

## PROJET DIT « MULTILINGUE »

Le projet proposé a pour but, à partir de données textuelles disponibles sur la toile, de produire des ressources linguistiques structurées. Ces dernières peuvent avoir une finalité lexicologique (voir [http://crim.fr/lexique\\_ri.html](http://crim.fr/lexique_ri.html) pour un exemple de lexique français-arabe des Relations Internationales), en particulier pour des couples de langues français+une langue enseignée à l'INALCO.

Mais on peut envisager aussi, à partir de données textuelles brutes, de poser un problème linguistique et d'y apporter une réponse à l'aide de méthodes et outils issus de la linguistique-informatique : on pourrait par exemple étudier, à partir d'un corpus parallèle ou comparable français-anglais, des phénomènes comme les équivalences de traduction : GN en français—GV en anglais (voir exemples en annexe).

De même, on pourrait s'interroger, dans le cadre de la traduction, sur des phénomènes comme le choix GN discret/non-discret en anglais pour traduire des GN français pluriels : en particulier, comment le genre textuel influence-t-il ce choix (voir exemples en annexe) ?

Dans tous les cas, ce projet fera l'objet d'une présentation en ligne et offrira une valeur ajoutée, non seulement dans la compréhension des phénomènes linguistiques étudiés, mais aussi dans le choix des outils utilisés. Pour ces derniers, en particulier dans le cas de langues dites « peu dotées » en outils d'ingénierie linguistique, la présentation en ligne permettra de fournir une documentation et des liens vers les outils, ainsi qu'un mode d'emploi clair et une évaluation succincte.

### ETAPES (15 SEMAINES)

**(les chiffres entre parenthèses correspondent à une progression par semaine)**

Tout au long du projet, on utilisera un outil permettant au groupe de communiquer (montrer son code pour demander où est l'erreur, donner adresse de sites utiles, présenter une expression régulière vraiment incompréhensible, connaître le code de l'espace.....).

-constitution raisonnée de corpus en fonction de la finalité retenue (1-3)

lexiques multilingues en ligne :	corpus multilingues, parallèles ou comparables, domaine
glossaire spécialisé :	documentation technique, extraction de définition
terminologie :	domaine restreint à choisir, technique d'extraction
traductologie :	problème posé
linguistique :	problème posé

-méthodologie de la constitution selon finalité (1-3)

quels textes ? quels genres ? quelles sources (journaux, magazines, romans, blogs...) ? source unique (facilite la phase de pré-traitement du corpus) ? comment assurer la couverture maximale du domaine considéré ? quel est le biais apporté par la collecte de données numériques ? quelles caractéristiques lexicales et grammaticales peuvent aider à classer par « genre » ?

-étalon de mesure (1-3)

comment mesurer la pertinence des données réunies ? quelles données pourraient servir de références ? comment évaluer la qualité des données (en particulier dans des domaines comme la traduction) ?

-qualité du corpus (1-3)

Où trouve-t-on des corpus parallèles de qualité ? Sont-ils disponibles ? Quel format de fichier est le plus susceptible de refléter une certaine qualité (.html, .pdf...) ? Les outils destinés à mettre le corpus au format texte existent-ils ? Les fichiers sont-ils utilisables (cf. fichiers verrouillés)

-pré-traitement du corpus en vue de traitements linguistiques (3-4)

encodage standard (selon outils à utiliser, syntaxe par exemple prend du iso-latin-1 en entrée) de documents venant de sources multiples

-conservation des traces de chaque étape de traitement (3-6)

capacité de renvoyer au corpus de départ (.html), au corpus au format texte, au corpus étiqueté, au corpus aligné. Choix d'un format pour la conservation(XML...). Voir par exemple [http://www.crim.fr/monde\\_diplo](http://www.crim.fr/monde_diplo) qui présente les données sous différents formats, selon l'étape de traitement linguistique.

-débalisage ou traduction en format texte (4-5)

quels outils pour quels formats ? quelles contraintes, en particulier en termes d'encodage (le débalisage ne doit pas corrompre les fichiers) ? quelles sont les options des outils pour traiter les pages .html contenant des graphiques, tableaux, colonnes ?

-correspondance de paires de fichiers (4-6)

faut-il un alignement brutal de chaque paire de fichiers débalisés ? Ou doit-on garder trace de la structure HTML pour réaliser un meilleur alignement (titres, liens hypertexte...)

-étiquetage morpho-syntaxique (5-7)

2 langues étiquettent-elles de façon similaire des phénomènes de surface qui semblent identiques ? Comment limiter les distorsions créées par les outils de façon à ce que les conclusions linguistiques finales restent valides ? Comment se présentent les entrées des étiqueteurs (fichier texte pour Cordial, une phrase par ligne pour le tagger de Brill dans laquelle les marques de ponctuation sont précédées et suivies d'un blanc, ...) ? Quels sont les traitements informatiques à mettre en œuvre pour avoir des données au bon format ? Quels outils existent pour mettre ces données au format : outils du shell (sed, tr...), langage de programmation (Perl) ? Comment se présentent les sorties des étiqueteurs ?

-extraction terminologique (7-10)

Quelle est la structure d'un terme d'une langue à l'autre ? Peut-on envisager une correspondance de structures de termes (p. ex. 'pouvoir d'achat, liberté d'opinion, table des négociations, conseil de sécurité' sont tous des syntagmes du type 'Nom de Nom', mais cette structure constante se retrouve-t-elle en langue-cible ?) Quels sont les extracteurs

terminologiques existants ? Ecrire des programmes permettant d'extraire des patrons morpho-syntaxiques à partir d'un texte étiqueté.

Voir [http://www.crim.fr/patrons\\_verbaux\\_pour\\_extraction.html](http://www.crim.fr/patrons_verbaux_pour_extraction.html) pour un exemple de patrons verbaux.

Evaluer et trier les sorties.

Voir [http://www.crim.fr/resultat\\_extraction.html](http://www.crim.fr/resultat_extraction.html) pour un exemple de sortie non triée.

-alignement de phrases (9-10)

Quelles sont les méthodes classiques d'alignement (statistiques, linguistiques, cognats...) ? Ecrire un programme permettant de séparer un texte en phrases terminées par un point. Quelles difficultés ?

Voir un exemple d'alignement manuel permettant l'extraction de syntagmes nominaux à [http://www.crim.fr/tableau\\_de\\_correspondance\\_noms.html](http://www.crim.fr/tableau_de_correspondance_noms.html)

Quels sont les résultats produits par un alignement fruste (chaque phrase de la langue-source est alignée par défaut avec chaque phrase de la langue-cible, en commençant en haut du fichier).

-alignement de mots/syntagmes (11-12)

Outils existants ? Choix de la méthode.

-mise en ligne, corrections, organisation du projet en site (12-15)

ANNEXES (syntagmes proposés hors contexte, mais exemples réels disponibles)

<b>NOMS EN FRANÇAIS</b>	<b>VERBES EN ANGLAIS</b>
à défaut, faute de quoi	failing that
à l'approche de	as sthg nears
à l'instigation de	at the urging of
activité législative	law-making
affaires/possessions	belongings
alarmiste	alarm-raising
ancestral	time-honored
applicatif	application-oriented
artisanal	home-made
attribution	granting
au bas mot	that's a conservative estimate
au fil des mois/au fil du temps	as months went by/as time goes by
au grand dam de	causing the anger of
aucune idée	search me
autodidacte	self-made-man
avant la date fixée	before the deadline runs out
beuverie	binge-drinking
bouclage	cordonning/sealing off
calciné	burnt-down
casanier	stay-at-home
citadins	city-dwellers
combats	fighting
combines/micmacs	wheeling and dealing
comité permanent	standing committee
concentration (militaire)	build-up
condoléances	be sorry for sbdy's loss
conduite	driving
construction navale	ship-building
contingentement	quota-setting
contrebande	smuggling
dans la mesure du possible	if I could help it
date de péremption	the sell-by/best-before date
de bon/mauvais augure	that bodes well/ill of
de son propre fait	of one's own making
démantèlement	dismantling
dépenses	spending
déplacé	uncalled-for
diplômé	degree-holder (GB)
discours	what they say
économie du savoir	the knowledge-based economy
emballages	packaging
en baisse	flagging
en convalescence	recovering

en gestation/en devenir/potentiel/virtuel	in the making
en herbe/naissant	budding
en-cas	snacking
éphémère	short-lived
escalade	rock-climbing
euthanasie	mercy-killing
exceptionnel/inédit	unheard-of
externalisation	outsourcing
financement	funding/financing
fixation	setting
flottement (hésitation)	dithering
funambulisme	tight-rope walking/a balancing act
fusillade	shooting
grabataire	bed-ridden
grande époque	those were the days
impossible	no can do
inauguration (pol.)	swearing-in ceremony
indécis	fence-sitter
innovant	ground-breaking
intact	unimpaired
jeux de hasard	gambling
la collecte	gathering
le moins-disant social	social dumping
le nième jour consécutif	the nth day running
législateur/parlementaire	law-maker
les mains vides	empty-handed
levée (d'une mesure)	lifting
maintien de l'ordre	policing/law enforcement
majeur/à grande échelle/véritable	full-blown
majorité	come of age
même constatation pour	the same goes for
minutieux/méticuleux	painstaking
mise au point	get your facts straight
mise en chômage technique	idling
mobilisation	rallying around
multiplication	ever-increasing number of
multiplication	ever-increasing number of
numéro vert	call toll-free
opérations/actes	dealings
optimiste	upbeat/sanguine
par le biais de X	X-brokered
parachutisme	sky-diving
pendaison de crémaillère	house-warming party
perdu	God-forsaken
petite phrase	soundbite
planification	planning
pointage	clocking in/out
président en exercice	acting president

prolifération	mushrooming
qualités	things going for
réchauffement de la planète	global warming
récidive	reoffending
recours	resorting
recyclage de (déchets)	recycling, reprocessing
rédaction	drafting
rééchelonnement	rescheduling
réformiste	reform-minded
relâchement	let-up
répartition de X	the way X breaks down
représentant de l'ordre	law-enforcer
résultats	findings
rodéos (en voiture)	joyriding
sans-opinion	don't know (DK)
sans-papiers	undocumented
selon le cas	as the case may be
si nécessaire	if need be
soins prodigués aux malades	patient care
sous conditions de ressources	means-tested
sous couvert d'anonymat	on condition he not be identified
sous le regard de X	while X looked on
sous les yeux de	as X was looking on
strident	high-pitched
surpopulation carcérale	prison overcrowding
système de suivi	tracking device
terme générique	catch-all phrase
terne	lacklustre
timide (incertain)	faltering
traçabilité	track-and-trace (techniques)
transsexualisme	gender-swapping
vieillesse	ageing
volontaire	can-do

<b>SYNTAGMES FRANÇAIS (DISCRETS)</b>	<b>SYNTAGMES ANGLAIS (NON- DISCRETS)</b>
actions (mesures)	action
affaires	business
affaires (les)	business
allégements fiscaux	tax relief/tax breaks/tax cuts
applaudissements	applause
atermoiements/tergiversations	procrastination
belle-famille	in-laws
bureaucratie	bureaucrats
capacités	ability
capacités de production non-utilisées	spare capacity
capitaux	capital
changements	change
changer de camp	switch allegiances/sides
Clergé	clerics
combats	fighting
combines/micmacs	wheeling and dealing
commentaires	comment
commentaires (conjectures)	speculation
conflits	conflict
conjectures	speculation
conséquences	fallout
contestations	protest
dans ses pensées	deep in thought
de plus en plus de critiques	a growing amount of criticism
déchets	waste
dégâts	damage
délits	crime
dépenses	spending
des expériences	experience
des faits	fact
des morts	loss of life
des traitements médicaux	medical treatment
détails	detail
devoirs	homework
dissensions	dissent
divergences	disagreement
efforts	effort
électorat	voters
emballages	packaging
embouteillages	congestion
émettre des critiques	level criticism
ennuis	trouble
entourage	his closest advisers
état-major/direction	leadership/leaders

faux-semblants	pretence
heures supplémentaires	overtime
incertitudes	uncertainty
inquiétudes	concern
insultes	abuse
investissements	investment
jeux de hasard	gambling
la culture	the arts
le consensus Les Echos	Les Echos' panel of economic forecasters
le jury	they
le moindre effort	the least amount of effort
les activités	activity
les analyses	analysis
les horaires aménagés	flexitime
les secours	help
les urgences	an emergency room
logiciels	software
loisirs	leisure, entertainment
louanges	praise
luxue de détails	wealth of detail
médicaments	medication/medicine
munitions	ammunition
mutations	change
négligences	neglect
peines	punishment
peu d'indications	little sign
polémiques	controversy
potins mondains	celebrity gossip
progrès	progress
provoquer des débats	stir debate
recherches	research
recoupements	overlap
réformes agraires	land reform
remords	remorse
renseignements	intelligence
réticences	reluctance
rires	laughter
selon les plans/comme prévu	according to plan
signes	sign
soins	care
soins gratuits aux malades de longue durée	free long-term care
soins prodigués aux malades	patient care
somme de connaissances	amount of knowledge
Sornettes	nonsense
Spécificités	distinctiveness
Suffrages	vote
Témoignages	testimony
terres agricoles	farmland

Transports	transport
Travaux	work
travaux ménagers	housework
turbulences (remous)	turmoil
Violences	violence