Filières TAL et Ingénierie Linguistique de Paris III Sorbonne nouvelle, Paris X Nanterre, INALCO (Institut National des langues et civilisations orientales)

http://www.cavi.univ-paris3.fr/ilpga/plurital/

JOURNEE DU 19 NOVEMBRE 2004 14 HEURES

CONTENU :

ORGANISATION PAR GROUPES : CHAQUE GROUPE ASSISTE A UNE SEANCE DE 50 MINUTES ET REJOINT ENSUITE LA SALLE FIXEE CI-DESSOUS

Les groupes ont été formés sur des critères purement linguistiques (présence de locuteurs arabophones, chinois, japonais dans chaque groupe).
Soyez à l'heure car nous devons libérer la salle 124 à 17 heures.

le groupe 1 commence en 124, puis va en 204 et termine en 304
le groupe 2 commence en 204, puis va en 304 et termine en 124
le groupe 3 commence en 304, puis va en 124 et termine en 204

GROUPE 1	GROUPE 2	GROUPE 3
Rémi Quesnel	Lim Lay-Chan	Blandine Deleuze
Dafir Bouchra	Rachid Belmouhoud	Rym Ben Haddada
Anne-Myrtille Renoux	Masaru Tomimitsu	Cyril Rassou
Pierre-Julien Pera	Marie-Véronique Leroi	Sacha Nguyen
Roland Côté	Gwendoline Fox	Anne Garcia-Fernandez
Margaretha Lenting	Frédérique Bénard	Olga Kussy
Jesir Vargas	Anghel Nicoleta-Ramona	Jean-Louis Garçon
Sarra El Ayari	Audrey Vittecoq	Sandra Petel
Oscar Garcia	Sonia Krivine	Aleksandra Sondermeijer
Sophie Chevrier	Zita Bognar	Delphine Lagarde
Aurélie Cousseau	Priscille Sabin	

CREATION D'UNE PAGE MULTILINGUE

Construire une page web contenant des définitions en arabe, anglais, chinois, japonais du mot laïcité, définitions qui seront accompagnées de leurs traductions françaises respectives.

Compétences mises en œuvre

-application du cours sur normes et standards
-manipulation des outils simples d'acquisition et normalisation de corpus
-prise de conscience des spécifications relatives à la construction de pages multilingues
-réalisation de la même page (un tableau) sous Windows et Linux pour prendre conscience de la diversité des OS, des éditeurs, des navigateurs (en particulier le « copier » sous WINDOWS semble s'adapter au format de la page où l'on utilise le « coller », d'où un illusoire sentiment de simplicité, alors que les choses sont moins simples sous LINUX, mais aussi plus claires)

Ressources : http://crim.inalco.fr/corpus/plurital/lay_[ar|ja|zh|en]_fr*

WINDOWS (SALLE 304)	**LINUX (SALLE 204)**
aller à l'URL crim.inalco.fr/corpus/plurital/*.htm	idem
comment les pages s'affichent-elles ? constater le mélange de langues dans la page lay_zh_fr, pourquoi l'affichage est-il correct ? pourquoi l'arabe s'affiche-t-il correctement ? aller voir source du document rappeler brièvement structure pages html	idem affichage avec les navigateurs présents : -konqueror -netscape
comment ces pages sont-elles codées (charset) ? le charset de la source est-il bien celui choisi par le navigateur ? afficher lay_en_fr avec codage utf-8/avec codage windows (cp-1252)	idem
expliquer l'importance des outils pour récupérer un gros corpus à l'aide de HTTRACK, aspirer ces pages en suivant les étapes suivantes -projet : nom de nouveau répertoire accueillant pages -action : copie automatique de sites web -définir les options : pour experts : mode de parcours : rester sur le même répertoire -OK, suivant, Terminer	récupération du corpus : créer répertoire de travail mkdir mon_rep aller dans ce répertoire cd mon_rep enregistrer corpus avec ligne de commande wget –r –l 2 –A lay*.htm crim.inalco.fr/corpus (voir pages d'aide avec man wget pour plus de détails, -r indique récursivité, -l profondeur d'enregistrement à partir de la racine) vérifier que les bons fichiers ont été enregistrés (ceux contenant l'expression régulière lay)
aller dans le répertoire choisi pour accueillir pages, enlever les fichiers inutiles	vérifier que les bons fichiers ont été enregistrés (ceux content l'expression régulière lay), éventuellement enlever les scories
ouvrir pages arabe, chinoise, japonaise, lay_en_fr -avec bloc-notes -avec word -avec word-pad -avec navigateurs (IE/Netscape)	afficher avec less, cat, cat -A ouvrir pages arabe, chinoise, japonaise, lay_en_fr -avec emacs -avec vim -avec gedit -avec navigateurs (Konqueror/Netscape)
aller à l'URL crim.inalco.fr/corpus/plurital_illegal/*.htm	aller à l'URL crim.inalco.fr/corpus/plurital_illegal/*.htm
afficher pages, aller voir source	afficher pages, aller voir source

pourquoi un affichage avec entités ?	pourquoi un affichage avec entités ?
parler brièvement des outils disponibles sur les machines pour normaliser des corpus : - detagger (équivalent de html2txt.pl sous linux) - unifier (en gros équivalent de recode sous linux) - ici, étant donné la simplicité du code html, on pourrait envisager une commande simple utilisant sed	outils pour corpus : -en particulier html2txt.pl (site CPAN http://www.cpan.org/) -recode (pour changer codage fichier si non-conforme, pour créer page multilingue en utf-8 par exemple)
passer à la réalisation de la page : code html minimum : html, title, head, body	idem
pour le tableau, balises minimum : table, tr, td (mettre cellpadding cellspacing en fin de parcours)	idem
2 colonnes : une avec les définitions en chinois, arabe, anglais, japonais, l'autre avec les traductions en français de ces définitions	idem
parler des spécifications du w3c pour les pages multilingues comme un attribut lang à chaque changement de langue, l'attribut dir=rtl pour l'arabe ou autres langues de ce type	idem
écrire dans bloc-notes la structure du tableau nu	créer le tableau avec gedit, attention il faudra sans doute utiliser recode pour les fichiers non UTF-8 avec une syntaxe du type recode windows-1252..UTF-8<fichier>nouv_fichier
remplir le tableau par copier-coller	idem
signaler qu'un gros tableau devrait être rempli automatiquement par ouverture de fichiers et remplissage de chaque cellule avec chaque ligne des fichiers (par exemple)	idem
afficher tableau résultant, rectifier format avec les attributs cellpadding, cellspacing si nécessaire, donner au fichier source nom dérivé du nom de l'élève ou du groupe d'élèves l'ayant réalisé	idem
uploader le tableau à l'adresse avec le nom retenu http://www-new.biomath.jussieu.fr/cgi-bin/depot-plurital.cgi	idem
afficher sur site distant	idem

LISTE DES ELEVES PLURITAL

Nom	Mail	Lieu	Séance 1	Séance 2
Cyril Rassou	cyrilrassou@voila.fr			
Pera Pierre-Julien	pejipera@hotmail.com			
Marie-Véronique LEROI	marivliax@yahoo.fr	INALCO		
Sacha NGUYEN	sacha_nguyen@yahoo.fr			
Côté Roland	cote.roland@uqam.ca			
Gwendoline FOX	gwendoline_fox@yahoo.fr	P3		
Garcia-Fernandez Anne	bobette_anne@yahoo.fr			
Louise DELEGER	louisedeleger@hotmail.com			
Lay-chan Lim	lim_lay_chan@yahoo.fr			
Margaretha LENTING	lentinglamarck@aol.com			
Frédérique Bénard	fred-benard@freesurf.fr	P3		
Masaru Tomimitsu	mtomimitsu@free.fr			
Rym Ben Haddada	rymbh@noos.fr			
Olga KUSYY	olga_kusyy@yahoo.fr			
JESIR VARGAS	jvargas@acm.org	P3		
Nicoleta-Ramona ANGHEL	rami78_ro@yahoo.ro			
Blandine Deleuze	blandine_deleuze@hotmail.com			
Dafir Bouchra	dafir_bouchra@yahoo.com			
Jean Louis Garçon	ziel@club-internet.fr			
Belmouhoub Rachid	belmouhoub.rachid@libertysurf.fr			
Sarra El Ayari	aie.carumba@wanadoo.fr	P3		
Audrey Vittecoq	audreyvittec@aol.com			
Rémi Quesnel	remiquesnel@hotmail.com			
Sandra Petel	petelsandra@yahoo.fr			
OSCAR GARCIA	osqvar@hotmail.com	P3		
Sonia Krivine	sonia.krivine@free.fr			
Aleksandra Sondermeijer	a_nikoloska@hotmail.com			
Sophie Chevrier	chevriersophie@yahoo.fr			
Zita BOGNAR	zita.bognar@wanadoo.fr			
Delphine Lagarde	lagadella@yahoo.fr	P3		
Anne-Myrtille Renoux	am_renoux@yahoo.fr			
Aurélie Cousseau	mini_aurelie78@hotmail.com			
Priscille Sabin	cypriscille@hotmail.com			

fichiers html non-conformes aux spécifications du w3c, poids des fichiers faits avec word est énorme, déconseillé

bloc-notes : OK
wordpad : OK
word : affiche comme page web, rajout de balises
arabe : attention, affichage à l'envers, rétablissement dans le navigateur avec dir=rtl

HTTRACK

montrer le site avant (pages cachées, pas de liens à ces pages dans site), plurital et plurital_illegal

faire nouveau projet : laïcité (crée dossier de ce nom)
signaler aux étudiants où vont se retrouver les fichiers enregistrés (chemin de base)

action : copie automatique de sites web
adresse Web : crim.inalco.fr/corpus/plurital/
paramètres de la copie du site : définir les options : rien (trop de fic
+crim.inalco.fr/corpus/plurital/lay*.htm* : même chose)
pour experts : rester sur le même répertoire OUI
règles de filtrage +crim.inalco.fr/corpus/plurital/lay*.htm* (inutile, rester sur le même repertoire suffit NON)
il reste des scories (car pas de liens internes ?)

copier-coller doit marcher ? oui mais l'arabe s'affiche pas dans le bon sens il faut rajouter une balise pour le sens (?) laquelle

gestion des cellules : comment faire ligne blanche/