sur Internet en swahili !

Open Office est disponible en swahili !

L'anglais, *lingua franca* incontournable ?

Comment s'approprier une technologie si l'on ne lit pas l'anglais ?

Comment faire progresser une norme si l'on ne s'exprime pas en anglais ?

Normes :

- XML et HTML : traductions fragmentaires bénévoles
- UNICODE : traduction partielle
- TEI : en anglais
- TEI Lite : traduction
- (X)CES : en anglais
- TMF: en anglais

Logiciels :

- *MKLC* : en anglais
- *Keyman* : en anglais
- *Toolbox* : en anglais
- *XML Spy* : en anglais
- *Oxygen* : en français

Quelles solutions pour permettre l'appropriation ?

- Davantage et mieux enseigner l'anglais ?
- Renforcer la stratégie du pré carré ?
- Organiser des sommets, « coquetèles » et vitrines technologiques ?
- Au minimum, permettre un dialogue international !

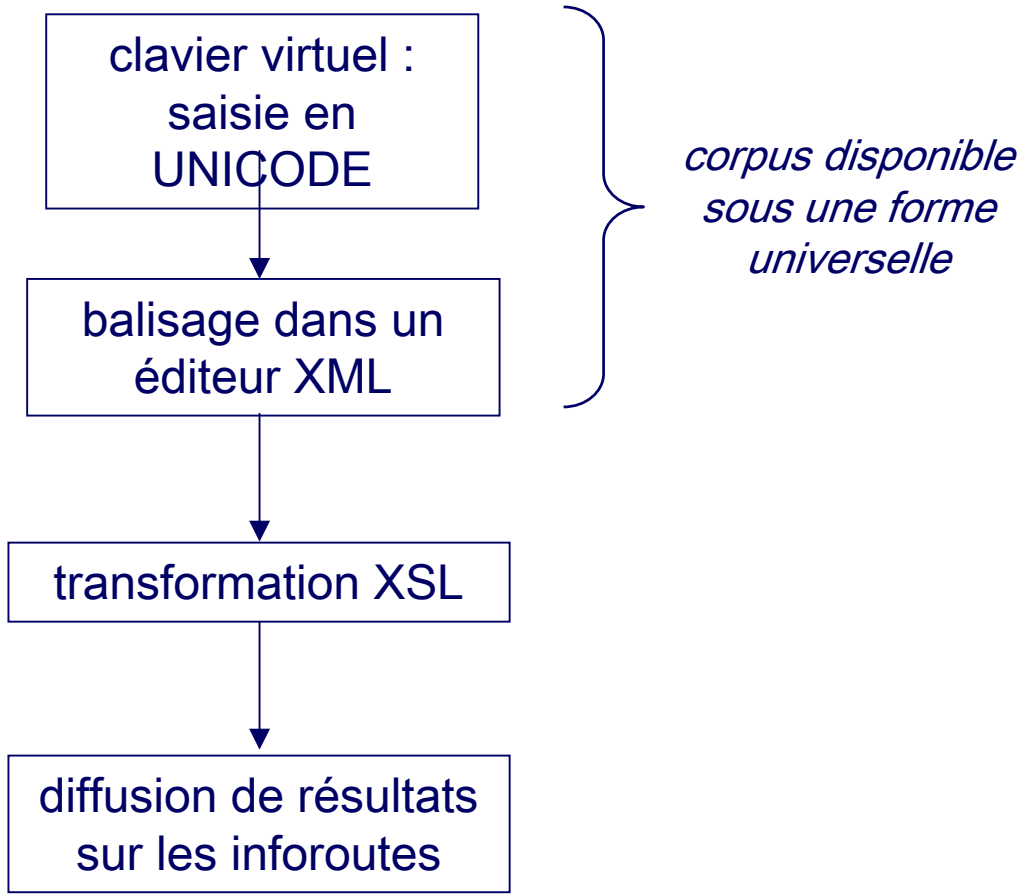
Plaidoyer pour une cellule de traduction « inforoutes »

Faire de la francophonie un moteur du progrès en :

- ❖ assurant une traduction professionnelle et une diffusion et une actualisation rapides des normes en français
- ❖ permettant une traduction en anglais des réactions et demandes d'aménagement des chercheurs francophones
- ❖ localisant les logiciels incontournables (*Toolbox !*)

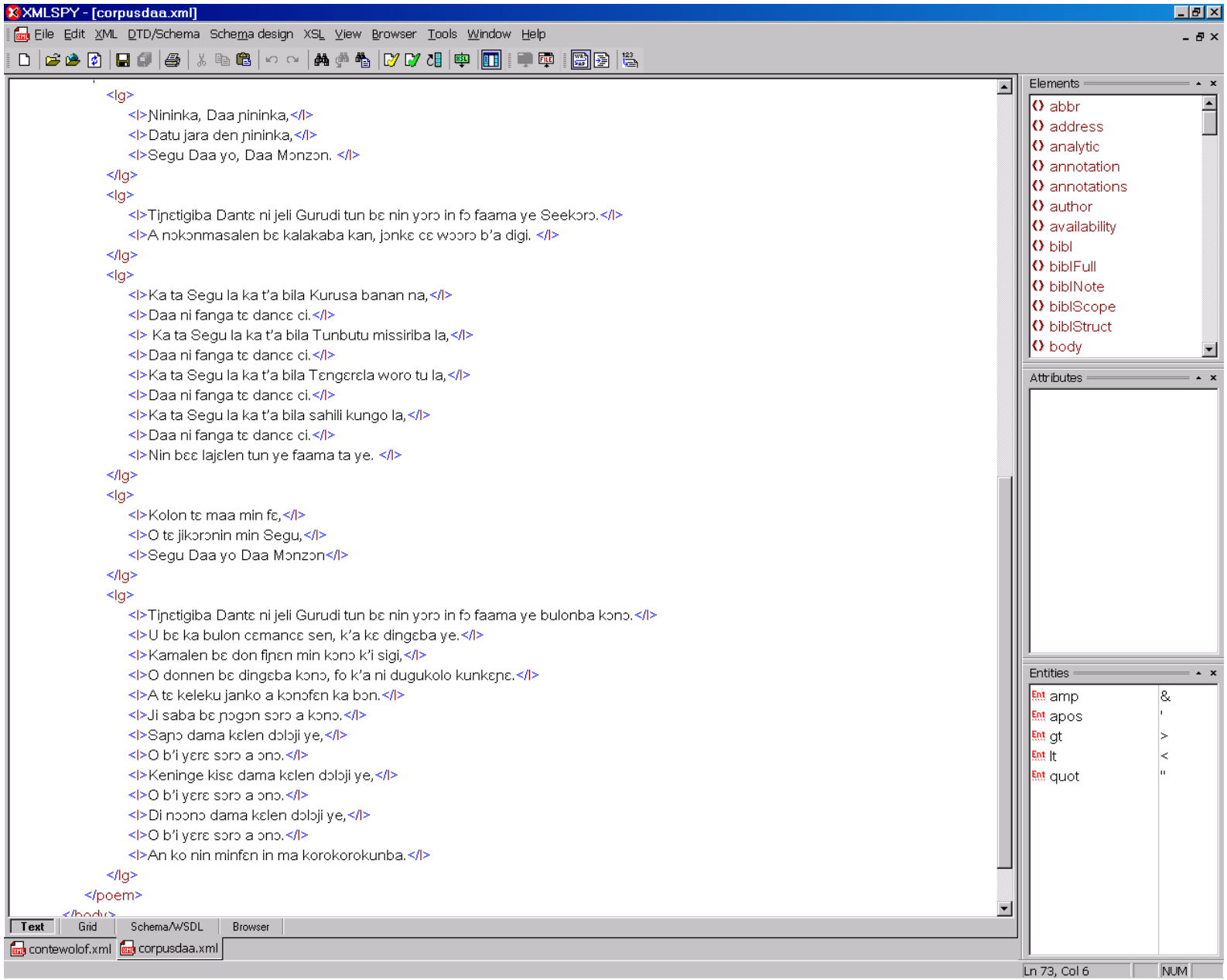
Un coût dérisoire pour une solution efficace !

Quelle chaîne de traitement du corpus textuel ?*

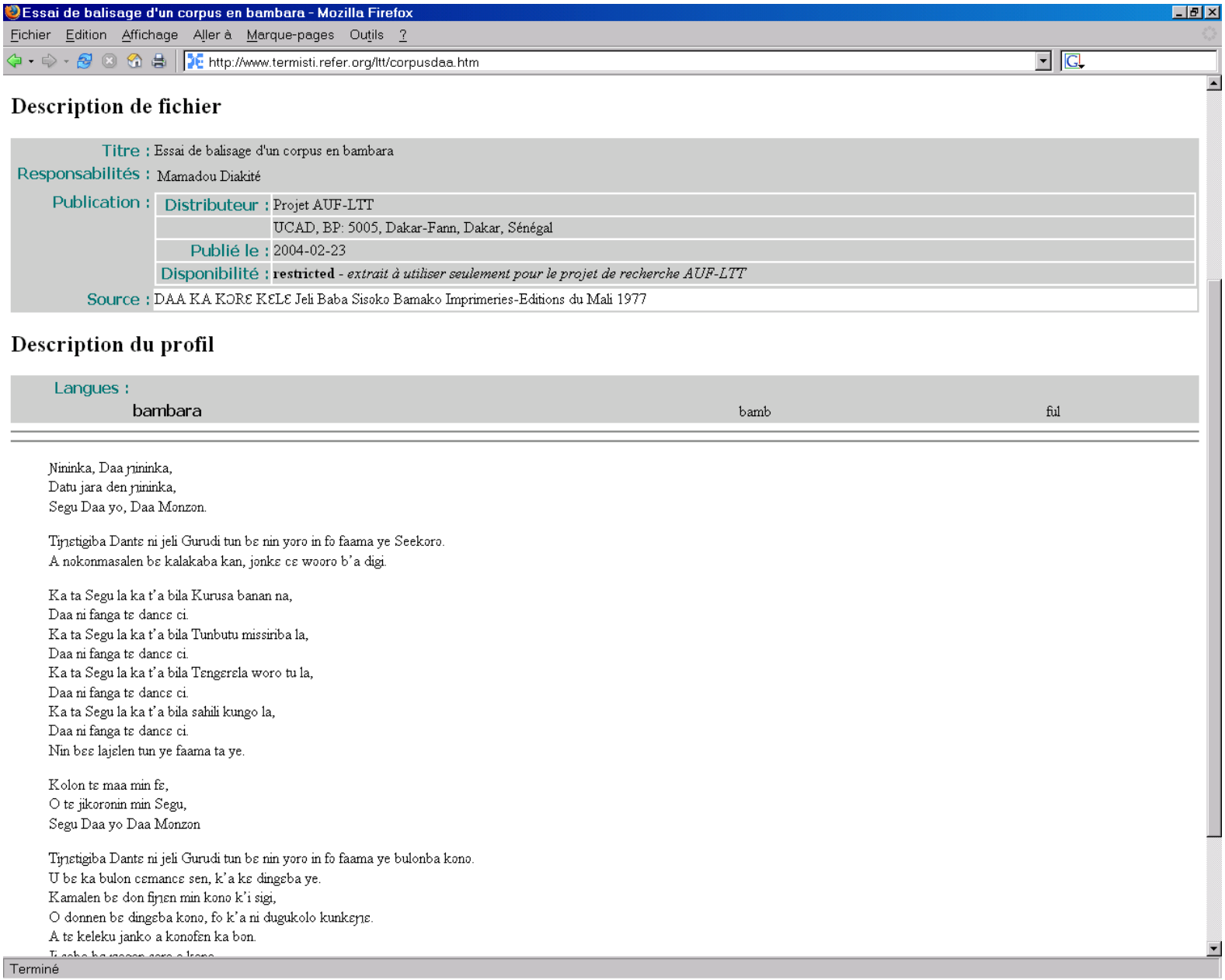


* Le traitement des bases de données lexicales et terminologiques n'est pas envisagé ici.

Exemple de fichier XML



Exemple de conversion vers HTML



Normes expérimentées dans le cadre du projet

- *Text Encoding Initiative* (TEI et TEI Lite)
- *Corpus Encoding Standard for XML* (XCES)

Validation du processus

- Barrière de la langue anglaise
- Catégories de données prises en compte à adapter aux genres relevant de l'oralité
- Simplifier le travail du linguiste :
  - Activité de balisage manuel lourde, sinon fastidieuse
  - Transformations XSL probantes mais résultats à améliorer
  - XSL difficile à maîtriser au-delà du bricolage

➔ Nécessité de simplifier le processus
pour aboutir à une technologie plus appropriée

Pour une chaîne simplifiée

- Passerelle TEI - *Open Office* :
  - associer style et catégorie de données
  - de *Writer* vers la TEI et de la TEI vers *Writer*

➔ démonstration probante, mais peu de styles disponibles

- Avantages d'une telle démarche :
  - propreté du fichier de traitement de texte
  - feuilles de style XSL déjà rédigées
  - moins de logiciels à maîtriser
  - seule une initiation à la gestion de styles est nécessaire

- Exemple d'application type : projet **Cyberdocs** (Cyberthèses)

  À sans doute adapter en fonction des critères propres aux technologies appropriées

Concrètement

Permettre une appropriation et une interaction effectives

Information et sensibilisation

- Répertoire commenté de ressources francophones disponibles
- Cellule de traduction et de localisation

Outils directement utilisables

- Mise en ligne de fiches et modèles pratiques
- Diffusion des technologies UNICODE

Développements locaux

- Adaptation des normes d'échange (linguiste)
- Adaptation des outils (informaticien)

Revoir les diapos : www.termisti.refer.org/ltt/ltt.htm