Méthodes pour informatiser les langues et les groupes de langues " peu dotées "

Vincent Berment

UNIVERSITÉ JOSEPH FOURIER, GRENOBLE 1
UFR D'INFORMATIQUE ET MATHÉMATIQUES APPLIQUÉES

THÈSE

présentée et soutenue publiquement le 18 mai 2004 par

Vincent BERMENT

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Spécialité

INFORMATIQUE

MÉTHODES POUR INFORMATISER DES LANGUES ET DES GROUPES DE LANGUES « PEU DOTÉES »

Jury :

M.	Bruno	OUDET	Président
M.	Yves	LEPAGE	Rapporteur
M.	Jean	VÉRONIS	Rapporteur
M.	Christian	BOITET	Directeur
M.	Gilles	DELOUCHE	Examinateur
M.	Mathieu	LAFOURCADE	Examinateur
M.	Claude	DEL VIGNA	Invité

THÈSE PRÉPARÉE AU SEIN DU GETA, LABORATOIRE CLIPS (IMAG, UJF, INPG & CNRS)

Remerciements

Au moment où ce travail s'achève, je tiens à remercier :

Monsieur Bruno Oudet, professeur à l'Université Joseph Fourier, à l'origine du Chapitre Français de l'Internet Society et de la Fête de l'Internet, qui m'a fait l'honneur de présider le jury.

Messieurs Yves Lepage, chargé de recherche aux laboratoires *traduction et langue parlée* de la société ATR à Kyoto, et Jean Véronis, professeur à l'Université de Provence, qui ont accepté avec bienveillance d'être rapporteurs et m'ont prodigué des remarques pertinentes qui m'ont permis d'améliorer ce document.

Messieurs Gilles Delouche, Claude Del Vigna et Mathieu Lafourcade, qui ont accepté de participer à mon jury de thèse.

Monsieur Gilles Delouche, président de l'INALCO, qui fut mon inoubliable professeur de siamois voilà déjà dix ans et qui, malgré sa charge nouvelle de président, m'a aidé avec beaucoup d'attention lorsqu'il s'est agi de préciser la structure des syllabes siamoises.

Monsieur Claude Del Vigna, ingénieur de recherche au CNRS, qui a encouragé mes efforts depuis mon DEA et a contribué très activement à l'étude d'un aspect théorique de la segmentation syllabique présentée dans cette thèse.

Monsieur Mathieu Lafourcade, maître de conférence à l'Université Montpellier 2, qui a accompagné de sa bonne humeur ma première conférence qui se tenait à Penang et qui m'a initié, à cette occasion, à la notion de vecteur sémantique.

Monsieur Christian Boitet, professeur à l'Université Joseph Fourier, pour l'ampleur du sujet qu'il m'a offert d'explorer. Sans son initiative, cette thèse ne serait pas. Je le remercie très chaleureusement aussi pour les nombreux conseils prodigués pendant la thèse et qui m'ont appris tant de choses ainsi que pour la grande disponibilité et l'enthousiasme dont il n'a cessé de faire preuve.

J'exprime aussi toute ma reconnaissance aux personnes qui m'ont aidé, à un moment où à un autre de cette thèse, qui à appréhender une notion, qui une subtilité linguistique ou informatique, et en particulier à Michel Antelme, Patrick Beaudouin, Pascal Berment, Louis-Jean Calvet, Kim Chuah Choy, Éric Duboscq, Michel Fanton, Michel Ferlus, Colette Grinevald, Paul Hector, Michel Ilkiewicz, Lamvieng Inthamone, Mustafa Jabbar, Marilyn Mason, Thakkhinh Jacqmin, Igor Mel'čuk, Jean-Claude Meunier, Tai-Luc Nguyen, Alain Polguère, Sunee Pongpinigpinyo, Pierre Sein-Aye, Christian Thomas, Houmphanh Thongvilu, Roland Touchais, Dominique Vaufreydaz, Romain Wong ainsi qu'à tous les membres du GETA.

Enfin, j'ai une pensée affectueuse pour Christine qui a accepté avec beaucoup de gentillesse et de patience toutes les contraintes qu'imposait ce travail.

Table des matières

Table des figures

Index

Introduction

INTRODUCTION

SITUATION ET MOTIVATIONS

Ce mémoire s'inscrit dans un large mouvement international qui vise à ce que chaque peuple puisse disposer de tous les moyens pour communiquer dans sa langue. Dans les siècles précédents, affirmer ou défendre une langue passait par d'autres moyens : fixer une orthographe, construire des dictionnaires monolingues ou bilingues, recueillir des traditions orales ou encore élaborer des polices d'imprimeur.

Aujourd'hui, le développement des ordinateurs personnels et celui des réseaux font de l'informatique un instrument pour écrire et communiquer au même titre que le papier l'est depuis Cai Lun et l'imprimerie depuis Gütenberg. Traitements de texte et courriers électroniques sont devenus des outils de langue largement répandus. En dépit du caractère manifestement politique de ce mouvement d'affirmation des langues — si l'on s'accorde, avec Hannah Arendt ([Arendt 1995]), pour dire que « la politique repose sur un fait : la pluralité humaine » — l'idée s'impose alors qu'aux moyens traditionnels doivent s'ajouter les outils informatiques appropriés sans lesquels les buts visés ne peuvent plus être atteints. L'informatisation occupe ainsi une place essentielle dans cette vaste mobilisation culturelle et linguistique.

PROBLÉMATIQUE ET INTÉRÊT DE NOTRE TRAVAIL

Mais les langues ne sont pas égales devant le processus d'informatisation et les populations parlant des langues mal dotées ont un accès limité à ces nouveaux moyens, limitation pouvant aller d'une simple gêne à une incapacité totale. Les Nations Unies, élément central dans le mouvement de protection de la diversité linguistique, ont progressivement pris en compte dans leur démarche la dimension informatique et, parallèlement, de nombreuses initiatives souvent artisanales furent organisées pour informatiser des langues « peu dotées », en particulier par des groupes de développement travaillant en réseau.

Nous avons développé, avant cette thèse, plusieurs traitements de texte grand public pour le laotien, langue peu dotée informatiquement et s'écrivant avec un système d'écriture spécifique. S'appuyant sur cette expérience, la présente thèse propose une réflexion plus approfondie sur les stratégies et méthodes d'informatisation tout en se plaçant dans la perspective plus générale de l'informatisation des langues, dans le but de dégager une méthodologie multidisciplinaire pouvant s'appliquer à d'autres langues que le laotien ainsi qu'à des groupes de langues. Cette réflexion vise donc l'optimisation de l'effort d'informatisation. En effet, l'informatisation des langues peu dotées n'est pas tant une difficulté sur le plan informatique qu'une question de moyens humains et financiers pour permettre à ces populations de se munir des moyens adaptés à leurs écritures et à leurs langues.

MÉTHODOLOGIE

Le titre de cette étude évoque l'informatisation *en général* des langues *en général*. Un tel thème aurait, de loin, dépassé le cadre d'une thèse de doctorat et risqué de l'éloigner de sa matière — l'informatique — s'il avait dû être traité dans sa totalité. De nombreux compromis ont dû être consentis pour faire tenir l'étude dans son cadre. Pour résoudre cette difficulté, la diversité et la généralité du sujet ont été abordées à travers une réflexion sur les méthodologies et les techniques à mettre en œuvre pour diminuer les coûts de développement. Nous avons alors appliqué ces principes — quand cela était possible — à une « langue test » : la langue laotienne. La méthodologie proposée pour l'informatisation d'un groupe de langues a été, quant-à elle, déduite des taux de réutilisation constatés lors de plusieurs développements dérivant de logiciels existants.

ORGANISATION DE LA THÈSE

Ce mémoire est constitué de trois parties et de quatorze annexes. Dans la **première partie**, nous présentons les contextes linguistique, politique et technique de l'informatisation des langues peu dotées. En particulier, nous situons la question par rapport au mouvement des Nations Unies pour la protection des minorités et du patrimoine linguistique de l'humanité ainsi que par rapport à celui de l'internationalisation croissante et de plus en plus performante des systèmes d'exploitation (Windows, Linux...). Ce paysage général est complété par un panorama des acteurs et projets de l'informatisation des langues peu dotées. Il présente, en particulier, les intervenants mettant en œuvre cette informatisation et comment ils le font. Nous présentons six méthodes ou stratégies techniques adaptées aux difficultés des langues peu dotées et destinées à en aider l'informatisation. Enfin, nous présentons comment plusieurs de ces méthodes ont été mises en œuvre et évaluées sur la langue laotienne.

La **deuxième partie** est consacrée à la description technique des développements réalisés sur la langue laotienne. Ils mettent en œuvre plusieurs des principes exposés dans la première partie. Les développements réalisés couvrent essentiellement des services de traitement du texte, de dictionnaire électronique et d'aide à la traduction humaine. En conclusion de cette deuxième partie, nous présentons les dynamiques développées en parallèle de nos travaux sur le laotien, en particulier les expériences participatives menées et les groupes formés autour des concepts Pak Lao, LaoUniKey et LaoLex.

Nous revenons dans la **troisième partie** à un point de vue plus général, en tentant de dégager une méthodologie pour l'informatisation d'un groupe de langues. Nous l'appliquons alors au groupe des langues à écritures non segmentées d'Asie du Sud-Est, créant pour cela les outils permettant de segmenter leurs textes et les expérimentant sur les écritures birmane, khmère, laotienne et siamoise (thaïe). Nous concluons cette troisième partie avec une « étude de cas » offrant une vision concrète de ce que pourrait être un grand projet d'informatisation.

Quatorze **annexes** sont proposées.

L'annexe 1 propose des éléments classés sur des techniques d'informatisation des langues.
L'annexe 2 présente le tableau ISO 639 des codes de langues.
L'annexe 3 présente le tableau ISO 15924 des codes de systèmes d'écriture.
L'annexe 4 présente un tableau langue - systèmes d'écriture.
L'annexe 5 présente les principales familles de langues.
L'annexe 6 présente un tableau langues – familles – nombre de locuteurs – codes Ethnologue.
L'annexe 7 présente les parties du discours utilisées dans LaoDict (niveau catégorie).
L'annexe 8 présente les parties du discours, avec exemples, du laotien (niveau sous-catégorie).
L'annexe 9 présente le schéma XML Papillon pour le laotien.
L'annexe 10 présente les exemples d'articles (lexies) de LaoDict.
L'annexe 11 présente la grammaire des grammaires syllabiques.
L'annexe 12 présente l'article « Ambiguïtés irréductibles dans les monoïdes de mots ».
L'annexe 13 présente une licence GPL type.
L'annexe 14 présente un programme Prolog de génération des syllabes laotiennes.

Elles sont de natures diverses et réunissent dans un même document de nombreuses informations souvent éparses.

PREMIERE PARTIE

METHODES D'INFORMATISATION D'UNE LANGUE PEU DOTEE

I. MÉTHODES D'INFORMATISATION D'UNE LANGUE PEU DOTÉE

I.1 CONTEXTE DE L'INFORMATISATION DES LANGUES PEU DOTÉES

I.1.1 Terminologie

I.1.1.1 INFORMATISATION D'UNE LANGUE

Du mot *informatisation*, le Grand Robert de la Langue Française donne la définition : « *Introduction dans une activité des méthodes informatiques* ». Idéalement, informatiser une langue c'est donc mettre à la disposition de l'utilisateur humain tous les moyens dont il a besoin dans sa langue, qu'elle soit écrite ou non : dialogue avec la machine, outils pour écrire ou lire un texte (« en local »), envoyer un courrier électronique (« en réseau »), traduction informatisée dans une autre langue, etc.

Voici plus précisément les ressources et les logiciels que nous retiendrons ici comme cadre de l'informatisation d'une langue :

ressources :

- ⇨ **dictionnaires** :
  - bilingues,
  - d'usage,

logiciels :

- ⇨ **logiciels de traitement de la langue écrite**[1],
  - saisie et visualisation,
  - recherche et remplacement de texte,
  - sélection du texte[2],
  - tri lexicographique,
  - correction orthographique,
  - correction grammaticale,
  - correction stylistique,

- ⇨ **logiciels de traitement de l'oral** :
  - synthèse vocale,
  - reconnaissance de la parole,

- ⇨ **logiciels de traduction automatique et d'aide à la traduction de l'écrit et de l'oral**,

- ⇨ **logiciels de reconnaissance optique de caractères** (ROC),

- ⇨ **logiciels fournissant des services avancés**,
  Sont classés dans cette catégorie les logiciels peu répandus ou encore à un stade de recherche. Par exemple : saisie manuscrite, résumé automatique, génération de phrases, interrogation de bases de données en langage naturel, saisie prédictive, transcription phonétique...

[1] Il s'agit des éditeurs de texte et de leurs applications : outils de bureautique, navigateurs Internet, messageries électroniques...
[2] En particulier pour les systèmes d'écriture « non segmentés », c'est à dire des écritures ne séparant pas les mots les uns des autres. Dans ces systèmes d'écritures, la sélection des mots (à la souris ou au clavier) est plus complexe que la simple recherche des espaces ou des signes de ponctuation qui les encadrent.

⇨ **logiciels existants adaptés**.
Il s'agit de logiciels réalisés à l'origine pour des langues bien dotées et adaptés à d'autres langues, mais avec des modifications ne nécessitant pas de techniques de traitement des langues, par exemple : traduction des menus et messages dans la nouvelle langue, adaptations culturelles[1], calculs avec les chiffres vernaculaires, choix de polices compatibles avec l'encodage et la technologie d'affichage[2]. Parmi ces logiciels, citons les calculatrices, les gestionnaires de base de données, les tableurs et les outils de planification ainsi que des services tels que recherche, dépouillement et indexation de l'information. Nous incluons aussi dans cette catégorie les logiciels de comptabilité dont les spécificités, qui peuvent être très significatives, ne sont pas dues à la langue mais à la législation. Notons que l'adaptation d'un traitement de texte à une langue (par exemple l'adaptation de Word au laotien avec LaoWord, voir le chapitre I.3.2) ne sera pas classée dans cette catégorie de logiciels du fait que ses apports (clavier virtuel, sélection du texte, traduction de mots, tri lexicographique...) sont des traitements concernant la langue.

I.1.1.2 LANGUE « PEU DOTÉE »

On trouve dans la littérature, scientifique ou non, plusieurs termes pour désigner des langues moins bien informatisées que les grandes langues véhiculaires (l'anglais, l'espagnol, le français...). Les anglo-saxons emploient fréquemment les termes de *less prevalent language* (langue parmi les moins répandues) ou de *minority language* (langue minoritaire ou de minorité). Ces termes, qui ne sont pas directement liés au niveau d'informatisation, revêtent un sens fluctuant. Dans son éditorial du numéro spécial d'Elsnews[3] consacré à l'informatisation des langues minoritaires d'Europe ([Sampson 2001]), Geoffrey Sampson s'interroge sur le sens du terme *minority language*, tentant quatre définitions et concluant qu'aucune n'était entièrement satisfaisante. Ainsi, *minority language* pourra désigner aussi bien des langues parlées par quelques locuteurs seulement que des langues de minorités, qui ne sont pas nécessairement des langues avec peu de locuteurs. Par exemple, le hindi, qui est la langue d'une minorité au Royaume Uni[4], compte 366 millions de locuteurs dont c'est la première langue et 487 millions dont c'est au moins la seconde (à comparer aux chiffres équivalents pour le français : 77 et 128 millions).

Ces termes renvoient généralement à des causes de la faible informatisation mais ne la caractérisent pas. Dans la plupart des cas, ces causes (faible nombre de locuteurs, langue mal décrite...) conduisent les grands éditeurs de systèmes d'exploitation et de logiciels à ne pas intégrer les langues dans leur plan de développement. La possibilité d'exploiter un ordinateur dans ces langues – et donc l'accès aux avantages de l'informatique par le plus grand nombre – est alors tributaire d'une filière de développement parallèle (ou de l'apprentissage d'une langue comme l'anglais). Nous appellerons ces langues informatiquement peu dotées les **langues-π** (**π** pour **p**eu ou **p**as dotées), par opposition aux **langues-τ** (**τ** pour **t**rès bien dotées) et aux **langues-μ** (**μ** pour **m**oyennement dotées)[5].

[1] Par exemple, les formats de date, d'heure ou de nombre.
[2] Voir le chapitre I.1.4.3.
[3] Lettre du réseau européen pour les technologies de la langue humaine.
[4] Par exemple, dans son article intitulé « *Machine translation and minority languages* » présenté à l'*Aslib's Translating & the Computer Conference* (http://www.ling.lancs.ac.uk/monkey/ihe/mille/somers.htm, novembre 1997), Harold Somers utilise les termes de *minority languages* et de *exotic languages* pour parler de plusieurs langues parlées en Grande Bretagne, incluant l'arabe et les principales langues indiennes. Ce sens donné à « *minority language* » est aussi employé par le projet MILLE dont le site héberge l'article de Somers.
[5] Il est remarquable que certaines ex-langues-π ou μ aient récemment connu une informatisation significative, comme c'est le cas du basque, grâce à l'action d'un groupe de recherche de l'université de San Sebastian. Ces langues ont pour nous un intérêt certain car elles ont été des langues peu dotées qui ont résolu le problème de leur informatisation.

Nous proposons la technique suivante pour évaluer de manière quantitative le degré d'informatisation d'une langue. À chaque service ou ressource, un groupe d'utilisateurs représentatifs des locuteurs de la langue attribue un niveau de criticité C_k et une note N_k, la moyenne pondérée des notes — appelée **indice-σ** — reflétant leur **s**atisfaction globale[1]. Nous adoptons alors les conventions suivantes :

- ⇨ **langues-π** : moyenne entre 0 et 9,99 (peu dotées),
- ⇨ **langues-μ** : moyenne entre 10 et 13,99 (moyennement dotées),
- ⇨ **langues-τ** : moyenne entre 14 et 20 (très bien dotées).

Une langue-π est ainsi définie comme une langue dont l'indice-σ, n'atteignant pas 10/20, est encore insuffisante aux yeux de ses évaluateurs.

L'existence d'indices quantitatifs ne doit pas faire oublier que les frontières entre langues-π, μ et τ restent imprécises, ne résultant que de données subjectives. De plus, ces indices-σ peuvent être influencés négativement par des éléments non entièrement techniques comme le packaging, la documentation ou tout autre élément dont la qualité serait jugée insuffisante par rapport aux standards que connaissent les utilisateurs.

	Services / ressources	**Criticité C_k (0 à 10)**	**Note N_k (/20)**	**Note pondérée (C_kN_k)**
Traitement du texte				
	Saisie simple			
	Visualisation / impression			
	Recherche et remplacement			
	Sélection du texte [2]			
	Tri lexicographique			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral				
	Synthèse vocale			
	Reconnaissance de la parole			
Traduction				
	Traduction automatisée			
ROC				
	Reconnaissance optique de caractères			
Ressources				
	Dictionnaire bilingue			
	Dictionnaire d'usage			
Total		**ΣC_k**		**ΣC_kN_k**
Moyenne (/20)				**ΣC_kN_k / ΣC_k**

Figure 1 : Tableau d'évaluation du niveau d'informatisation d'une langue

[1] La criticité est une mesure de l'importance relative d'un service pour un groupe d'évaluation donné. Notons que les points de vue nécessairement personnels des populations évaluant les criticités et les notes rendent la méthode subjective.

[2] Pour les langues à écriture non segmentée, ce service n'est pas fourni directement par les classes de fenêtres d'édition standard.

Par rapport aux ressources et aux logiciels retenus pour définir l'informatisation d'une langue en général, nous avons exclu de ce tableau les logiciels fournissant des services avancés et les logiciels adaptés. Cela ne signifie pas que ces logiciels ne puissent pas être développés pour les langues-π — nous avons par exemple développé une transcription phonétique du laotien — ni qu'il faille viser pour elles une informatisation de qualité médiocre, mais simplement que nous les excluons de notre analyse. Les raisons en sont les suivantes.

- ⇨ Les logiciels de la première catégorie (services avancés) sont peu répandus ou répondent à des besoins moins élémentaires, y compris pour des langues bien dotées. Ils sont donc davantage impliqués dans l'informatisation des langues-τ[1].
- ⇨ Les logiciels de la seconde catégorie (logiciels adaptés) ne nécessitent pas de développements particuliers dus à la langue mais uniquement des adaptations.

À titre d'exemple, nous avons demandé à Pierre Sein-Aye, pionnier de l'informatisation du birman, et à Michel Antelme, responsable de l'enseignement du khmer à l'INALCO, de compléter ces tableaux d'évaluation du niveau d'informatisation pour le birman et pour le khmer, en fonction de leur connaissance des logiciels et des ressources existants. Leurs évaluations sont contenues dans les deux tableaux ci-dessous.

	Services / ressources	Criticité (0 à 10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	16	160
	Recherche et remplacement	8	0	0
	Sélection du texte	8	16	128
	Tri lexicographique	6	0	0
	Correction orthographique	4	0	0
	Correction grammaticale	4	0	0
	Correction stylistique	2	0	0
Traitement de l'oral				
	Synthèse vocale	2	0	0
	Reconnaissance de la parole	2	0	0
Traduction				
	Traduction automatisée	6	0	0
ROC				
	Reconnaissance optique de caractères	8	0	0
Ressources				
	Dictionnaire bilingue	8	0	0
	Dictionnaire d'usage	4	0	0
Total		**82**		**448**
Moyenne (/20)				**448 / 82 = 5,46**

Figure 2 : Tableau d'évaluation du niveau d'informatisation pour le birman

[1] Nous avons cependant hésité pour certains services tels que la saisie manuscrite ou la saisie prédictive qui entrent actuellement en force dans les PDA (Personal Digital Assistant) et les téléphones portables. À titre d'exemple, ces deux technologies sont en cours d'expérimentation au NECTEC (centre de recherches situé à Bangkok) pour la langue thaïe qui est une langue déjà assez bien dotée. L'évolution rapide des technologies et des marchés rend très fluctuante la frontière entre service avancé et service répandu.

	Services / ressources	Criticité (0 à 10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	14	140
	Recherche et remplacement	8	12	96
	Sélection du texte	6	12	72
	Tri lexicographique	5	0	0
	Correction orthographique	2	0	0
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	4	32
ROC				
	Reconnaissance optique de caractères	9	0	0
Ressources				
	Dictionnaire bilingue	10	4	40
	Dictionnaire d'usage	10	0	0
Total		**88**		**540**
Moyenne (/20)				**540 / 88 = 6,14**

Figure 3 : Tableau d'évaluation du niveau d'informatisation pour le khmer

I.1.2 Les langues : de l'ordre dans la diversité

I.1.2.1 LANGUES DANS LE MONDE

La base de données « Ethnologue » (http://www.ethnologue.com) recense plus de 6800 langues dans le monde, réparties géographiquement de la manière suivante.

	Langues	**Pourcentage**
Amérique	1 013	15 %
Afrique	2 058	30 %
Europe	230	3 %
Asie	2 197	32 %
Pacifique	1 311	19 %
TOTAL	**6 809**	

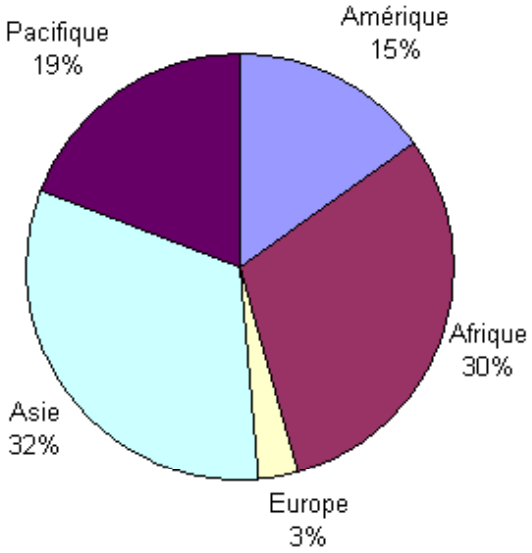


Figure 4 : Répartition des langues dans le monde (source Ethnologue, 2000)

Réunissant de nombreuses informations sur les langues du monde, cette base de données est le résultat de plus de cinquante années d'un travail de synthèse débuté par Richard Pittman. Ce travail s'appuie sur de nombreuses sources incluant aujourd'hui 828 publications et sur de nombreuses contributions individuelles bénéficiant de l'expérience de terrain de SIL International[1], organisme très lié à Ethnologue. Les informations fournies sur le site Ethnologue sont aussi diffusées sous forme papier [Grimes & Grimes 2000]. En dépit de la précision affichée, ces chiffres sont le résultat de travaux difficiles, en particulier du fait de l'hétérogénéité des sources et de la délicate distinction entre langue et dialecte. Généralement, seules les langues font l'objet de statistiques (voir, par exemple, [Breton 1995]) mais leur définition s'appuie sur des critères variables et subjectifs comme la volonté des locuteurs, l'existence d'une tradition d'écriture et de littérature, la non-intercompréhension avec d'autres langues, le statut institutionnel et l'enseignement ([Dubois et al. 1994]).

[1] SIL International — ex Summer Institute of Linguistics — a commencé son activité en 1934 (voir les sites www.sil.org et www.ethnologue.org et le chapitre I.1.3.2.3).

En fait, les organismes ayant mené à bien cette tâche de dénombrement au niveau mondial sont rares et la base de données Ethnologue, gérée par SIL International depuis 1971, est aujourd'hui la source de statistiques sur les langues la plus citée, en dépit des sérieuses critiques qui sont faites à cet organisme ([Calvet 1987], pages 205 à 217). Ethnologue a néanmoins eu un précurseur illustre : « Les langues du monde » ([Meillet & Cohen 1952]). Dans cet ouvrage collectif qui compile un important travail de recensement et de description des langues, Marcel Cohen confirme avec modestie et prudence plusieurs estimations antérieures donnant une fourchette de 2500 à 3500 langues dans le monde, en se limitant à celles « qui ne sont pas de simples parlers locaux ». Ce travail collectif publié en 1952 avait débuté en 1938 et repartait d'une première édition publiée en 1924. Les chercheurs ayant collaboré à ce travail de pionniers sont des linguistes de grand renom : Benveniste, Bloch, Caquot, Cohen, Delafosse, Demiéville, Denys, Dumézil, Faublée, Guiart, Haguenauer, Haudricourt, Jakobson, Lacombe, Leenhardt, Loukotka, Maspéro (Henri), Perrot, Rivet, Sauvageot, Schmidt, Sinor, Stresser-Péan, van Bulck et Vendryes. Quelques travaux comparables ont eu lieu depuis, en particulier ceux de Charles et Florence Voegelin [Voegelin & Voegelin 1977], de Heinz Kloss et Grant McConnel [Kloss et al. 1974-1984], de Marius Sala et Ioana Vintila-Radulescu [Sala & Vintila-Radulescu 1984], de Merritt Ruhlen [Ruhlen 1987] ou de David Dalby [Dalby D. 1999]. D'autres travaux, plus détaillés mais sur moins de langues, offrent une vision plus concrète des langues et des familles de langues, par exemple [Malherbe 1983], [Campbell 1991], [Comrie 1991], [Moseley & Asher 1994], [Katzner 1995] et [Dalby A. 1998].

Certains auteurs parlent de beaucoup moins de langues que les 6809 citées par Ethnologue : David Dalby et Claude Hagège en retiennent cinq mille ([Dalby 1999], [Hagège 2000]), Charles et Florence Voegelin environ quatre mille cinq cents ([Voegelin & Voegelin 1977]), Michel Malherbe et Marius Sala / Ioana Vintila-Radulescu trois mille ([Malherbe 1983], [Sala & Vintila-Radulescu 1984]). Le nombre plus important avancé par Ethnologue s'explique en partie par la différence de point de vue sur ce qu'est une langue, leur énumération comprenant « nombre de parlers qu'ailleurs on classerait dialectes » ([Breton 1995], page 115) mais aussi par une amélioration, au fil des années, de notre connaissance des langues de certaines parties du monde comme l'île de Nouvelle-Guinée où l'on dénombre aujourd'hui 1086 langues ([Grimes & Grimes 2000]). Ainsi, le nombre de langues ne cesse d'augmenter dans les éditions successives d'Ethnologue — 6 528 en 1992, 6 703 en 1996, 6 784 en 1999 ([Breton 2003], page 15) — malgré la disparition de dix langues ([Calvet 2002][1]) à vingt-cinq langues ([Hagège 2000]) chaque année.

Sur le plan normatif, deux listes de langues du monde ont été réunies dans un standard, l'ISO 639[2], qui fournit des codes pour les noms de langues. La première liste (ISO 639-1[3]), dont les codes sont sur deux lettres, a pour domaine d'application la terminologie. La seconde (ISO 639-2[4]), sur trois lettres, couvre la documentation bibliographique et la terminologie, incluant ainsi la première. Ce standard ISO 639-2, qui ne contient que 470 codes (voir la liste en annexe A.2) parmi lesquels se trouvent des codes de langues mortes ou correspondant à des familles de langues, est loin de couvrir toutes les langues dénombrées par Ethnologue. C'est pourquoi cette organisation utilise sa propre liste de codes sur trois lettres pour désigner les langues de façon non ambiguë. Cette liste est téléchargeable (http://www.ethnologue.com/codes/) et peut être utilisée comme base pour d'autres travaux comme ont déjà choisi de le faire les projets *OLAC*[5] (*Open Language Archives Community*), *the Linguist List*[6] et *Rosetta*[7].

[1] Page 116, citant le numéro d'avril 2000 du Courrier de l'UNESCO
[2] Ce standard est le résultat du groupe de travail commun constitué par l'ISO TC37/SC2 et l'ISO TC46/SC4. Voir http://linux.infoterm.org/iso-e/i-iso.htm.
[3] Standard maintenu par le Centre d'Information International pour la Terminologie (http://linux.infoterm.org/).
[4] Standard maintenu par la Librairie du Congrès, http://www.loc.gov/standards/iso639-2/ et dérivé de la liste MARC de ce même organisme (voir http://www.loc.gov/marc/ et http://www.loc.gov/marc/languages/). Les débuts de la liste de codes MARC remontent à 1968.
[5] Voir http://www.language-archives.org/ et aussi http://www.language-archives.org/wg/language-codes/.
[6] Voir http://www.linguistlist.org/.
[7] Voir http://www.rosettaproject.org/.

I.1.2.2 ORALITÉ ET « LITTÉRISATION »

Plus encore que de dénombrer les langues elles-mêmes, il semble qu'il soit difficile de faire la part de celles qui sont écrites et de celles qui ne le sont pas. Nous n'avons pu trouver que des statistiques non étayées et très variables — selon les sources, de 67 à 90 % des langues seraient non écrites, ce qui donnerait en moyenne seulement 1500 langues écrites, sur les 6 809 langues du monde ! L'Alliance Biblique Universelle indique, quant à elle, que la Bible a été traduite au moins partiellement dans 2 355 langues[1]. Le rapport sur les langues du monde préparé par Linguapax pour l'UNESCO clarifiera peut-être la situation (voir le chapitre I.1.3.1.2.1). Ce rapport, demandé en 1996, est attendu au printemps 2004.

Louis-Jean Calvet nuance d'ailleurs les situations de « scripturalité » et d'oralité en distinguant [Calvet 1987, p. 59-62] :

- ⇨ les sociétés à tradition écrite ancienne où la langue écrite est la langue parlée (français, même s'il y a une différence dans le style entre langue écrite et langue parlée),
- ⇨ les sociétés à tradition écrite ancienne où la langue écrite n'est pas la langue parlée (arabe),
- ⇨ les sociétés à langue récemment écrite (certaines situations post-coloniales),
- ⇨ les sociétés de tradition orale.

Dans les deux dernières situations, le recours à l'écrit n'est pas naturel, comme le montre l'exemple donné [Calvet 1987, p.61] de l'évolution post-coloniale de l'Algérie et du Mali, le premier ayant remplacé le français par l'arabe à l'école et le second commençant à peine à enseigner certaines langues africaines.

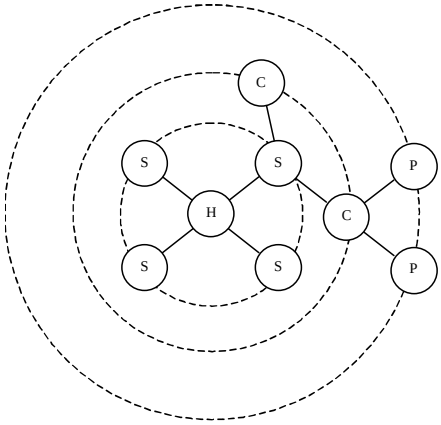
Roland Breton rappelle, quant à lui, que « la 'littérisation' d'une langue (le fait de lui donner une écriture adaptée à sa phonologie, codifiée et normalisée par des règles d'orthographe), est un premier pas technique » qui doit être accompagné d'une volonté à la fois de la part de ses locuteurs de l'utiliser et de celle « des institutions d'organiser et de financer son enseignement » ([Breton 2003], page 43). Notons que cette première phase technique est rarement entièrement réalisée, en particulier pour la normalisation de l'orthographe. Louis-Jean Calvet cite le cas du mot « huit » en mandingue qui s'écrit *segin* au Mali, *seyin* en Guinée et *séegin* au Burkina Faso ([Calvet 1999], page 220). À l'intérieur d'un même pays à tradition grammairienne comme la France, de nombreux mots s'écrivent de plusieurs manières comme *rancart* (rancard, rencart, rencard) ou encore *gnole* (gniole, gnôle, gniaule)[2].

Dans notre travail, nous nous sommes limité aux langues écrites, laissant de côté langues de tradition orale, langues artificielles et langues des signes. Citons seulement ici, en ce qui concerne les langues non écrites, le projet « de constitution et de diffusion d'une archive de documents linguistiques son/texte » de la fédération de recherche *Typologie et universaux linguistiques*, composée en particulier du LACITO (http://lacito.vjf.cnrs.fr/), du LLACAN (http://llacan.cnrs-bellevue.fr/) et du CRLAO (http://www.ehess.fr/centres/crlao/crlao.html).

[1] Statistiques 2003. Voir http://www.biblesociety.org/latestnews/latest273-slr2003stats.html.
[2] Exemples donnés par Nina Catach (PUF 1978), citée dans [Perret 1998], page 136.

I.1.2.3 MODÈLE GRAVITATIONNEL ET CONFIGURATION GÉNÉTIQUE

Sur les 6809 langues parlées dans le monde, plus de la moitié est parlée par moins de 10 000 locuteurs représentant au total moins de 0,3 % de la population du monde. Plusieurs linguistes se sont intéressés à des questions de politique linguistique et de protection des langues minoritaires. Louis-Jean Calvet a proposé quelques instruments « ayant pour but de mettre de l'ordre dans ce désordre babélien » [Calvet 2002, p. 26-34]. Parmi ces instruments, « le modèle gravitationnel part du principe que les langues sont reliées entre elles par des personnes bilingues et que les systèmes de bilinguisme sont déterminés par des rapports de force ». Les langues que Calvet appelle périphériques graviteront ainsi autour de langues centrales qui seront des pivots pour communiquer avec d'autres langues. Ces « cent à deux cents langues centrales » graviteront à leur tour autour d'une langue supercentrale (une dizaine, incluant espagnol, français, hindi-urdu, arabe, indonésien-malais, russe, portugais) ou de l'anglais, la langue hypercentrale, pivot des langues supercentrales.



H : langue hypercentrale (l'anglais)

S : langues supercentrales (≈ 10)

C : langues centrales (100 à 200)

P : langues périphériques (> 6 000)

Figure 5 : Modèle gravitationnel

Cette description permet de faire ressortir l'importance du bilinguisme dans les rapports entre les langues. Par exemple, lorsque l'on cherche des ressources pour informatiser une langue minoritaire, on s'aperçoit que les dictionnaires bilingues ne sont pas forcément faits avec des langues de la même famille mais plutôt avec des langues centrales ou supercentrales. Par exemple, il existe bien plus de dictionnaires laotien-français ou laotien-anglais que de dictionnaires laotien-shan ou laotien-yuon, le shan et le yuon étant pourtant des langues voisines du laotien mais avec lesquelles le laotien a peu de contacts. De nombreux dictionnaires bilingues ont d'ailleurs été le fait de puissances coloniales ou impérialistes souhaitant étendre leur sphère d'influence. Le modèle gravitationnel peut aussi donner une indication sur la méthode à suivre pour informatiser une langue et suggère que les langues centrales auront intérêt à être informatisées en priorité, permettant ensuite un accès plus facile aux langues périphériques associées. De plus, les difficultés linguistiques rencontrées pourront être mieux surmontées par des locuteurs natifs parlant une ou plusieurs langues pivots.

Parmi les autres instruments proposés par Louis-Jean Calvet, nous retiendrons la *configuration génétique* qui concerne les langues liées par une parenté génétique, c'est à dire les langues dérivées d'une même souche[1] (langues de la famille indo-européenne, sino-tibétaine...). Des développements réalisés pour une langue-τ pourront ainsi être assez facilement adaptés à une langue-π ou μ suffisamment proche. La proximité devra cependant être évaluée au coup par coup, l'appartenance à une même famille, voire à un même groupe, n'étant pas systématiquement suffisante pour être exploitable. Un tableau rappelant les principales familles de langues est proposé en annexe A.5. Il est tiré de [Grimes & Grimes 2000].

[1] [Dubois et al. 1994]. Une autre parenté linguistique — l'affinité entre langues — provient du contact géographique entre langues de familles *a priori* différentes. Du fait que ces langues viennent de substrats différents, les ressemblances syntaxique et lexicale qui les lient seront généralement plus limitées et ainsi plus difficiles à exploiter.

I.1.2.4 DROITS DES LANGUES ET COROLLAIRE INFORMATIQUE

I.1.2.4.1 Droits linguistiques

Outre les aspects techniques des langues, nous nous sommes intéressé à leur dimension juridique et aux conséquences de cette dimension sur leur informatisation.

L'article 5 de la **déclaration universelle des droits linguistiques (**[CMDL 1996]**)**[1]**,** proclamée en 1996, affirme le principe « que les droits de toutes les communautés linguistiques sont égaux et indépendants du statut juridique ou politique de leur langue en tant que langue officielle, régionale ou minoritaire ».

Dans son article 40, elle proclame :

« Toute communauté linguistique a le droit de disposer, dans le domaine de l'informatique, d'équipements adaptés à son système linguistique et d'outils de production dans sa langue, afin de profiter pleinement du potentiel qu'offrent ces technologies pour l'autoexpression, l'éducation, la communication, l'édition, la traduction, et en général le traitement de l'information et de la diffusion culturelle. »,

et dans son article 47, alinéa 2, que :

« Tout membre d'une communauté linguistique a le droit de disposer dans sa langue de tous les moyens que requiert l'exercice de l'activité professionnelle, comme par exemple les documents et les livres de consultation, les instructions, les imprimés, les formulaires et équipements, les outils et les programmes informatiques. ».

Malgré les prestigieuses signatures apposées à cette déclaration, comme par exemple celles de Noam Chomsky, Nelson Mandela, Desmond Tutu, Shimon Peres, Yasser Arafat ou encore du Dalaï Lama, cette reconnaissance d'un droit des langues n'a cependant pas le caractère juridique qu'à ce niveau seules les Nations Unies peuvent conférer à un texte. En effet, même si les déclarations et recommandations des Nations Unies ne sont que du « droit vert » — elles n'ont pas force de loi — elles forment un code de bonne conduite qu'il convient que les États suivent car elles reflètent la volonté de la communauté internationale, et ce d'autant plus qu'elles ont été adoptées à une forte majorité.

Les Nations Unies se sont intéressées au droit des minorités linguistiques[2] depuis très longtemps puisque la **déclaration universelle des droits de l'homme** proclamée en 1948 affirme que « chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente Déclaration, sans distinction aucune, notamment [...], de langue, [...] ou de toute autre situation ». D'autres étapes importantes ont suivi. De nombreux pactes, recommandations, conventions et déclarations concernant les droits linguistiques des peuples ont été énoncés depuis. Une synthèse de ces éléments se trouve dans le document « L'éducation dans un monde multilingue » de l'UNESCO[3]. Nous citerons deux déclarations pour souligner l'évolution récente.

[1] http://www.linguistic-declaration.org/versions/frances.pdf, http://www.egt.ie/udhr/udlr-fr.html.
Cette conférence organisée à l'initiative du Comité des traductions et droits linguistiques du PEN Club International et du CIEMEN (Centre International Escarré pour les Minorités Ethniques et les Nations), avec le soutien moral et le support technique de l'UNESCO.

[2] Le livre de Joseph Yacoub « Les minorités dans le monde » ([Yacoub 1998]) propose une étude très documentée du problème des minorités, analysé globalement puis région par région, et développant la question des droits linguistiques.

[3] Texte : http://unesdoc.unesco.org/images/0012/001297/129728f.pdf.

En 1992, alors que l'Europe diffusait sa charte des langues régionales ou minoritaires (voir ci-dessous), l'ONU adoptait la **déclaration des droits des personnes appartenant à des minorités nationales ou ethniques, religieuses et linguistiques**[1] qui a pour but de protéger les minorités, y compris sur le plan linguistique. Cette déclaration demande en particulier que les Etats protègent l'existence et l'identité [...] linguistique des minorités (article 1.1) et, plus loin (article 4.2), [qu'ils] prennent des mesures pour créer des conditions propres à permettre aux personnes appartenant à des minorités d'exprimer leurs propres particularités et de développer leur culture, leur langue, leurs traditions et leurs coutumes, sauf dans le cas de pratiques spécifiques qui constituent une infraction à la législation nationale et sont contraires aux normes internationales.

En 2001, la **déclaration universelle de l'UNESCO sur la diversité culturelle**[2] allait beaucoup plus loin en demandant à œuvrer à la sauvegarde du patrimoine linguistique de l'humanité et en s'engageant dans la voie de l'informatique, affirmant (Article 12) :

Les États membres s'engagent à prendre les mesures appropriées pour diffuser largement la Déclaration [...] et pour encourager son application effective, en coopérant notamment à la réalisation des objectifs suivants : [...]

> 5. sauvegarder le patrimoine linguistique de l'humanité et soutenir l'expression, la création, et la diffusion dans le plus grand nombre possible de langues,
>
> 6. encourager la diversité linguistique - dans le respect de la langue maternelle - à tous les niveaux de l'éducation, partout où c'est possible, et stimuler l'apprentissage du plurilinguisme dès le plus jeune âge[3], ...
>
> 9. encourager l'« alphabétisation numérique » et accroître la maîtrise des nouvelles technologies de l'information et de la communication, qui doivent être considérées aussi bien comme des disciplines d'enseignement que comme des outils pédagogiques susceptibles de renforcer l'efficacité des services éducatifs,
>
> 10. promouvoir la diversité linguistique dans l'espace numérique et encourager l'accès universel, à travers les réseaux mondiaux, à toutes les informations qui relèvent du domaine public,
>
> 11. lutter contre la fracture numérique - en étroite coopération avec les institutions compétentes du système des Nations Unies - en favorisant l'accès des pays en développement aux nouvelles technologies, en les aidant à maîtriser les technologies de l'information et en facilitant à la fois la circulation numérique des produits culturels endogènes et l'accès de ces pays aux ressources numériques d'ordre éducatif, culturel et scientifique, disponibles à l'échelle mondiale.

Les **déclarations des Nations Unies**, sans aller aussi loin que la **déclaration universelle des droits linguistiques** qui prône explicitement l'égalité des langues, montrent donc une volonté claire de sauvegarder l'ensemble du patrimoine linguistique actuel. Ainsi, l'UNESCO s'est-elle plus récemment inquiétée de la disparition accélérée de langues qui, selon les prévisions les plus pessimistes, pourrait avoir atteint 90% des langues existant actuellement d'ici la fin du siècle[4]. En effet, 96% des habitants de la planète ne parlant que 4% des langues, l'augmentation des échanges en langues véhiculaires, en désenclavant les régions les plus isolées, risque fort de faire reculer la pratique des langues vernaculaires. Le rapport d'experts et de linguistes remis à l'UNESCO le 25 mars 2003 recommande la mise en œuvre de plans d'action dans ce domaine.

[1] Résolution 47/135. Texte : http://www.droitshumains.org/Biblio/Trait_internat/Discrim_12.htm.
[2] Texte : http://www.unesco.org/culture/pluralism/diversity/html_fr/decl_fr.shtml.
[3] Il semble paradoxal d'encourager l'apprentissage de plusieurs langues dès le plus jeune âge car cela entraînera plutôt un abandon des langues périphériques au profit des langues centrales, supercentrales et de l'anglais.
[4] Une étude de l'UNESCO a permis de tirer un premier bilan sur les langues les plus menacées. [Wurm 2001] en présente une répartition géographique pour plus de 700 langues dont 13 en France métropolitaine.

Malgré sa portée géographique plus restreinte, la charte des langues régionales ou minoritaires, adoptée par le Comité des ministres du **Conseil de l'Europe** en 1992, est allé plus loin dans le domaine juridico-linguistique. Traité destiné à reconnaître, sauvegarder et promouvoir les langues minoritaires européennes[1], la Charte est entrée officiellement en vigueur le 1er mars 1998 et dix-sept États sont à ce jour contractants du traité : Allemagne, Arménie, Autriche, Chypre, Croatie, Danemark, Espagne, Finlande, Hongrie, Liechtenstein, Norvège, Pays-Bas, Royaume-Uni, Slovaquie, Slovénie, Suède et Suisse. Onze autres n'en sont que signataires : Azerbaïdjan, France, Islande, Italie, Luxembourg, Malte, Moldova, Pologne, Roumanie, République tchèque, Macédoine, Russie, Ukraine.

Au-delà de celles des Nations Unies et du Conseil de l'Europe, ce sont les politiques linguistiques de cent vingt pays et de quatre-vingt-cinq États non souverains qui ont été décrites dans [Gauthier et al. 1994] par François Gauthier, Jacques Leclerc et Jacques Maurais. Le site de Jacques Leclerc, « L'aménagement linguistique dans le monde »[2], offre une présentation mise à jour de ces politiques linguistiques des pays du monde.

I.1.2.4.2 Facette informatique des droits linguistiques

Bien que l'informatisation apparaisse en filigrane dans ces textes et en particulier dans la déclaration universelle sur la diversité culturelle, rien d'explicite n'y est dit sur les moyens à mettre en œuvre pour y parvenir. La prise en compte d'un volet informatique, nécessaire pour l'application de ces principes, semble avoir du mal à se concrétiser. Le rôle de l'UNESCO dans l'informatisation des langues apparaît à partir de 1995, en particulier dans les décisions adoptées par le Conseil Exécutif à sa 150e session tenue en octobre 1996[3].

Un « projet de recommandation concernant la promotion et l'usage du multilinguisme et l'accès universel au cyberespace »[4] a été adopté lors de la 32e session de la conférence générale de l'UNESCO qui a eu lieu en **septembre-octobre 2003**. Il avait été présenté pour la première fois en 1997 lors de la 29e session.

Ce texte, dans sa dernière mouture, stipule en particulier :

1. Les secteurs public et privé et la société civile, aux niveaux local, national, régional et international, devraient s'efforcer de fournir les ressources nécessaires et prendre les mesures requises pour atténuer les obstacles linguistiques et promouvoir l'interaction humaine sur Internet en encourageant **la création et le traitement des contenus éducatifs, culturels et scientifiques sous forme numérique**, et l'accès à ces contenus, de façon à assurer que toutes les cultures puissent **s'exprimer et avoir accès au cyberespace dans toutes les langues, y compris les langues autochtones**.

2. Les Etats membres et les organisations internationales devraient encourager et appuyer le renforcement des capacités de production de **contenus locaux et autochtones sur Internet**.

[1] http://conventions.coe.int/Treaty/FR/searchsig.asp?NT=148&CM=8&DF=22/12/03.
[2] http://www.tlfq.ulaval.ca/axl/, hébergé sur le site du « Trésor de la Langue Française au Québec ».
[3] http://unesdoc.unesco.org/images/0010/001044/104447F.pdf page 33.
[4] http://www.unesco.org/general/fre/about/circulars/cl3653.pdf.

3. Les Etats membres devraient formuler des **politiques nationales appropriées sur la question cruciale de la survie des langues dans le cyberespace** en vue de promouvoir l'enseignement des langues, y compris les langues maternelles, dans le cyberespace. L'appui et l'assistance internationale aux pays en développement devraient être renforcés et élargis pour faciliter la **conception de matériel librement accessible sur l'enseignement des langues sous forme électronique** et **l'amélioration des compétences humaines dans ce domaine**.

4. Les Etats membres, les organisations internationales et les entreprises spécialisées dans les technologies de l'information et de la communication devraient **encourager la recherche-développement, suivant des modalités de collaboration participative, pour la mise au point de systèmes d'exploitation, de moteurs de recherche et de navigateurs Web dotés de grandes capacités multilingues, ainsi que leur adaptation aux conditions locales**. Ils devraient appuyer les efforts internationaux de coopération relatifs aux **services de traduction automatisée accessibles à tous**, ainsi qu'aux **systèmes linguistiques intelligents** tels que ceux qui remplissent des **fonctions multilingues de recherche de l'information, de dépouillement/résumé et de reconnaissance de la parole**, tout en respectant pleinement le droit de traduction des auteurs.

Grâce à l'importance prise par Internet, les Nations Unies — qui voient l'informatisation essentiellement à travers ce prisme — se sont enfin dotés d'un premier outil normatif dans le domaine de l'informatisation des langues-π[1].

La contribution des Nations Unies à la législation en faveur des langues peu dotées informatiquement avait eu un précédent avec la **déclaration du Millénaire**[2]. Ce texte adopté par l'Assemblée Générale le **8 septembre 2000** (résolution 55/2) comporte, dans son chapitre « Développement et éradication de la pauvreté » point 20, un volet consacré à l'informatique : « We also resolve… to ensure that the benefits of new technologies, especially information and communication technologies, in conformity with recommendations contained in the ECOSOC 2000 Ministerial Declaration, are available to all. ». L'ECOSOC 2000[3] est le rapport E/2000/52 de la session du Conseil Economique et Social tenue à New York du 5 juillet au 1er août 2000 et qui énonce les éléments clés d'un programme d'action en faveur des technologies de l'information pour le développement. Il incluait dans ses recommandations des éléments en faveur du développement de contenu local, par exemple (point 16) : « Encourage cultural and linguistic diversity and local content in cyberspace, drawing on the local know-how and indigenous knowledge ».

Suite à cet événement fut créée, en **novembre 2001**, la **commission spéciale des Nations Unies pour les technologies de l'information et des télécommunications**[4] (*United Nations Information and Communication Technologies Task Force*), sur proposition du Conseil Économique et Social. Ce groupe, constitué de représentants gouvernementaux, de la société civile, du secteur privé et des organisations et agences du système des Nations Unies, a reçu pour mission de rapprocher les intérêts des uns et des autres pour réaliser les objectifs de la déclaration.

[1] L'informatisation de ces langues avait pourtant commencé depuis longtemps, en particulier celles à écritures non latines. Bien avant l'avènement d'Internet, elle répondait au simple besoin des locuteurs d'utiliser un ordinateur pour écrire courriers et informations diverses dans leur langue – ce qui reste d'ailleurs un besoin non satisfait aujourd'hui pour plusieurs systèmes d'écriture – et s'est développée considérablement avec l'apparition des micro-ordinateurs au début des années 80 puis des polices TrueType qui permettaient plus de souplesse que les polices bitmap et vectorielles.
[2] http://www.un.org/millennium/declaration/ares552e.pdf.
[3] http://www.un.org/documents/ecosoc/docs/2000/e2000-52.pdf.
[4] http://www.unicttaskforce.org/.

D'autres grandes institutions des Nations Unies contribuent à réaliser l'objectif « informatique » de la déclaration du Millénaire, comme la **Banque Mondiale** avec son département sur les technologies globales de l'information et des télécommunications[1] ou le **PNUD** à travers son programme de réseau de développement durable[2]. Le **sommet mondial sur la société de l'information**, en **décembre 2003**[3], fut l'occasion de communiquer sur l'importance de cette société pour le développement dans le monde[4].

D'autres textes ont vu le jour récemment, en particulier :

- ⇨ le « projet de charte sur la préservation du patrimoine numérique »[5] adopté, comme le « projet de recommandation concernant la promotion et l'usage du multilinguisme et l'accès universel au cyberespace », lors de la 32e session de la conférence générale de l'UNESCO, en **octobre 2003**,
- ⇨ le « projet de plan d'action (WSIS/PC-3/DT/5) » présenté à Genève en **décembre 2003**, lors de la première phase du sommet mondial sur la société de l'information, en accompagnement d'un « projet de déclaration de principe »[6].

Si ces textes, et les projets qui leur sont associés, encouragent la sauvegarde du patrimoine numérique ainsi que l'élaboration de logiciels et l'utilisation d'Internet dans les langues locales, ils n'incluent pas encore explicitement de volet « traitement des langues ».

[1] http://info.worldbank.org/ict/.
[2] http://www.sdnp.undp.org/.
[3] http://www.itu.int/wsis/.
[4] http://www.ict-4d.org/.
[5] http://unesdoc.unesco.org/images/0013/001311/131178f.pdf.
[6] http://www.itu.int/wsis/documents/doc_multi.asp?lang=fr&id=1104|1106.

I.1.3 Les acteurs : des projets et des hommes

I.1.3.1 LES DONNEURS D'ORDRE

I.1.3.1.1 Typologie

Beaucoup de projets de développement, et en particulier dans le domaine de l'informatique et des télécommunications, reposent sur des fonds publics ou sur des dons. Lorsque les projets sont le fait d'entités politiques (États, Union Européenne...), les objectifs sont généralement de trois ordres :

- ⇨ l'aménagement de leur politique linguistique interne,
- ⇨ le développement d'outils linguistiques dans leur sphère d'influence,
- ⇨ la promotion d'une langue nationale à l'extérieur des frontières.

Il existe aussi des donateurs privés tels que les fondations Rockefeller (http://www.rockfound.org), Toyota (http://www.toyotafound.or.jp/etop.htm), Ford (http://www.fordfound.org/) ou MacArthur (http://www.macfound.org/).

De nombreux projets de développement dans le domaine des technologies de l'information et des communications sont recensés par AiDA (Accessible Information on Development Activities, http://aida.developmentgateway.org), « point d'information commun sur les activités des principaux donneurs internationaux, de certaines organisations de la société civile et de fondations privées ».

I.1.3.1.2 Actions des Nations Unies

I.1.3.1.2.1 L'UNESCO

Comme nous l'avons précisé au chapitre précédent, l'UNESCO a encore peu œuvré à la mise en place de projets d'informatisation, malgré son récent rapprochement avec SIL International.

Citons néanmoins la création de Linguapax[1] et le rapport sur les langues du monde qui lui a été demandé en 1996. Linguapax est un institut (organisation non gouvernementale) créé par l'UNESCO en 1987 et situé à Barcelone. Il a actuellement pour objectifs de :

- ⇨ Promouvoir l'information et la recherche sur les politiques linguistiques,
- ⇨ Conseiller les responsables de la politique linguistique des gouvernements nationaux ou régionaux,
- ⇨ Encourager l'éducation multilingue et perfectionner les méthodes d'apprentissage.
- ⇨ Relier l'éducation multilingue à la culture de la paix,
- ⇨ Offrir une assistance technique aux communautés linguistiques minoritaires ou minorisées,
- ⇨ Défendre les droits linguistiques comme des droits individuels et collectifs de la personne,
- ⇨ Faciliter la gestion de la diversité des langues dans les villes réunissant plusieurs communautés linguistiques,
- ⇨ Contribuer à la présence du multilinguisme dans le cyberespace.
- ⇨ Créer de nouveaux mécanismes de collaboration entre ONG spécialisées.

[1] http://www.unesco.org/education/educprog/linguapax/homepage/menupage.htm.

Linguapax a par exemple organisé un congrès mondial sur les politiques linguistiques en avril 2002 à Barcelone dans lequel l'un des ateliers était dédié aux technologies des langues et intitulé *« Les nouvelles technologies de l'information et les langues de démographie limitée et moyenne. Les défis lancés par les nouvelles technologies et la production de ressources linguistiques pour la promotion de la diversité linguistique et de la culture de la paix »*. Ses conclusions[1] incluent des recommandations intéressantes :

- ⇨ il est important de consolider les ressources existantes et les projets à venir devraient faire l'objet de collaborations,
- ⇨ les différentes organisations qui travaillent dans le domaine des langues et de la technologie (par exemple l'UNESCO, le projet Atlantis, etc.) devraient travailler en étroite coopération afin de mettre les ressources à la disposition d'autres usagers potentiels,
- ⇨ nous croyons que Linguapax devrait s'impliquer dans la coordination des initiatives visant à élaborer des bases de données linguistiques mondiales, qui sont actuellement dispersées,
- ⇨ Linguapax devrait encourager la réutilisation des ressources existantes et favoriser les effets multiplicateurs, c'est à dire qu'un même logiciel puisse être utilisé pour produire du matériel dans de nombreuses langues,
- ⇨ nous recommandons aussi que les logiciels libres soient utilisés pour rendre les ressources accessibles au public le plus large possible.

L'apport le plus important de Linguapax sur le plan pratique sera probablement le « rapport sur les langues du monde »[2] que l'ancien Directeur Général de l'UNESCO, Federico Mayor, lui a demandé en 1996, lors d'un séminaire sur les politiques linguistiques organisé à Leioa au Pays basque. Il s'agit d'élaborer « [un rapport] qui décrive notre richesse [linguistique] et explique les problèmes qui affectent les langues dans les différentes régions du monde. ». Plus précisément, les objectifs du rapport sont les suivants :

- ⇨ décrire l'état des langues du monde,
- ⇨ analyser et évaluer les problèmes qui affectent les langues,
- ⇨ faire des recommandations destinées à conserver et à valoriser le patrimoine linguistique mondial.

L'élaboration du rapport sur les langues du monde, coordonné par UNESCO Etxea, le centre UNESCO du Pays basque, se trouve dans la phase d'édition. Il est attendu au printemps 2004.

Un autre programme de l'UNESCO, lancé fin 1999 et intitulé *Initiative B@bel* [3] est davantage orienté vers l'informatisation des langues. Nous reproduisons ci-dessous sa description par l'UNESCO.

L'information du domaine public est un bien public mondial ; à moins d'un soutien actif du public, l'offre de ce bien sera insuffisante. Sachant cela, l'UNESCO se fixe pour objectif principal de redéfinir l'universalisation de l'accès à l'information dans toutes les langues dans le cyberespace, en encourageant (1) l'élaboration d'instruments (systèmes de traduction ; outils terminologiques ; protocoles ; etc.) qui faciliteront la communication multilingue dans le cyberespace ; (2) l'attribution d'une juste part des ressources publiques aux fournisseurs d'information publique et (3) l'élargissement de l'accès à l'information et aux connaissances relevant du domaine public qui existe dans toutes les langues.

[1] http://www.linguapax.org/congres/Conclusions/con5_fr.html.
[2] http://www.unesco.org/education/educprog/linguapax/homepage/report.htm, http://www.linguapax.org/fr/monfr.html.
[3] http://webworld.unesco.org/imld/babel_fr.html.

Le programme « Initiative B@bel » se propose d'y parvenir en exécutant, aux échelons national et international, des activités concrètes visant à développer le multilinguisme sur les réseaux d'information et à encourager un véritable partenariat entre les pouvoirs publics, les entreprises et la société civile. Il pourrait s'orienter dans plusieurs directions, à savoir :

- ⇨ la création d'une infrastructure : mise en place de chaires UNESCO, association des universités à l'industrie en vue d'intensifier la recherche-développement sur les moteurs de recherche multilingues, les passerelles multilingues, les bibliothèques et archives virtuelles, etc.;
- ⇨ la mise au point d'outils multilingues : adaptation des systèmes d'indexation multilingue des sites Web, thésaurus, normes, lexiques et outils terminologiques existant dans l'Union européenne, à l'UNESCO, à l'ISO, à l'UNU, à l'Union latine, à Infoterm, etc. dans d'autres langues, y compris les langues locales ;
- ⇨ le renforcement de l'interopérabilité : appui à la mise au point d'outils de traduction automatique, notamment à la production de logiciels de traductique gratuits, à l'application des travaux des écoles de traduction aux pages Web, à l'élaboration en ligne d'encyclopédies multilingues, au perfectionnement des routeurs, etc. ;
- ⇨ l'élaboration de politiques et règlements nationaux et internationaux : promotion de l'emploi de nombreuses langues sur les réseaux d'information, de l'enseignement en ligne des langues étrangères dans le cadre des systèmes éducatifs, de l'élaboration de sites Web multilingues (récompensés par un prix du site Web, etc.).

Plus concrètement, un document qui nous a été transmis par Paul Hector — « *assistant programme specialist* » à la division Société de l'Information à l'UNESCO et travaillant sur le projet *Initiative B@bel* — fait état de plusieurs projets en cours :

- ⇨ standardisation de l'éthiopien (codage, méthode de saisie, translittération),
- ⇨ bibliothèque audio pour des langues en danger du Caucase (abkhazien, bats et laz),
- ⇨ dictionnaire (abkhazien-géorgien),
- ⇨ plate-forme de gestion de contenu multilingue,
- ⇨ outil de développement et d'affichage multilingue pour Internet,
- ⇨ implémentation de modules de traduction automatique sur le site *Initiative B@bel*.

Plusieurs autres projets y sont mentionnés pour 2004-2005, en particulier :

- ⇨ la réalisation d'un CD-Rom d'enseignement du multilinguisme pour les 12-17 ans,
- ⇨ le développement d'orthographes pour les langues non écrites,
- ⇨ la poursuite des travaux précédents.

I.1.3.1.2.2 Le PNUD et ses partenaires

Parmi les autres organismes de l'ONU, le PNUD (projet des Nations Unies pour le développement[1]), a réalisé des actions dans le domaine de l'informatisation des langues peu dotées. Voici comment Mark Malloch Brown, administrateur du PNUD, le décrit.

[1] http://www.undp.org/french/.

En tant que réseau de développement des Nations Unies, le PNUD aide les pays en développement à élaborer leurs propres solutions aux problèmes nationaux et mondiaux au moyen de programmes et de services novateurs. Nous intervenons dans le monde entier pour mettre en rapport les pays donateurs et récipiendaires, le secteur public et le secteur privé, les conseils de politique et les ressources de programmes. Nos travaux sont de plus en plus fortement axés sur la coopération Sud-Sud, grâce à laquelle les pays en développement établissent entre eux des relations de partenariat. Par des dialogues, des échanges et des réseaux informatiques, nous aidons les autorités gouvernementales et les organisations à partager leur expertise, à nouer des liens et à susciter des opportunités.

Cet organisme dédié au développement suit la déclaration du Millénaire des Nations Unies et a, en particulier, pour mission « d'encourager la diversité linguistique et culturelle ainsi que le contenu local du cyberespace, en puisant dans les savoir-faire locaux et les connaissances indigènes » (voir le chapitre I.1.2.4.2). Grâce à ses bureaux dans plus de cent pays, il dispose de l'avantage unique d'être en contact direct avec les populations et de pouvoir ainsi donner des réponses mieux adaptées à leurs problèmes.

Le PNUD a diffusé en juillet 2001 le rapport final de la *Digital Opportunity Initiative*, un projet destiné à « évaluer au niveau mondial l'étendue du fossé en matière d'information et de connaissances et ses implications, de proposer un plan d'action pratique pour éliminer ce fossé dans les dix prochaines années, et d'élaborer de nouveaux projets novateurs exemplaires qui constituent des progrès décisifs dans le monde entier ». Bien qu'intéressant, ce rapport se borne, dans le domaine de l'informatisation des langues, à conseiller le recours à des partenaires externes :

« **Language Compatibility**. In many developing countries, problems also arise because standard fonts for local languages are unavailable. External partners (public, private and citizens in diaspora) can play a key role in this area. »

Le PNUD peut réaliser des projets directement mais fait aussi appel à des partenaires. Parmi ceux-ci, citons le GKP[1] (Global Knowledge Partnership), l'IDRC à travers le programme de financement du Pan Asia Networking[2] (International Development Research Center), ainsi que l'APDIP[3] (Asia-Pacific Development Information Programme), lui-même une initiative régionale des Nations Unies. Citons par exemple le projet de réalisation de polices de caractères ourdous[4] réalisé par la *National University of Computer and Emerging Sciences* à Lahore au Pakistan et financé conjointement par l'IDRC, l'APDIP, l'UNDP et l'APNIC.

L'ONU, à travers l'UNESCO et le PNUD, s'intéresse depuis 2001 aux logiciels libres et *Open Source* (FOSS, Free and Open Source Software). Outre le portail « *Free Software* »[5] que ces deux organismes leur consacrent, des projets existent déjà, comme l'initiative « *International Open Source Network* »[6] qui encourage les politiques dans ce domaine au niveau de la zone Asie du Sud-Est.

[1] http://www.globalknowledge.org/.
[2] http://www.panasia.org.sg/.
[3] http://www.apdip.net.
[4] http://www.apdip.net/ictrnd/nafees.asp.
[5] http://www.unesco.org/webworld/portal_freesoft/index.shtml.
[6] http://www.iosn.net/.

I.1.3.1.3 Actions de l'Union Européenne

Le traité instituant la Communauté Européenne[1] reconnaît douze langues : allemand, anglais, danois, espagnol, français, finnois, grec, italien, irlandais, néerlandais, portugais et suédois qui sont aussi langues officielles et de travail à l'exception de l'irlandais, en vertu du règlement n° 1 du conseil de la CEE adopté en avril 1958 et amendé lors des élargissements successifs de l'Europe[2]. **L'Union Européenne est ainsi l'organisation internationale la plus consommatrice de traduction**. Normand Labrie ([Labrie 1999]) considère que, chaque année, la seule Commission Européenne utilise plus de cent mille hommes-jours d'interprète et traduit plus d'un million de pages. Des projets européens ont été lancés pour résoudre les problèmes de plurilinguisme de l'Europe. En particulier, de 1982 et 1993, neuf États membres de la Communauté Européenne ont coopéré au **projet européen Eurotra**. Il s'agit d'un très ambitieux programme ayant pour buts :

- ⇨ la réalisation d'un prototype de système de traduction automatique pour les neuf langues officielles de la CEE de cette époque (allemand, anglais, danois, espagnol, français, grec, italien, néerlandais, et portugais),
- ⇨ la stimulation de la recherche en traitement des langues dans les États membres.

Puis, de 1994 à 1996, le **projet MULTEXT**[3] a vu la réalisation d'outils d'annotation de corpus pour six langues majeures de l'Union Européenne : l'allemand, l'anglais, l'espagnol, le français, l'italien et le néerlandais. Il a fait l'objet de plusieurs suites qui ont étendu son champ à plusieurs langues de l'Union : catalan, occitan et suédois (projets MULTEXT-CATALOC et MULTEXT-SW), de l'Europe Centrale et Orientale : bulgare, estonien, hongrois, roumain, slovène et tchèque (projet MULTEXT-EAST) mais aussi hors d'Europe : bambara, kikongo, et swahili (projet ALAF[4]).

En ce qui concerne les langues minoritaires, l'intérêt de l'Europe apparaît fin 1979 avec un projet de résolution présenté au Parlement Européen par John Hume et qui réclamait des droits pour les langues et les cultures régionales. Le rapport demandé à Gaetano Arfé sur ce point fut remis au Parlement pour sa session plénière d'octobre 1981 et la **résolution Arfé** fut adoptée par le Parlement. La **résolution Kuijpers** (du nom de son auteur, Willy Kuijpers), plus précise que la précédente sur les actions à entreprendre, fut adoptée par le Parlement Européen en 1987. Suite à chacune de ces résolutions, un organisme, financé en partie par la Commission Européenne, fut créé :

- ⇨ en 1982, le **BELMR**, un bureau européen pour les langues les moins répandues[5],
- ⇨ en 1989, le **réseau Mercator**[6], un réseau de recherche et d'information sur les langues régionales et minoritaires de l'Union Européenne.

Au début des années 90, le projet de **charte européenne des langues régionales ou minoritaires** confirme cette orientation en faveur des minorités linguistiques. À titre d'exemple, le financement du BELMR par l'Union Européenne est passé de cent mille à quatre millions d'euros entre 1982 et 1997. En 1992, l'Union Européenne lance un appel d'offres destiné à discerner le potentiel des groupes linguistiques minoritaires et qui a donné naissance au **projet Euromosaic**[7]. Ce projet réalisa un rapport général intitulé *Euromosaic : Production et reproduction des groupes linguistiques minoritaires au sein de l'Union Européenne*, qui fut publié en 1996 par la Commission.

Euromosaic a été prolongé début 2001 — année européenne des langues — sous la forme du projet Atlantis[8] (*Academic Training, Languages, And New Technologies in the Information Society*) qui réalisa un rapport des ressources et technologies disponibles pour les différentes langues minoritaires

[1] Dans sa version amendée par le traité d'Amsterdam (1997).
Voir http://europa.eu.int/eur-lex/fr/treaties/dat/amsterdam.html#0173010078.
[2] http://www.france.diplomatie.fr/europe/fran_euro/fra07.html.
[3] http://www.lpl.univ-aix.fr/projects/multext/.
[4] ALAF est une action de recherche partagée de l'AUPELF-UREF.
[5] http://www.eblul.org/.
[6] http://www.mercator-central.org/index-fr.htm.
[7] http://www.uoc.edu/euromosaic/web/homefr/index1.html.
[8] http://www.uoc.edu/in3/atlantis/fr/, rapport final : http://www.uoc.edu/in3/atlantis/eng/final_report.html.

de l'Union[1]. Cependant, les projets de traitement des langues-π européennes sont surtout limités à des projets de localisation comme le projet DART de localisation de langues celtiques[2] qui inclut une version « localisée » du navigateur Opera ainsi qu'une base terminologique en informatique pour les langues bretonne, irlandaise, gaélique d'Écosse et galloise.

En effet, après les importants projets d'informatisation des années 80 et du début des années 90, clairement tournés vers les problèmes posés par le plurilinguisme européen, les projets ont pris une orientation plus politique et internationale. Par exemple, parmi les récents appels d'offres de l'Union Européenne, voici les projets proposés par l'office de coopération EuropeAid, office mettant en œuvre l'aide extérieure :

- ⇨ Asia IT&C (http://europa.eu.int/comm/europeaid/projects/asia-itc/html/main.htm) qui finance des partenariats Europe-Asie dans le domaine des technologies de l'information,
- ⇨ Asia-Invest (http://europa.eu.int/comm/europeaid/projects/asia-invest/html2002/main.htm) qui finance des coopérations Europe-Asie dans le domaine industriel,
- ⇨ Asia-Link (http://europa.eu.int/comm/europeaid/projects/asia-link/index_en.htm) qui finance le développement et le renforcement des réseaux universitaires entre l'UE et l'Asie du Sud, l'Asie du Sud-Est et la Chine,
- ⇨ AUNP (http://europa.eu.int/comm/europeaid/projects/aunp/index_en.htm) qui finance le renforcement des coopérations universitaires entre l'UE et l'ASEAN.

Cependant, ces projets européens ne favorisent pas particulièrement les travaux dans le domaine du traitement des langues[3]. Ils sont par ailleurs lourds à gérer au niveau de centres de recherche, tant lors de la phase d'appel d'offres que dans celui de réalisation. Ces lourdeurs les rendent, de surcroît, mal adaptés à des partenaires de pays du Sud. Citons néanmoins le thème « technologies de la société de l'information » du sixième PCRD (programme cadre de recherche et développement) de l'Union Européenne[4], qui prévoit que des travaux en reconnaissance de la parole et en interfaces homme-machine évoluées bénéficient d'une part de son budget qui est globalement de 3 600 millions d'euros pour la période 2002-2006.

Une autre action intéressante est celle du forum **ELSNET**[5] (*European Network in Language and Speech*). Cet organisme créé en 1991 est l'un des réseaux d'excellence[6] établis dans le cadre du programme ESPRIT — programme européen de recherche sur les technologies de l'information et des télécommunications — et couvrant ensemble ce domaine. À ce titre, ELSNET est financé par la Commission Européenne. Son objectif principal est de rapprocher les acteurs des technologies du langage. Entre autres services, il diffuse un grand nombre d'informations[7] et de ressources[8], publie la revue ELSNews, propose une liste de diffusion et recense les thèses sur le traitement des langues. Avec **Euromap**[9], organisme créé en 1996 avec pour mission d'aider l'industrialisation de la recherche faite en technologie des langues, ELSNET est l'un des principaux soutiens financiers du projet HLTCentral[10] consacré à la diffusion d'informations sur Internet dans ce domaine. Notons encore l'existence d'outils d'aide à la traduction[11] et en particulier Eurodicautom, un dictionnaire terminologique proposé sur le site de l'Union Européenne et fournissant des traductions de termes techniques dans les onze langues officielles de l'Union[12].

---

[1] http://www.uoc.edu/in3/atlantis/eng/final_report.html.
[2] http://ww2.eblul.org:8080/eblul/Public/projets_en_cours/projets_precedents/dart/view.
[3] À notre connaissance, au moins trois projets proposés dans ce domaine — CurricInfo, MIAM et PapilloNet — ont été refusés lors des appels d'offres d'octobre 2002, de mai 2003 et de juin 2003.
[4] Voir le site CORDIS http://www.cordis.lu/fp6 (COmmunity Research & Development Information Service).
[5] http://www.elsnet.org/.
[6] http://www.newcastle.research.ec.org/, http://www.newcastle.research.ec.org/esp-syn/ltr-ne-index.html.
[7] Incluant une liste de produits : http://www.elsnet.org/productslist.html.
[8] http://www.elsnet.org/resources.html.
[9] http://www.hltcentral.org/htmlengine.shtml?id=56.
[10] http://www.hltcentral.org/. Voir aussi http://www.unesco.org/webworld/portal_observatory/Access_-_Applications/Multilingualism/Standards/European_Postions/index.shtml.
[11] http://europa.eu.int/comm/translation/reading/articles/tools_and_workflow_en.htm.
[12] http://europa.eu.int/eurodicautom/login.jsp.

I.1.3.1.4 Financements d'États

I.1.3.1.4.1 France

Le rapport[1] adressé en avril 1999 par Bernard Cerquiglini, alors directeur de l'INaLF, aux ministres de l'Education, Claude Allègre, et de la Culture, Catherine Trautmann, dénombre soixante-dix-sept « langues parlées par des ressortissants français sur le territoire de la République »[2] :

- ⇨ vingt-cinq en métropole : dialecte allemand d'Alsace et de Moselle, basque, breton, catalan, corse, flamand occidental, francoprovençal, occitan (gascon, languedocien, provençal, auvergnat-limousin, alpin-dauphinois), franc-comtois, wallon, picard, normand, gallo, poitevin-saintongeais, bourguignon-morvandiau, lorrain, berbère, arabe dialectal, yiddish, romani chib, arménien occidental,
- ⇨ treize dans les DOM : créoles à base lexicale française (martiniquais, guadeloupéen, guyanais, réunionnais), créoles bushinenge à base lexicale anglo-portugaise de Guyane (saramaca, aluku ou njuka ou paramaca), langues amérindiennes de Guyane (galibi ou kalina, wayana, palikur, arawak proprement dit ou lokono, wayampi, émerillon), hmong,
- ⇨ trente-neuf dans les TOM : nyelâyu, kumak, caac, yuaga, jawe, nemi, fwâi, pije, pwaamei, pwapwâ, dialectes de la région de Voh-Koné, cèmuhî, paicî, ajië, arhâ, arhö, ôrôwe, neku, sîchë, tîrî, xârâcùù, xârâgùrè, drubéa, numèè, nengone, drehu, iaai, fagauvea, tahitien, marquisien, langue des Tuamotu, langue mangarévienne, langue de Ruturu (Iles Australes), langue de Ra'ivavae (Iles Australes), langue de Rapa (Iles Australes), walissien, futunien, shimaoré, shibushi.

Ce rapport avait été demandé en décembre 1998 pour préparer la ratification de la charte européenne des langues régionales ou minoritaires. Bien que signée par le gouvernement le 7 mai 1999, la charte n'a pas été ratifiée, le Conseil Constitutionnel ayant déclaré celle-ci contraire à la Constitution, le « droit imprescriptible à pratiquer une langue régionale ou minoritaire, non seulement dans la vie privée mais également dans la vie publique » du préambule de la charte étant contraire à l'article 2 de la Constitution[3] (« Le français est la langue de la République. »).

De fait, nous n'avons pas trouvé trace de projets d'informatisation d'une des langues minoritaires parlées en France qui soit une initiative étatique, la tradition française favorisant la **francophonie**. Mais, dans ce cadre de la promotion du français, qui est langue officielle dans trente-sept États et l'une des principales langues internationales, la France a été à l'origine de projets d'informatisation de langues-π ou μ en relation avec le français, par le biais de programmes (ACCT, AUPELF-UREF/AUF...) dans lesquels il s'agit, par exemple, de construire des ressources lexicales (projets de dictionnaires français-malais, français-arabe, wolof-français-arabe ou encore le projet ALAF, suite du projet MULTEXT citée précédemment).

[1] http://www.culture.gouv.fr/culture/dglf/lang-reg/rapport_cerquiglini/langues-france.html.

[2] Pour la France, Ethnologue dénombre, quant à lui, vingt-huit langues (noms donnés tels que sur le site Ethnologue) :
alemannisch, auvergnat, navarro-labourdin (basque), souletin (basque), breton, caló, catalan-valencian-balear, corse, dutch, esperanto, franco-provençal, french, gascon, greek, interlingua de iala, italian, languedocien, ligurian, limousin, luxembourgeois, picard, portuguese, provençal, balkan (romani), sinte (romani), vlax (romani), castillian, flamand,
et indique la présence de trente-sept autres :
Adyghe, Algerian Spoken Arabic, Judeo-Moroccan Arabic, Judeo-Tunisian Arabic, Moroccan Spoken Arabic, Tunisian Spoken Arabic, Armenian, Assyrian Neo-Aramaic, Brao, Western Cham, Chru, Western Farsi, Standard German, Hmong Daw, Iu Mien, Kabuverdianu, Kabyle, Central Khmer, Khmu, Kirmanjki, Kurmanji, Laz, Lesser Antillean créole French, Mandjak, Nhang, Tachelhit, Tai Dam, Tai Don, Tai Nüa, Central Atlas Tamazight, Tarifit, Tay, Turkish, Vietnamese, Wolof, Yeniche, Western Yiddish.

[3] Cet alinéa a été ajouté à la Constitution par la loi constitutionnelle 92-554 du 25 juin 1992.

I.1.3.1.4.2 Canada

L'assemblée législative du Nunavut, région du Canada nouvellement promue au rang de territoire (le 1er avril 1999), a permis la réalisation de polices de caractères Unicode pour l'Inuktitut, langue des Inuits, ainsi qu'un clavier virtuel permettant de les utiliser[1].

I.1.3.1.4.3 Japon

Le centre de la coopération internationale pour l'informatisation (Center of the International Cooperation for Computerization ou CICC, http://www.cicc.or.jp/english/index.html) est « une organisation à but non lucratif destinée à aider les pays en développement à réaliser leur informatisation ». Ce centre a par exemple réalisé de 1987 à 1993 le projet de traduction automatique « Research Cooperation on Machine Translation System Among Neighboring Countries » avec la Chine, l'Indonésie, la Malaisie et la Thaïlande. Les logiciels de traduction automatique au Japon étant réalisés par des entreprises privées, NEC fut associé à la Thaïlande, Hitachi à l'Indonésie, Fujitsu à la Malaisie et les trois firmes à la République Populaire de Chine. Les résultats de ces travaux ont été centralisés au Japon. Seule la Thaïlande, grâce à la présence de doctorants thaïs chez NEC au Japon[2], a pu développer un système de traduction automatique qui a survécu au projet CICC (Prasit, fondé sur le système PIVOT / Crossroad de NEC).

I.1.3.2 LES PRODUCTEURS DE LOGICIELS

I.1.3.2.1 Grands éditeurs de logiciel et industrie des langues

I.1.3.2.1.1 Éditeurs de systèmes d'exploitation et d'outils de base

Citons d'abord l'oligopole des **éditeurs de systèmes d'exploitation** du secteur privé (Microsoft, Apple, SUN...) ou du domaine des logiciels ouverts (GNU/Linux). Ils fournissent en particulier les systèmes d'exploitation, les navigateurs et suites bureautiques, outils nativement multilingues, tant au niveau saisie et affichage – ces outils ont bien sûr accompagné l'évolution vers Unicode – qu'à des niveaux plus linguistiques. En particulier, Windows (depuis Windows NT 3.1), Macintosh (depuis MacOS 8.5) et Unix/Linux intègrent Unicode et de nombreuses polices de caractères sont disponibles, rendant *a priori* possible l'existence de documents réellement multilingues[3].

Ces acteurs commencent à s'intéresser aux langues-μ, voire aux langues-π, au moins à celles que Louis-Jean Calvet appelle les langues « centrales »[4]. Les fenêtres d'édition disponibles permettent aujourd'hui de réaliser une saisie adaptée à des écritures diverses (voir I.1.4.2). Cependant, si l'on y regarde de plus près, une suite bureautique aussi répandue que Microsoft Office XP ne dispose d'outils linguistiques que pour 48 langues. L'exemple n'est pas un cas particulier et l'on peut dire que moins de 1% des langues sont ainsi correctement informatisées par les produits de ces grands du secteur informatique[5].

Les systèmes d'exploitation fournissent ainsi le support sur lequel l'industrie des langues et les développeurs n'appartenant pas au secteur privé peuvent développer leurs logiciels.

[1] http://www.assembly.nu.ca/unicode/fonts/.
[2] Virach Sornlertlamvanich en particulier (au NECTEC, http://www.links.nectec.or.th/virach/).
[3] Voir le chapitre I.1.4.2
[4] Voir le chapitre I.1.2.3.
[5] Les langues bien informatisées sont généralement celles de communautés à fort potentiel économique, mais les critères utilisés pour décider d'informatiser une langue sont plus complexes. Ils incluent plus généralement la notion de retour sur investissement, et une langue « facile » à informatiser aura donc plus de chance.

I.1.3.2.1.2 Industrie des langues

On trouve souvent, dans la littérature traitant d'informatisation des langues, les « numéronymes » G11n, (*Globalization*) I18n (*Internationalization*) et L10n (*Localization*)[1]. Précisons ces termes.

- ⇨ La **globalisation**, G11n, est l'ensemble des moyens permettant à une entreprise d'être adaptée à un marché mondial. Cela inclut l'adaptation de son marketing, de ses forces de vente et de son service après-vente.
- ⇨ L'**internationalisation**, I18n, est le processus permettant à un produit de s'adapter à la langue et aux conventions culturelles de sa cible, sans nécessiter de conception supplémentaire.
- ⇨ La **localisation**, L10n, est l'adaptation linguistique et culturelle d'un produit à une cible à laquelle il sera vendu et par laquelle il devra être utilisé. La localisation s'oppose donc à l'internationalisation en ce qu'elle est une modification d'un produit alors que la technique de l'internationalisation est de faire un produit le plus neutre possible dès sa conception, limitant le temps nécessaire à la localisation.

Ces termes recouvrent un pan important d'activité du secteur privé appelé l'industrie des langues. Cependant, mises à part de rares entreprises qui ciblent quelques niches en croissance dans ce domaine, l'industrie des langues contribue peu à l'informatisation des langues-π. Citons la société américaine MIT2[2] qui a réalisé *créoleConvert* et *créoleScan*, des outils de standardisation orthographique et de reconnaissance optique de caractères pour le créole haïtien [Mason 2000]. Sa présidente, Marilyn Mason, défend l'idée que les langues-π ont un réel potentiel commercial à condition de viser des secteurs limités qu'il faut identifier au cas par cas pour chaque langue [Mason 2002]. Notons l'existence de normes[3] dans le domaine de la spécification de conventions culturelles (ISO/CEI 14652), du tri (ISO/IEC 14651) et de l'internationalisation (ISO/CEI 15435).

I.1.3.2.2 Développeurs de logiciels Open Source ou libres

I.1.3.2.2.1 Projet GNU et philosophie des logiciels libres

La philosophie des logiciels libres remonte au moins au projet GNU. Ce projet, lancé par Richard Stallman en 1984[4], a contribué au succès de Linux grâce aux logiciels GNU venus compléter le noyau créé par Linus Torvalds (bibliothèque glibc, compilateur gcc, Emacs...). L'idée de base est que les sources des logiciels, souvent gratuits[5], sont distribués en même temps que les exécutables, permettant ainsi à toute personne, en respectant les termes d'une licence — par ecemple la *GNU Public License* (GPL) — d'améliorer le logiciel. Une condition généralement demandée est que toute évolution notable soit mutualisée, ce qui confère à ces licences un comportement qualifié par certains de viral[6]. La licence GPL est donnée en annexe A.13 ainsi que sur le site http://opensource.org/licenses/ qui recense au total plus de 50 modèles de licences[7].

---

[1] Le mot *Localization* est constitué de la lettre *L* puis des dix lettres *ocalizatio* puis de la lettre *n*. Notons que l'on trouve encore d'autres tels numéronymes, en particulier T9n (*Translation*) qui forme avec les trois autres le sigle GILT parfois utilisé dans l'industrie des langues. Voir par exemple les sites http://www.i18nguy.com/ et http://www.lisa.org/info/faqs.html.

[2] http://hometown.aol.com/mit2usa/Index2.html.

[3] http://anubis.dkuug.dk/jtc1/sc22/wg20/docs/projects.html.fr.

[4] Le Gnu's Not Unix (GNU) est un projet de développement d'un système d'exploitation de type Unix en logiciel libre (http://www.gnu.org/).

[5] Bien que Linux soit un système d'exploitation gratuit, sa distribution est généralement payante du fait du service après vente, de la documentation papier et des outils auxiliaires qu'elle inclut.

[6] L'article d'Éric Raymond « The Cathedral and the Bazaar » démontre avec enthousiasme l'efficacité du travail coopératif dans le domaine du logiciel (http://www.tuxedo.org/~esr/writings/cathedral-bazaar/, version française sur http://www.linux-france.org/article/these/cathedrale-bazar/cathedrale-bazar.html).

[7] Le mémoire de DEA d'Éric Di Filippo ([Di Filippo 1999]) développe les aspects juridiques du logiciel libre.

Devant le succès de GNU/Linux, de nombreux groupes de développement logiciel utilisant ces principes ont vu le jour. Parmi les plus connus, le World Wide Web Consortium (W3C), projet de développement de technologies web, développe des logiciels ouverts[1] et libres (http://www.w3.org/). Nous donnons ci-dessous quelques exemples de projets de type GNU, certains visant l'informatisation de langues-π. Ce sont, en particulier, les projets de localisation de Linux et d'OpenOffice.org (OOo)[2].

I.1.3.2.2.2 Quelques projets de type GNU

Le site SourceForge.net (http://sourceforge.net/) recense un très grand nombre de projets Open Source. En septembre 2003, il en héberge 45 607 dans 52 langues différentes[3]. Parmi ces projets, plus de trente-cinq mille sont en anglais, puis onze langues ont plus de cent projets totalisant environ dix mille projets (allemand, brésilien, chinois simplifié, espagnol, français, hollandais, italien, japonais, polonais, russe et suédois). Puis viennent quarante langues qui ont moins de cent projets faisant environ mille projets au total (afrikaans, arabe, bengali, bosniaque, bulgare, catalan, chinois traditionnel, coréen, croate, danois, espéranto, finnois, grec, hébreu, hindi, hongrois, islandais, indonésien, javanais, latin, letton, macédonien, malais, marathi, norvégien, penjabi, persan, portugais, roumain, serbe, slovaque, slovène, tamoul, tchèque, telugu, thaï, turc, ukrainien, urdu, vietnamien). Ces statistiques donnent une image du dynamisme des communautés linguistiques. Il est remarquable que moins du quart des projets soit dans une autre langue que l'anglais et que n'apparaissent que des langues officielles. D'autant que ces projets ne sont pas, pour la plupart, des projets d'informatisation des langues mais simplement des projets dont l'interface utilisateur est dans une langue particulière.

Parmi les projets d'informatisation des langues, nous trouvons le plus souvent des projets de localisation. Par exemple, plusieurs projets de localisation de Linux existent :

- ⇨ dans plusieurs langues de l'Inde et du Bangladesh (projet IndLinux, http://indlinux.org/),
- ⇨ dans plusieurs langues indiennes (http://dot.kde.org/1053201681/),
- ⇨ en tamoul (http://www.tamillinux.org/),
- ⇨ en bengali (projet Ankur http://www.bengalinux.org/),
- ⇨ en laotien (projet LaoNux, http://laonux.muanglao.com/),
- ⇨ KDE dans plusieurs langues (http://i18n.kde.org/teams/index.php),

des projets de développement d'applications telles que Mozilla (http://www.mozilla.org/),
ainsi que des projets de localisation d'OpenOffice.org, par exemple :

- ⇨ OpenOffice.org hindi (http://hi.openoffice.org/),
- ⇨ OpenOffice.org thaï (http://www.pladao.org, http://th.openoffice.org/).

IBM s'est fortement engagé vers l'Open Source et propose des bibliothèques de services pour gérer Unicode sur plusieurs systèmes d'exploitation différents. Ce projet s'appelle *International Components for Unicode* ou *ICU* (http://oss.software.ibm.com/icu/). Il propose aussi une liste des paramètres de localisation nationaux (http://oss.software.ibm.com/cgi-bin/icu/lx/en_US/utf-8/).

---

[1] L'Open Source Initiative (OSI) est un groupement à but non lucratif dont le but est de promouvoir le logiciel dit ouvert (http://www.opensource.org/).

[2] OpenOffice.org (http://www.openoffice.org/) est un projet de suite bureautique concurrente de Microsoft Office. Il provient de la mise dans le domaine ouvert par Sun Microsystems du produit StarOffice, après le rachat de ce logiciel à la société StarDivision.

[3] Ces langues sont : afrikaans (11 projets), allemand (3453 projets), anglais (35081 projets), arabe (31 projets), bengali (3 projets), bosniaque (2 projets), bulgare (32 projets), catalan (35 projets), chinois (simplifié) (133 projets), chinois (traditionnel) (85 projets), coréen (36 projets), croate (10 projets), danois (43 projets), espagnol (1352 projets), espéranto (20 projets), finnois (34 projets), français (2127 projets), grec (27 projets), hébreu (47 projets), hindi (12 projets), hollandais (254 projets), hongrois (60 projets), islandais (2 projets), indonésien (19 projets), italien (373 projets), japonais (322 projets), javanais (1 projets), latin (8 projets), letton (10 projets), macédonien (1 projets), malais (7 projets), marathi (2 projets), norvégien (31 projets), penjabi (3 projets), persan (14 projets), polonais (150 projets), portugais (60 projets), portugais (brésilien) (245 projets), roumain (41 projets), russe (638 projets), serbe (7 projets), slovaque (27 projets), slovène (7 projets), suédois (114 projets), tamoul (19 projets), tchèque (78 projets), telugu (6 projets), thaï (19 projets), turc (57 projets), ukrainien (23 projets), urdu (7 projets), vietnamien (22 projets)

I.1.3.2.3 SIL International

SIL International (nouveau nom du *Summer Institute of Linguistics*), qui œuvre dans le domaine de la description des langues depuis 1934, a su s'adapter au processus d'informatisation. Il a développé divers logiciels et polices de caractères qu'il diffuse via Internet, environ quatre cents polices (incluant quelques écritures absentes d'Unicode) et soixante logiciels[1] (http://www.sil.org/computing/).

SIL International poursuit actuellement neuf projets :

- ⇨ **Non-Roman Script Initiative** : groupe travaillant à faciliter l'utilisation d'écritures non-latines et complexes dans les études de linguistique, de traduction, de littérature et de publication,
- ⇨ **TECKit** : une boîte à outils pour la conversion en Unicode, soit dans une application ou directement dans des fichiers texte ou SFM,
- ⇨ **XSEM** : modèle d'encodage XML pour l'écriture,
- ⇨ **Graphite** : un moteur d'affichage pour des systèmes d'écriture non-latins complexes,
- ⇨ **Fieldworks** : les fondements de la prochaine génération de logiciels pour les langues,
- ⇨ **LinguaLinks**, un système d'aide à la productivité informatique pour linguistes de terrain,
- ⇨ **CELLAR** : Computing Environment for Literary, Linguistic, and Anthropological Research,
- ⇨ **Hermit Crab** : un analyseur et un générateur morphologiques pour de la morphologie et de la phonologie classique,
- ⇨ **Speech Analysis Tools** : des outils d'enregistrement, transcription et analyse de fichiers son.

En janvier 2003, l'UNESCO et le SIL ont conclu un accord incluant sept objectifs dont l'utilisation de Graphite, le système du SIL concurrent de la technologie OpenType[2]. Le SIL devrait développer des outils pour la saisie et l'affichage des écritures complexes et adapter un navigateur pour lui permettre l'affichage de ces écritures.

I.1.3.2.4 SALTMIL

L'ISCA — *International Speech Communication Association* — est une association à but non lucratif créée à Grenoble en 1988 par René Carré[3]. Elle a fondé, en 1999, SALTMIL[4] (*Special Interest Group on Speech and Language Technology for Minority Languages*), un groupe d'intérêt spécial (SIG) pour les langues qu'elle nomme « minoritaires ». Ce groupe a pour but de promouvoir la recherche et le développement dans le traitement des langues les moins utilisées. Il est présent sous forme d'ateliers aux conférences LREC (*Language Resources and Evaluation Conference*) depuis la première de ces conférences en 1998 où fut proposée sa création :

- ⇨ 1998 : http://www.lrec-conf.org/lrec98/ceres.ugr.es/_rubio/elra/minority.html (Grenade Espagne)
- ⇨ 2000 : http://www.lrec-conf.org/lrec2000/www.cstr.ed.ac.uk/SALTMIL/lrec00.html, (Athènes Grèce)
- ⇨ 2002 : http://www.lrec-conf.org/lrec2002/lrec/wksh/WP15agendaF.html. (Las Palmas, îles Canaries, Espagne)

[1] Incluant Shoebox, une boîte à outils intégrée de gestion et d'analyse de données pour les linguistes sur le terrain. Sites web :
http://www.sil.org/computing/shoebox/,
http://llacan.cnrs-bellevue.fr/FRANCAIS/EnLigne/Shoebox/ShoeboxCadre.htm.

[2] http://www.sil.org/sil/news/2003/unesco.htm, http://www.sil.org/sil/news/2003/unesco2.htm.

[3] À l'origine, cette association s'appelait ESCA (*European Speech Communication Association*). Elle a changé de nom en 1999 lorsque son champ d'investigation s'est globalisé.

[4] http://isl.ntf.uni-lj.si/SALTMIL/.

I.1.3.2.5 Recherche universitaire

La recherche universitaire est très présente dans le traitement des langues. Citons le *Language Technology Institute*[1] de l'université Carnegie-Mellon à Pittsburg. Créé en 1996 à partir du *Center for Machine Translation*, lui-même créé en 1986, il continue de travailler dans le domaine de la traduction automatique mais étudie aussi le traitement automatique des langues naturelles, le traitement de la parole et la recherche d'information. Avec, fin 2003, cinquante-cinq doctorants et trente-deux étudiants en master, cet institut a la taille critique nécessaire pour réaliser ses nombreux projets ambitieux[2] (dont sept en traduction automatique). Cette taille critique et un financement de la DARPA ont permis la création du projet AVENUE/NICE[3] dans le cadre duquel des services de traduction automatique ont été développés pour trois langues-π, le mapudungun (Chili, 400 000 locuteurs), l'inupiaq (Alaska, 3 500 locuteurs) et le siona (Colombie, 300 locuteurs).

Cette dimension n'est cependant pas indispensable pour réussir une informatisation. Le groupe IXA de l'université du Pays Basque a ainsi réalisé de nombreux outils pour la langue basque, en particulier une base de données lexicales et un correcteur d'orthographe[4].

I.1.3.2.6 Développeurs isolés

De nombreuses initiatives plus ou moins significatives proviennent d'individus ou de petits groupes développant des logiciels, souvent pour une seule langue.

Il est remarquable que ces logiciels, généralement réalisés avec les seuls moyens des développeurs, soient quand même de très bonne qualité et très bien adaptés aux besoins des populations avec lesquelles ils sont en intime interaction. En voici quelques exemples.

- ⇨ Site de dictionnaires téléchargeables (Stéphane Rousseau ou Beaumont) http://www.freelang.com/,
- ⇨ Outils pour la langue khmère (Maurice Bauhahn) http://www.bauhahnm.clara.net/Khmer/Welcome.html,
- ⇨ Outils pour « la » langue créole (Marilyn Mason, The Creole Clearinghouse) http://hometown.aol.com/creoleCH/,
- ⇨ Travail universitaire sur l'amharique http://www.sciences.univ-nantes.fr/irin/taln2003/articles/alemu.pdf.

I.1.3.3 SOURCES D'INFORMATION ET OUTILS TÉLÉCHARGEABLES

Les sources d'information concernant l'informatisation des langues sont essentiellement :
des **conférences** :

- ⇨ l'atelier sur les langues minoritaires des conférences LREC (tous les deux ans depuis 1998) : http://www.lrec-conf.org/fr/index.html, http://www.lrec-conf.org/lrec98/ceres.ugr.es/_rubio/elra/minority.html, http://www.lrec-conf.org/lrec2000/www.cstr.ed.ac.uk/SALTMIL/lrec00.html, http://www.lrec-conf.org/lrec2002/lrec/wksh/WP15agendaF.html,
- ⇨ l'atelier associé à TALN 2003 « Traitement automatique des langues minoritaires et des petites langues » : http://www.sciences.univ-nantes.fr/irin/taln2003/page/acte_sommaire.html#atelier,
- ⇨ des listes de conférences se trouvent aux adresses : http://conferences.atala.org/conferences/conft.html, http://www.elsnet.org/cgi-bin/elsnet/events.pl.

[1] http://www.lti.cs.cmu.edu/.
[2] http://www.lti.cs.cmu.edu/Research/cmt-projects.html.
[3] http://www-2.cs.cmu.edu/~aria/avenue/, http://passau.is.cs.cmu.edu/dict.jsp.
[4] Voir [Agirre et al. 2001] et http://ixa.si.ehu.es.

des **informations en ligne** :

- ⇨ Language Technology World
  http://www.lt-world.org
  LT-World offre un grand nombre d'informations sur les technologies utilisées en traitement des langues. Il est réalisé par le *National Language Technology Competence Center* de DFKI, le centre allemand de recherche pour l'intelligence artificielle (*Deutsche Forschungszentrum für Künstliche Intelligenz GmbH*).
- ⇨ ELSNET (voir le chapitre I.1.3.1.3) :
  http://www.elsnet.org
- ⇨ ACL (Association for Computational Linguistics) :
  http://www.aclweb.org/
- ⇨ l'Open Language Archives Community :
  http://www.language-archives.org/
  L'OLAC est une plate-forme coopérative destinée à « créer une bibliothèque virtuelle mondiale des ressources langagières » ; créé en décembre 2000, ce projet récent réunit déjà plus de vingt participants (les *data providers*) dont les ressources sont accessibles à travers un *service provider* tel que « *the Linguist* » (http://www.linguistlist.org/olac/).
- ⇨ la Localization Industry Standards Association :
  http://www.lisa.org/
  LISA est une organisation professionnelle des métiers du GILT (Globalisation, Internationalisation, Localisation, Traduction) ; elle œuvre en particulier dans le domaine de la normalisation et édite le journal bimensuel LISA Newsletter.

des **bulletins électroniques** et des **forums de discussions** :

- ⇨ LN : LN@cines.fr (http://www.biomath.jussieu.fr/LN/LN-F/),
- ⇨ LN-FR : LN-FR@cines.fr (http://www.biomath.jussieu.fr/LN-FR/LN-FR/),
- ⇨ ELSNET : elsnet-list@elsnet.org (http://www.elsnet.org/list.html),
- ⇨ Unicode : unicode@unicode.org (http://www.unicode.org/consortium/distlist.html),
- ⇨ Linux-UTF8 : linux-utf8@nl.linux.org (http://mail.nl.linux.org/linux-utf8/).

des **thèses** :

- ⇨ liste de thèses en traitement automatique des langues du site de l'ATALA :
  http://www.biomath.jussieu.fr/ATALA/these/,
- ⇨ liste de thèses en traitement automatique des langues d'ELSNET :
  http://hltheses.elsnet.org/,
- ⇨ liste de thèses du CNRS :
  http://tel.ccsd.cnrs.fr.

des **liens** :

- ⇨ le répertoire de liens d'ELSNET vers des sites proposant des outils et des ressources : http://www.elsnet.org/toolslist.html,
- ⇨ le répertoire de liens thématiques du site de l'Observatoire Suisse des Industries de la Langue : http://www.issco.unige.ch/projects/osil/liste.html,
- ⇨ le répertoire de liens du site I18n Guy (GILT) : http://www.i18nguy.com/TranslationTools.html,
- ⇨ liste de documents sur les langues peu enseignées (dictionnaires, grammaires…) : http://www.dunwoodypress.com/home.htm,
- ⇨ recherche d'informations dans une langue : http://omniglot.com/links/searchengines.htm,
- ⇨ catalogue de la librairie Grant & Cutler : http://www.grantandcutler.com,
- ⇨ liste de liens en traitement des langues (ressources et institutions) : http://www.ims.uni-stuttgart.de/info/FTPServer.html,
- ⇨ liste de liens en traitement des langues : http://stp.ling.uu.se/~fredriko/Ling/lingLinks.html,
- ⇨ liste de liens en traitement des langues (inclut un volet sur les langues d'Afrique) : http://www.bisharat.net/links.htm.

Plusieurs sites Internet fournissent des informations techniques ou générales liées au traitement des langues et proposent parfois des outils en téléchargement, en particulier :

- ⇨ le répertoire d'outils pour le traitement automatique des langues du site de l'ATALA : http://www.biomath.jussieu.fr/ATALA/outil/,
- ⇨ le répertoire d'outils téléchargeables du site de l'université de Genève (ISSCO) : http://issco-www.unige.ch/tools/,
- ⇨ le répertoire d'outils (payants) de LATL.ch, société anonyme créée par Eric Wehrli et Luka Nerima : http://www.latl.ch/french/index.htm,
- ⇨ le répertoire d'outils téléchargeables de SIL International : http://www.sil.org/computing/,
- ⇨ le guide sur les systèmes d'écriture de SIL International : http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi.

I.1.4 La technologie : une informatisation par service

I.1.4.1 UNICODE ET POLICES DE CARACTÈRES

Jusqu'au milieu des années 1990, une multitude de codages était utilisée pour représenter les différents systèmes d'écriture existants, par exemple les standards ISO/IEC 8859-x, sur un octet, pour les principales écritures alphabétiques — écritures dérivées de l'écriture latine et que nous appellerons latines (8859-1, -2, -3, -4, -9), cyrillique (8859-5), arabe (8859-6), grec (8859-7), hébreu (8859-8)… — ou encore des standards sur deux octets pour les écritures idéographiques d'Extrême-Orient — chinois simplifié (GB2312, Chine populaire), chinois traditionnel (Big5, Taiwan), japonais (Shift_JIS ou EUC-JP)… Pour pouvoir visualiser un texte, il était donc nécessaire de savoir à quel codage il était associé et les textes multilingues étaient nécessairement balisés, par exemple par des séquences d'échappement comme celles de la norme ISO 2022. Aux difficultés nées de cette hétérogénéité venait s'ajouter l'existence de plusieurs codages différents pour un même système d'écriture (absence de standardisation). D'autres facteurs non techniques comme la mondialisation de l'économie et des échanges ou le développement d'Internet ont fini de clarifier l'existence d'un besoin pour un standard universel dans le domaine de la représentation de l'information textuelle.

Pour résoudre ces difficultés a été fondé en 1991 le consortium Unicode[1] qui vise à représenter tous les systèmes d'écriture existants de façon unifiée. Il a pour cela choisi d'abandonner le codage sur 8 bits et de créer un code pour chaque caractère à représenter[2]. Unicode est un consortium auquel participent les grands éditeurs de logiciel — Apple Computer, Hewlett-Packard, IBM, Microsoft, Oracle, Silicon Graphics, Sun Microsystems, Sybase ou encore Xerox, pour ne citer que les plus connus — ce qui assure une mise en œuvre dans les logiciels de ce standard *a priori*. C'est ce qui s'est passé : la première version du standard Unicode a été publiée en 1991 et les systèmes d'exploitation l'ont intégrée petit à petit. Aujourd'hui, bien que les textes représentés avec les codages des années 1980 continuent d'être correctement interprétés, ils sont souvent transformés en Unicode par les logiciels sans que l'utilisateur en ait conscience. Unicode est d'ailleurs le codage par défaut d'XML, langage qui utilise l'encodage UTF-8.

Le standard Unicode 3.0 [The Unicode Consortium, 2000] inclut trente-six systèmes d'écriture[3,4] :

- ⇨ **Amériques** : cherokee, syllabaire aborigène canadien,
- ⇨ **Afrique** : éthiopien,
- ⇨ **Europe** : arménien, cyrillique (*), géorgien, grec, latin (*), ogham, runique,
- ⇨ **Moyen Orient** : arabe (*), hébreu, syriaque, thaana,
- ⇨ **Asie du Sud** : bengali, devanagari, gurmukhi, gujrati, kannada, malayalam, oriya, cinghalais, tamoul, télougou,
- ⇨ **Asie du Sud-Est et Haute-Asie** : birman, khmer, laotien, thaï, tibétain,
- ⇨ **Asie Orientale** : bopomofo, han (idéogrammes unifiés : chinois, japonais, coréen), hangul et hangul jamo (syllabaire coréen), hiragana et katakana (syllabaires japonais), mongol, yi.

(*) : Ces systèmes d'écriture contiennent des variantes permettant l'adaptation à des écritures légèrement différentes.

[1] Unicode existait sous forme de projet depuis 1988.

[2] La version 4.0 du standard contient 1 114 112 codes.

[3] Les systèmes d'écriture pris en compte par Unicode sont présentés sur le site du consortium Unicode : http://www.unicode.org.

[4] La répartition dans le standard est la suivante. Les codes de **0000 à 1FFF** sont attribués aux **écritures** appelées **générales** dans le standard : latines, grecque, cyrillique, arménienne, hébreu, arabe, syriaque, thaana, devanagari, bengali, gurmukhi, gujrati, oriya, tamoule, télougou, kannada, malayalam, cinghalaise, thaïe, laotienne, tibétaine, birmane, géorgienne, hangul jamo, éthiopienne, cherokee, syllabaires aborigènes canadiens, ogham, runique, khmère, mongole. De **2000 à 27FF** se trouvent divers **symboles** (ponctuation, opérateurs mathématiques, symboles typographiques, flèches…) puis le **braille** de **2800 à 28FF**. Les **écritures d'Asie orientale** viennent ensuite de **2E80 à D7AF** (idéogrammes chinois, hiraganas, katakanas, yi, hangul). Un mécanisme permettant d'étendre la capacité de représentation d'Unicode à plus d'un million de caractères utilise les codes **D800 à DFFF** (**extensions** codées sur deux mots de 16 bits).

Il indique que quinze autres ne sont pas encore intégrés (entre parenthèses la langue associée au système d'écriture et les pays où elle est parlée) :

- ⇨ **Afrique** : tifinagh (tamazight ; Maroc),
- ⇨ **Moyen Orient** : samaritain (samaritain ; Israël),
- ⇨ **Asie du Sud** : chakma (chakma ; Bangladesh, Inde), lepcha (lepcha ; Bhutan, Inde, Népal), limbu (limbu ; Bhutan, Inde, Népal), meetai mayek (meetai ; Bangladesh, Inde), ol cemet (santali ; Inde), siloti nagri (sylhetti ; Bangladesh),
- ⇨ **Asie du Sud-Est** : batak (batak ; Philippines, Indonésie), bugi (bugi; Indonésie, Malaisie), cham (cham ; Cambodge, Thaïlande, Vietnam), hmong (hmong ; Laos, Chine), javanais (javanais ; Indonésie),
- ⇨ **Asie Orientale** : lisu (lisu ; Chine), ouïgour (ouïgour ; Chine, Kazakhstan, Kirghizistan).

Le standard Unicode reste incomplet, même avec ces systèmes d'écriture complémentaires. Par exemple, les écritures suivantes sont absentes du standard et de ses projets d'évolution : sibo, nakhi, tai dam, tai khao, tai deng, tai yo, lai pao, môn, tham, lü, khamti, shan, tai mau, tham laotien, tham isan, khün et jawi. Le standard Unicode prévoit cependant entre **E000 et F8FF** une **zone de 6400 codes destinés à l'usage privé**[1]. Cette réservation est définitive et la zone d'usage privé peut être utilisée de façon durable pour les systèmes d'écriture mal ou non pris en compte par Unicode. Un mécanisme d'intégration au standard est d'ailleurs possible lorsque l'usage d'un codage privé est devenu suffisamment répandu.

Plusieurs sources listent des systèmes d'écriture. Le projet de norme internationale ISO 15924[2], préparé par le comité technique ISO TC46/SC2, fournit une liste de codes pour différents systèmes d'écriture destinés à des usages en terminologie, en bibliographie, en lexicographie et en linguistique. Elle est aux écritures ce que la norme ISO 639 est aux langues. Le tableau des codes de cette norme est donné en annexe A.3. Un autre tableau est donné en annexe A.4. Il donne, pour 99 langues, le ou les systèmes d'écriture utilisés. Il est dérivé du standard Unicode 3.0. Plusieurs livres décrivent les systèmes d'écriture, en particulier [Février 1959], [Gelb 1973] et [Daniels & Bright 1996]. Enfin, le site Omniglot (http://omniglot.com/) présente différents systèmes d'écriture.

Les traitements réalisés par les logiciels et systèmes d'exploitation sur les codes — par exemple les méthodes de saisie ou le rendu à l'écran — ne font pas partie du standard Unicode qui ne fournit qu'un ensemble de couples codes-signification (par exemple, [0041 ; LATIN CAPITAL LETTER A]). Le rendu à l'écran est généralement réalisé par les systèmes d'exploitation, qui s'appuient pour cela sur les polices de caractères. Les polices de caractères ont évolué avec les possibilités des machines, passant de la technologie *bitmap* de la fin des années 1960, dans lesquelles les caractères sont représentés par une matrice de points (caractères de dimension fixe), à des technologies plus évoluées comme METAFONT, Type 1 ou TrueType, respectivement en 1983, 1985 et 1991. Toutes ces technologies ont en commun qu'à un caractère en mémoire est associé un glyphe (le dessin du caractère) et que les glyphes sont placés les uns à la suite des autres dans l'ordre des caractères du texte. Avec Unicode, une nouvelle technologie de polices a été nécessaire pour assurer le lien devenu plus complexe entre la représentation du texte en mémoire (une chaîne de caractères Unicode) et sa représentation visuelle. Cela est dû à une nouvelle approche dans laquelle plusieurs caractères peuvent correspondre à un seul glyphe, par exemple A (0041) + ◌̊ (030A[3]) = Å (00C5). De plus, les chaînes utilisées pour calculer les glyphes peuvent avoir subi un prétraitement réordonnant les caractères du texte en mémoire. Plusieurs technologies ont ainsi vu le jour, en particulier OpenType, AAT, Graphite, FreeType, ICU et Pango ([Bauhahn 2002]).

[1] Une autre zone à usage privé est prévue dans la zone étendue entre DB80:XXXX et DBFF:XXXX, XXXX étant compris entre DCOO et DFFF.

[2] Voir http://www.evertype.com/standards/iso15924/document/dis15924.pdf.

[3] Le caractère ◌, de code 25CC, indique que le signe ˚ est un signe suscrit.

I.1.4.2 BASE ASSURÉE PAR LES SYSTÈMES D'EXPLOITATION

Les systèmes d'exploitation actuels des micro-ordinateurs[1] intègrent la capacité Unicode en ce sens qu'ils présentent une interface de programmation compatible avec Unicode. Ils sont donc nativement multilingues, pour autant que le système d'écriture considéré soit dans Unicode et qu'une police de caractères existe et fonctionne pour ce système d'écriture[2]. Ces systèmes d'exploitation sont en particulier :

- ⇨ Windows depuis la version NT 3.1 en 1993[3],
- ⇨ MacOS à partir de la version 8.5 en 1998[4],
- ⇨ Linux à partir d'XFree86 4.0 en 2000[5].

L'élément de base permettant de créer du texte est la fenêtre d'édition (fenêtre dans laquelle on peut saisir du texte). Des fenêtres d'édition évoluées permettant l'édition dans plusieurs systèmes d'écriture contenus dans Unicode, voire dans tous, sont incluses dans les environnements de développement sous forme d'objets ou d'interfaces de programmation (API). L'utilisation de ces fenêtres d'édition permet un gain de temps considérable, ces objets étant devenus très complexes avec la prise en compte d'Unicode. Ils réalisent en effet les fonctions suivantes :

- ⇨ gestion des actions clavier et souris,
- ⇨ affichage du texte,
- ⇨ coupures de fin de ligne,
- ⇨ justification du texte,
- ⇨ gestion du mouvement du curseur,
- ⇨ sélection du texte (vidéo inversée),
- ⇨ copie-collage.

En plus de ces fonctions de base, les fenêtres d'édition courantes (HTML, RTF, Word...) gèrent l'association d'attributs — gras, italique, souligné, police... — à des parties de texte, grâce, généralement, à un balisage du texte.

---

[1] Nous nous intéressons principalement à cette catégorie de matériel qui constitue l'essentiel des plates-formes utilisées aujourd'hui par le grand public.

[2] Cela n'est cependant pas toujours le cas, même pour des écritures de langues officielles. Par exemple, la police Arial Unicode MS de Microsoft, qui contient pratiquement tout Unicode depuis sa création et en particulier le laotien, ne fonctionnait toujours pas pour cette dernière écriture à la sortie de Windows XP et d'Office 2002 (mauvaise compréhension du fonctionnement des voyelles suscrites et souscrites ainsi que des accents).

[3] Il s'agit surtout de la branche NT de Windows. L'autre branche (Windows 9x) a aussi une certaine capacité Unicode mais limitée à quelques API comme TextOutW qui permet l'affichage d'une chaîne Unicode. Les applications Unicode développées pour Windows 9x doivent ainsi traduire les chaînes Unicode en chaînes ANSI avant l'appel des fonctions.

[4] Le Macintosh avait cependant une capacité multilingue depuis 1992 (System 7) grâce à la technologie WorldScript qui utilise un codage propriétaire (ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/APPLE/). En 1995 (Système 7.5), la distribution de cette technologie fut intégrée à celle du système. À partir de 1996, les textes en Unicode purent être lus et enregistrés grâce à la conversion Unicode-WorldScript réalisée par le *Text Encoding Conversion Manager*. En 1998, la version 8.5 généralisa Unicode, permettant en particulier aux applications Macintosh de gérer les polices Unicode (y compris les polices Windows).

[5] Voir la page de Markus Kuhn : http://www.cl.cam.ac.uk/~mgk25/unicode.html.
Nota : Linux a choisi de rester « orienté huit bits » et d'utiliser UTF-8 pour faciliter cela. Cependant, bien que Ken Thompson ait inventé cet encodage UTF-8 dès 1992 (Ken Thompson, d'AT&T Bell Lab, est à l'origine d'Unix en 1969), la progression d'Unicode dans Linux est lente. Par exemple, si le **noyau Linux** permet, dès 1995, de passer les consoles texte en UTF-8, **XFree86** — l'interface graphique de Linux — ne permet d'utiliser Unicode et les polices TrueType qu'à partir de sa version 4.0, sortie en mars 2000 (le patch *xfsft*, de Mark Leisher et Juliusz Chroboczek, permettait, depuis déjà un an, d'utiliser des polices TrueType avec XFree86). La sortie, en juin 1999, de **QT 2** — générateur d'interfaces graphiques compatible Unicode — a permis à l'environnement **KDE** d'intégrer cette évolution dans sa version 2, en octobre 2000.

Ces fonctions, déjà assez lourdes à développer pour du texte en caractères latins, deviennent extrêmement complexes avec la prise en compte des contraintes liées à l'ensemble des systèmes d'écriture :

- ⇨ forme de caractères dépendant de leur voisinage (par exemple arabe, hébreu, thaï et hindi),
- ⇨ écritures sans séparateur entre mots (par exemple chinois, japonais [exemple ci-dessous], coréen, thaï et hindi),

ＮＴＴ西日本では、現在、Ｂフレッツのビジネスタイプ及びベーシックタイプ並びにフレッツ・ＡＤＳＬの電話と共用しないタイプにおいて、２４時間３６５日の故障修理を実施させていただく「サポートメニュー」を提供しておりますが、テナントビル等に入居しているビジネスユーザやマンション等に入居しているＳＯＨＯユーザ様等からのＢフレッツマンションタイプにおいても利用したいとのご要望にお応えするため、Ｂフレッツマンションタイプにおいても「サポートメニュー」を提供することとし、本日総務大臣に届出を行いました。

- ⇨ bidirectionalité (par exemple arabe [exemple arabe-latin ci-dessous], hébreu),

ملاحظة: باستخدام Microsoft Office Professional Edition 2003 ، يمكن استخدام Word لإنشاء مستندات محمية باستخدام IRM ومنح المستخدمين الآخرين الإذن بالوصول إلى المستندات وتعديلها. ويمكن أيضا تطبيق قوالب النهج على المستندات التي تقوم بإنشائها والمحمية باستخدام IRM. باستخدام Microsoft Office Standard Edition 2003 ، يمكنك قراءة المستندات المحمية باستخدام IRM ؛ كما يمكنك تعديلها ولكن مع وجود إذن للقيام بذلك.

- ⇨ écritures verticales (par exemple chinois [exemple ci-dessous], ouïgour).

序說
古早古早的古早以前
吾鄉的人們
開始懂得向上仰望
吾鄉的天空
就是那一副無所謂的模樣
無所謂的陰著或藍著
古早古早的古早以前
自吾鄉左側綿延而近的山影
就是一大幅
陰鬱的潑墨畫
緊緊貼在吾鄉人們的臉上
古早古早的古早以前
世世代代的祖先，就在這片
長不出榮華富貴的土地上
揮灑鹹鹹的汗水
繁衍認命的子孫
晨曦
雀鳥無關快樂不快樂的歌聲
還未醒來
吾鄉的婦女
已環坐古井邊
勤快地浣洗陳舊或不陳舊的流言
無關輝煌不輝煌的老太陽
還未爬上山頂
吾鄉的囝仔郎

Ainsi, des fonctions paraissant aussi basiques que la sélection de texte et même la gestion de la position du curseur deviennent de véritables casse-tête, en particulier avec des textes incluant à la fois les systèmes d'écriture arabe et latin.

Plusieurs classes de fenêtres permettent de gérer ces écritures complexes. Sous Windows, la classe *CRichEditCtrl*[1] encapsule le contrôle *Rich Edit* dont la version 3 couvre presque entièrement Unicode 3. Sous Linux/Unix, Windows et MacOS X, QT[2] propose la classe C++ *QTextEdit*[3] qui intègre plusieurs caractéristiques complexes comme la bidirectionalité (par exemple pour l'arabe et l'hébreu) et la césure des écritures sans séparateur entre mots (par exemple pour le chinois, le japonais, le coréen et le thaï). La bibliothèque Swing de Java offre plusieurs classes[4] dérivées de la classe de base *EditorKit*[5], permettant en particulier l'édition aux formats HTML (classe *HTMLEditorKit*) et RTF (classe *RTFEditorKit*).

[1] Cette classe fait partie de la bibliothèque *Microsoft Foundation Class* (MFC) de Microsoft. Voir la page http://msdn.microsoft.com/library/default.asp?url=/library/en-us/vcmfc98/html/mfchm.asp.
[2] http://doc.trolltech.com.
[3] http://doc.trolltech.com/3.2/scripts.html, http://doc.trolltech.com/3.2/qtextedit.html.
[4] http://java.sun.com/docs/books/tutorial/uiswing/components/generaltext.html.
[5] La classe *EditorKit* (http://java.sun.com/j2se/1.3/docs/api/javax/swing/text/EditorKit.html) est la composante « contrôle d'édition » de la classe *JTextComponent*.

De nombreuses applications compatibles avec Unicode ont été développées, en particulier des suites bureautiques et des navigateurs Internet. Certaines proposent des services linguistiques : détection automatique de la langue, formatage automatique de la date, coupure des mots en fin de ligne, segmentation (pour les écritures sans séparateur entre mots), correcteurs d'orthographe, de grammaire et de style, tri lexicographique, dictionnaire de synonymes, résumé automatique, etc.

Par exemple, Office XP[1], l'une des suites bureautiques les plus répandues, inclut des outils linguistiques pour quarante-huit langues[2]. Certaines de ces applications sont elles-mêmes des objets pouvant être utilisés comme plates-formes pour informatiser des langues-π.

Parmi les objets utilisés dans nos développements figurent la classe EDIT, fenêtre d'édition simple ne permettant d'afficher qu'une police à la fois, la classe *CRichEditCtrl*, fenêtre d'édition RTF[3] et l'application *Word*, traitement de texte incluant la très performante classe d'édition *_WwG*.

I.1.4.3 TECHNOLOGIES UTILISÉES POUR RÉALISER LES SERVICES ET DES RESSOURCES

Nous rappelons ci-dessous, sous forme de tableau, les ressources et logiciels que nous avons retenus pour définir l'informatisation d'une langue (voir I.1.1 et I.1.1.2) en précisant leur utilité.

	Services / ressources	**Commentaire**
Traitement du texte		
	Saisie simple	Base pour la communication et l'archivage (langues écrites).
	Visualisation / impression	
	Recherche et remplacement	
	Sélection du texte[4]	
	Tri lexicographique	Complément à l'édition de base, aide de premier niveau incluant les fonctions usuelles.
	Correction orthographique	
	Correction grammaticale	Complément à l'édition de base, aide secondaire en terme de criticité.
	Correction stylistique	
Traitement de l'oral		
	Synthèse vocale	Permet d'accéder à l'information et à l'expression aux locuteurs des langues non écrites ou aux illettrés et peut constituer une aide pour les personnes handicapées.
	Reconnaissance de la parole	
Traduction		
	Traduction automatisée	Permet d'accéder à l'information et à l'expression dans la langue considérée, en particulier sur Internet.
ROC		
	Reconnaissance optique de caractères	Gain de productivité lorsqu'une importante base papier préexiste.
Ressources		
	Dictionnaire bilingue	Outils à buts didactiques, professionnels ou pratiques.
	Dictionnaire d'usage	

Figure 6 : Utilité des éléments retenus pour l'informatisation d'une langue

[1] http://www.microsoft.com/office/previous/xp/multilingual/prooftools.asp.

[2] Ces langues sont les suivantes : allemand, anglais, arabe, basque, bulgare, catalan, chinois simplifié, chinois traditionnel, coréen, croate, danois, espagnol, estonien, finnois, français, galicien, gallois, grec, gujrati, hébreu, hindi, hollandais, hongrois, indonésien, italien, japonais, kannada, letton, lituanien, marathi, norvégien, polonais, portugais, portugais brésilien, penjabi, roumain, russe, serbe, slovaque, slovène, suédois, tamoul, tchèque, telugu, thaï, turc, ukrainien et vietnamien.

[3] RTF (Rich Text Format) est un format d'échange entre traitements de texte.

[4] Pour les langues à écriture non segmentée, ce service n'est pas fourni directement par les classes de fenêtres d'édition standard.

Nous proposons ci-dessous quelques exemples de développements qu'il est nécessaire de réaliser spécifiquement pour ces services ou ressources lorsque les logiciels disponibles sont insuffisants.

Saisie simple & visualisation / impression		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture contenant des symboles non disponibles (c'est à dire ni latins, ni cyrilliques, ...)	Police de caractères	A.1.3.2
Système d'écriture pour lequel il n'existe pas de clavier virtuel	Clavier virtuel	A.1.3.3
Système d'écriture vertical	Editeur adapté à l'écriture verticale	A.1.3.5

Recherche et remplacement		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Langue écrite dont les mots peuvent être écrits de plusieurs manières (orthographe non fixée ou problème inhérent à la méthode de saisie)	Outil normalisant toutes les formes possibles d'un mot	A.1.9.1

Sélection du texte		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture non segmenté dont la sélection de texte à la souris et au clavier est problématique	Segmenteur adapté	A.1.4.3

Tri lexicographique		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3
Langue écrite dont l'ordre lexicographique n'est pas pris en compte	Outil de tri adapté	A.1.3.4
Langue écrite ayant plusieurs ordres lexicographiques différents	Outil de tri paramétrable	A.1.3.4

Correction orthographique		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3
Langue écrite dont la correction orthographique n'est pas prise en compte par les logiciels existants	Outil de correction d'orthographe	A.1.4.2

Correction grammaticale & stylistique		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3
Langue écrite dont la correction grammaticale ou stylistique n'est pas prise en compte par les logiciels existants	Outil de correction grammaticale et / ou stylistique	A.1.4.2

Synthèse vocale		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3
Langue écrite pour laquelle il n'existe pas d'analyseur morphologique	Analyseur et générateur morphologiques pour la langue	A.1.4.3
Langue écrite pour laquelle il n'existe pas d'analyseur syntaxique	Analyseur syntaxique pour la langue	A.1.4.3
Langue écrite pour laquelle il n'existe pas de logiciel de synthèse vocale	Logiciel de synthèse vocale	A.1.5

Reconnaissance de la parole		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Langue écrite pour laquelle il n'existe pas de logiciel de reconnaissance de la parole	Logiciel de reconnaissance de la parole	A.1.6

Traduction automatisée		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Aide à la traduction		
Langue écrite pour laquelle il n'existe pas de logiciel d'aide à la traduction	Bases de données lexicales bilingue ou multilingue	A.1.8
Langue écrite pour laquelle il n'existe pas d'analyseur ni de générateur morphologique	Analyseur et générateur morphologiques pour la langue	A.1.4.3
Langue écrite pour laquelle il n'existe pas de logiciel d'aide à la traduction	Logiciel d'aide à la traduction (mot à mot ou utilisant des mémoires de traduction)	A.1.7
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3
Traduction automatique		
Langue écrite pour laquelle il n'existe pas de logiciel de traduction automatique	Logiciel et linguiciel de traduction automatique	A.1.7
Langue écrite pour laquelle il n'existe pas d'analyseur / générateur morphologiques	Analyseur et générateur morphologiques pour la langue	A.1.4.3
Langue écrite pour laquelle il n'existe pas d'analyseur et de générateur syntaxique	Analyseur et générateur syntaxiques pour la langue	A.1.4.3
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3

Reconnaissance optique de caractères		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Système d'écriture pour lequel il n'existe pas de logiciel de reconnaissance optique de caractères	Logiciel de reconnaissance des caractères	A.1.9.1
Système d'écriture non segmenté	Segmenteur adapté	A.1.4.3

Dictionnaire bilingue		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Langue écrite pour laquelle il n'existe pas de dictionnaire électronique bilingue	Base de données lexicales bilingue ou multilingue	A.1.8

Dictionnaire d'usage		
Type de langue ou de système d'écriture	*Développement*	*Voir annexe A.1*
Langue écrite pour laquelle il n'existe pas de dictionnaire d'usage	Base de données lexicales monolingue	A.1.8

I.2 MÉTHODES POUR L'INFORMATISATION DES LANGUES PEU DOTÉES

I.2.1 Une informatisation sous contraintes

L'informatisation des langues-π est souvent rendue difficile pour des raisons diverses, incluant :

- ⇨ les difficultés linguistiques :
  - o faible niveau de description de la langue,
  - o langue peu ou pas écrite,
- ⇨ les faibles nombre, revenu et compétences des locuteurs, ce qui entraîne :
  - o moins d'utilisateurs potentiels,
  - o moins de linguistes potentiels,
  - o moins de développeurs potentiels.

Ces difficultés doivent être évaluées au début du projet d'informatisation, de même que les ressources dont on dispose pour la langue : dictionnaires, grammaires, logiciels déjà existants, personnes compétentes et motivées pour mener à bien le projet... Le présent chapitre propose **six méthodes** pour réduire les difficultés et faciliter le projet.

I.2.2 Trouver des solutions adaptées : idées forces

I.2.2.1 BÉNÉFICIER DE DÉVELOPPEMENTS FAITS POUR DES LANGUES LIÉES

I.2.2.1.1 Présentation

Au chapitre I.1.2.3, nous avons vu que des langues voisines pouvaient être informatisées conjointement pour optimiser les ressources. Une autre façon de bénéficier des liens qui unissent une langue à une autre déjà bien informatisée consiste à utiliser les ressources de cette dernière pour faciliter les développements. Deux cas se présentent :

- ⇨ la langue bien informatisée est une langue proche sur le plan linguistique (lien **génétique** (langues de la même famille) : base commune pour le vocabulaire et la syntaxe, ou **typologique** (affinité entre langues) : emprunts terminologiques...,
- ⇨ la langue bien informatisée est une langue centrale, utilisée par la population dont la langue est à informatiser pour communiquer hors du contexte grégaire[1] du groupe (lien **gravitationnel** : base sociologique favorable, bilinguisme, dictionnaires et grammaires réalisés par le groupe dominant).

I.2.2.1.2 Digression

Le deuxième cas ci-dessus suggère que, dans une optique d'informatisation de nombreuses langues, procéder en commençant par les langues supercentrales et centrales devrait optimiser l'ensemble du processus. En effet, étant des pivots pour des langues périphériques, leur informatisation peut faciliter celle de ces dernières : les dictionnaires bilingues, lorsqu'ils existent — ou, à défaut, les compétences bilingues — sont en effet situés entre langues périphériques et langues pivot. De plus, les langues pivot, en permettant de faire dialoguer plusieurs populations entre elles, touchent davantage de personnes et ont un rôle que l'informatique peut utilement compléter (documents officiels...), alors que les langues périphériques ayant, quant à elles, souvent un rôle essentiellement grégaire, nécessitent moins d'être informatisées.

Ainsi, si rien n'empêche d'informatiser telle langue périphérique avant telle langue centrale (le basque peut être considéré comme périphérique et est bien informatisé grâce au dynamique groupe IXA de San Sebastian), la démarche langue centrale → langue périphérique correspond à une logique de « crucialité ». Par exemple l'informatisation du bengali est plus cruciale que celle du morvandiau (cité comme l'une des langues de la République dans le rapport Cerquiglini en 1999).

[1] Voir le glossaire.

Notons par ailleurs que le coût de l'informatisation des 6809 langues peut être exorbitant. Rien que la réalisation d'un dictionnaire papier et d'une grammaire de base coûte environ 150 000 € par langue pas encore décrite[1,2].

Source de l'estimation	Coût estimé	Coût en €
Foundation for Endangered Languages	35 000 £	53 000 €
Robert Dixon (the rise and fall of languages, 1997)	200 000 $	170 000 €
David Crystal (language death, 2000)	120 000 £	180 000 €

Figure 7 : Coût estimé de la réalisation d'un dictionnaire et d'une grammaire de base

David Crystal estimant le nombre de langues non décrites à trois mille, cela donne un coût d'environ cinq cent millions d'euros pour ces seuls éléments de base. Ce coût n'est d'ailleurs pas très élevé si on le compare au coût de l'informatisation elle-même. À titre d'exemple, Mathieu Lafourcade ([Lafourcade 1994], page 1) estime le temps nécessaire à l'élaboration des linguiciels d'un système de traduction automatique à plus de cinquante hommes-ans par couple de langues.

Ces coûts sont à rapprocher des PIB des États les plus riches en langues : 387 langues en Inde, 279 au Cameroun, 219 en République Démocratique du Congo, 202 en République Populaire de Chine, 169 aux Philippines, 137 en Tanzanie, 107 au Myanmar (données Ethnologue).

Ainsi, vouloir informatiser toutes les langues du monde ne semble pas être une ambition réaliste, d'autant que les langues évoluent et qu'il ne suffit pas de les informatiser toutes : il faut aussi maintenir les logiciels pour qu'ils parlent toujours la langue[3]. Une informatisation de qualité pour les quatre-vingt-dix à cent langues officielles du monde[4] représente déjà un objectif ambitieux.

I.2.2.1.3 Langues liées génétiquement ou typologiquement

Ce cas vise à tirer parti des ressemblances qu'une langue-π a avec une langue déjà informatisée. Il a été étudié par Michael Paul dans le cas de projets de traduction automatique [Paul 2001].

En Europe, des systèmes de traduction automatique de l'espagnol vers deux autres langues apparentées génétiquement — le catalan [Canals-Marote et al. 2001] et le galicien [Diz 2001] — sont déjà opérationnels et servent en particulier à produire des éditions, dans ces deux langues, de quotidiens espagnols très lus.

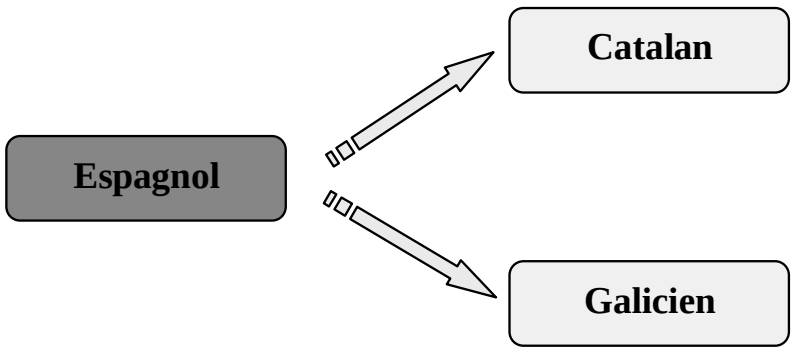


Figure 8 : Le catalan et le galicien bénéficient de leur parenté avec l'espagnol

[1] Noter l'existence de projets de description de ces langues comme http://www.hrelp.org/doc_home.htm.
[2] Exemples cités dans [Calvet 2002], page 107.
[3] Pour cet aspect maintenance, seuls des systèmes auto-apprenants pourraient parvenir à l'informatisation de toutes les langues du monde, par exemple via Internet si tant est que toutes les langues y soient un jour significativement représentées.
[4] Voir http://fr.wikipedia.org/wiki/Langue_officielle et http://www.tlfq.ulaval.ca/axl/.

En Asie, le système ATLAS-I de Fujitsu réalise la traduction entre les langues coréenne et japonaise grâce à la ressemblance très forte des systèmes morphologiques et surtout des syntaxes de surface des deux langues, bien que celles-ci ne soient pas apparentées. Cette parenté typologique provient de la profonde influence que le coréen, et à travers lui le chinois, a eue sur le japonais.

I.2.2.1.4 Langues liées « gravitationnellement »

Les locuteurs des langues-π communiquent généralement avec l'extérieur de leur groupe à travers une langue véhiculaire dite langue pivot et qui peut être centrale ou supercentrale. Bien que n'étant pas nécessairement de la même famille linguistique, il est fréquent qu'elles partagent des éléments qui sont de nature à faciliter l'informatisation de la langue-π : système d'écriture, éléments de vocabulaire, dictionnaires bilingues. Les synergies avec ces langues pivot seront donc à évaluer et devront conduire à une stratégie de développement les incluant.

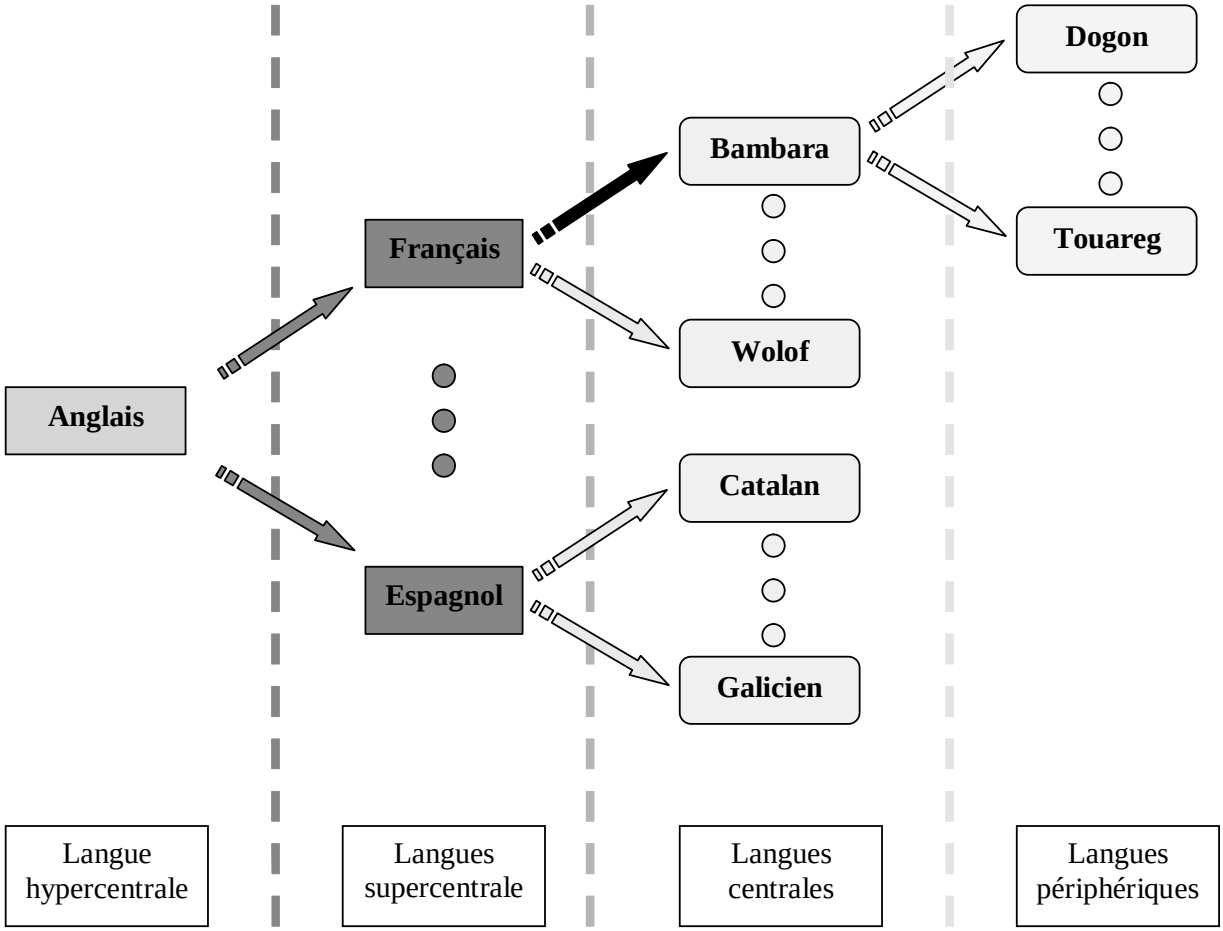


Figure 9 : Le français est une langue pivot pour le bambara et le wolof

Sur le plan technique, les outils permettant d'exploiter la parenté linguistique sont multiples. Parmi eux, l'analogie peut être utile en traduction automatique, par exemple pour la création de néologismes ou la traduction de noms propres ([Boitet 1997]) ou encore pour traiter divers problèmes de traduction lorsque l'on est en présence de phrases alignées ([Lepage 2000]).

I.2.2.2 S'INTÉGRER À DES PROJETS ET ENVIRONNEMENTS GÉNÉRIQUES, OPEN SOURCE ET À PIVOT

I.2.2.2.1 Le concept

Une part importante du temps nécessaire pour informatiser une langue est consacrée à la réalisation de l'environnement — création d'interfaces homme-machine, mise en place de bases de données... — qui peut être évité en intégrant un projet multilingue ayant déjà réalisé ces tâches périphériques de manière générique. Le caractère libre ou Open Source de ces projets peut aussi se révéler très avantageux.

En fonction du domaine visé, ces environnements génériques peuvent offrir des spécificités réduisant le temps de développement. Par exemple, dans le domaine de la traduction automatique ou dans celui des bases de données lexicales, la réalisation d'un système peut être significativement accélérée par l'adoption d'une architecture pivot. Dans une telle architecture, le développement d'une interface (avec le pivot) donne accès aux autres langues du projet.

I.2.2.2.2 Quelques exemples

Papillon[1] et UNL[2] sont deux exemples de projets multilingues incluant ou ayant inclus des langues-π. Par exemple, après une période initiale au cours de laquelle seules des langues bien dotées ont été impliquées, des langues moins bien dotées comme le mongol et le lituanien ont été étudiées dans le cadre du projet UNL.

Dans le domaine des logiciels Open Source ou libres, de nombreux groupes se sont d'ores et déjà créés pour adapter des logiciels comme la suite bureautique multi-plates-formes (Windows, Linux, Mac OS X / X11 et Solaris) OpenOffice.org (http://www.openoffice.org/) pour laquelle il existe déjà dix-huit adaptations : allemand, anglais, chinois simplifié, chinois traditionnel, coréen, danois, espagnol, français, hollandais, hindi, italien, japonais, portugais, portugais brésilien, suomi, suédois, thaï et turc, ou encore l'environnement graphique KDE, l'une des interfaces de Linux, pour lequel il existe soixante-dix-neuf groupes de traduction[3].

I.2.2.2.3 Avantages décisifs

Ce regroupement de plusieurs projets en un seul permet, en plus de la factorisation des efforts de développement pour les parties génériques, de mettre en contact des personnes ayant à résoudre les mêmes difficultés, ce qui limite les temps de recherche de solutions tout en augmentant la qualité grâce à l'émulation naissant du travail en commun.

[1] http://vulab.ias.unu.edu/papillon/.
[2] http://www.unl.ias.unu.edu/.
[3] http://i18n.kde.org/teams/index.php, ces langues sont Afrikaans, Albanian, Arabic, Azerbaijani, Basque, Belarusian, Bengali, Bosnian, Brazilian Portuguese, Breton, British English, Bulgarian, Catalan, Chinese Simplified, Chinese Traditional, Croatian, Czech, Danish, Dutch, Esperanto, Estonian, Faroese, Farsi, Finnish, French, Frisian, Galician, German, Greek, Gujarati, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Irish Gaelic, Italian, Japanese, Korean, Kurdish, Lao, Latvian, Lithuanian, Macedonian, Malagasy, Malay, Maltese, Maori, Marathi, Mongolian, Northern Sami, Norwegian Bookmal, Norwegian Nynorsk, Occitan, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Sorbian, Spanish, Swati, Swedish, Tagalog, Tajik, Tamil, Telugu, Thai, Tibetan, Turkish, Ukrainian, Venda, Vietnamese, Walloon, Welsh, Xhosa, Zulu.

I.2.2.3 RECOURIR À LA CONTRIBUTION LINGUISTIQUE GÉNÉRALISÉE : MUTUALISATION

I.2.2.3.1 Le concept

Selon notre point de vue, des ressources linguistiques de qualité peuvent être obtenues efficacement grâce à un **travail coopératif sur Internet** ([Boitet 1999]), remplaçant une équipe travaillant localement par un groupe de travail distribué, gratuit et potentiellement de beaucoup plus grande taille. Cette idée d'une « contribution linguistique généralisée » sur le web était au centre du projet Montaigne proposé en 1995 par le GETA et la société SITE-Eurolang mais qui n'avait pu, à cette époque, être réalisé faute de financement[1].

I.2.2.3.2 Quelques exemples

Ce concept de contribution linguistique généralisée a fait l'objet de plusieurs réalisations. Les laboratoires NII au Japon et NECTEC en Thaïlande ont ainsi développé un dictionnaire japonais-thaï[2] et la société Oki a mis en œuvre ce concept dans son site yakushite.net pour les traducteurs de japonais en anglais et vice-versa[3] [Shimohata 2001] sous la forme d'un site coopératif permettant à des traducteurs anglais-japonais d'utiliser le système de traduction automatique par patrons tout en consultant et enrichissant des dictionnaires plus ou moins spécialisés, organisés selon la hiérarchie des communautés d'utilisateurs du site.

À un autre niveau, l'organisation Open Language Archives Community[4] (OLAC) offre une plate-forme coopérative pour « créer une bibliothèque virtuelle mondiale de ressources de langues ». Démarré en décembre 2000, ce projet réunit des participants appelés « fournisseurs de données » (*data providers*) dont les ressources sont accessibles via des « fournisseurs de services » (*service providers*) tels que « *The Linguist* » [5]. D'autres projets, comme la *Lesser-Used Languages Software Developers' Association*[6] ou le projet Rosetta,[7] sont moins institutionnels mais remplissent aussi ce rôle de point de ralliement pour des initiatives isolées.

I.2.2.4 RECYCLER LES DICTIONNAIRES EXISTANTS

La réalisation d'une base lexicale de qualité est une tâche lourde dans le processus d'informatisation d'une langue. Une alternative à la construction complète et coûteuse d'une base lexicale est le retraitement de fichiers contenant des dictionnaires destinés à être utilisés sous forme papier. Les dictionnaires papier répondant à des exigences différentes des bases lexicales, la récupération n'est pas immédiate. Pour faciliter ce travail de récupération, il est possible d'utiliser des outils tels que *RecupDic* dans lequel la structure textuelle des dictionnaires est décrite formellement de façon à automatiser leur transformation en base lexicale [Doan-Nguyen 1998]. Notons qu'un tel processus ne peut souvent être mis en œuvre qu'après une étape de « normalisation » semi-automatique ad hoc, destinée à réduire le « bruit » [8].

---

[1] Il s'agissait de mettre gratuitement à disposition de communautés thématiques françaises, via le web, des aides à la traduction de et vers le français, et en particulier des mémoires de traduction et une version de l'environnement spécifique EuroLang Optimizer ajouté à des texteurs comme Word, WordPerfect et Interleaf. Le financement demandé était de l'ordre de 2 hxa, nécessaires pour adapter l'architecture du produit commercial de départ.

[2] Projet Saikam, http://saikam.nii.ac.jp/.

[3] http://www.yakushite.net/.

[4] http://www.language-archives.org/.

[5] http://www.linguistlist.org/olac/.

[6] http://www.egt.ie/lulsda/index.html.

[7] http://www.rosettaproject.org/.

[8] Par exemple, un blanc non gras entre deux mots formant un « mot-vedette » comme « base lexicale » peut créer une rupture invisible conduisant à « récupérer » seulement « base » comme mot-vedette, puis à d'autres erreurs dans la suite de l'article à récupérer.

I.2.2.5 Recourir aux diasporas et à Internet pour rapprocher les acteurs

I.2.2.5.1 Le concept

Les diasporas jouent un rôle important dans le développement de leurs pays d'origine. Formées dans des pays à technologie avancée, sensibilisées à leur culture d'origine dont elles sont séparées, motivées par l'idée d'apporter quelque chose à leurs compatriotes restés au pays, elles constituent une population essentielle dans le processus d'informatisation des langues peu dotées. Réunis sur des forums spécialisés dans leur région d'origine, leurs membres parlent en particulier des outils qui existent pour apprendre, écrire et traduire leur langue, et parfois des groupes se forment spontanément pour créer des polices, des dictionnaires ou des claviers virtuels.

I.2.2.5.2 Quelques exemples

Cette motivation des diasporas a récemment été canalisée par les Nations Unies qui ont créé successivement le *Digital Diaspora Network – Africa* (juin 2002, http://www.ddn-africa.org/) et le *Digital Diaspora Network – Latin America & caraïbebean* (janvier 2003, http://www.ddn-lac.org/). La mission de ces réseaux est de promouvoir le développement de leur zone d'action respective en s'appuyant sur l'expertise des diasporas.

I.2.2.6 Informatiser des groupes de langues liées génétiquement

L'acquisition simultanée de plusieurs langues proches — comme l'espagnol, l'italien, le portugais, le français... — demande beaucoup moins d'effort que celle des langues prises isolément. C'est ce qu'ont pu analyser les partenaires du **projet Galatea**[1] sur les langues romanes depuis le début du projet en 1991[2]. Galatea vise à « prendre appui sur la parenté des langues romanes pour en favoriser l'apprentissage et l'utilisation ».

Les didacticiens de Galatea préconisent un apprentissage « consécutif référencé » pour optimiser le processus d'enseignement, l'effort consistant à une assimilation par contraste. L'apprenant, stimulé par sa connaissance préalable d'une langue, acquiert la nouvelle langue en la comparant avec cette langue déjà connue, n'ayant besoin, pour cela, que de **mémoriser des différences**.

La réalisation de dictionnaires ou de logiciels pour un groupe de **n langues apparentées** peut, de la même manière, bénéficier d'une similitude entre les langues. Si l'on compare les quatre locutions proposées sur la page http://www.grenoble.iufm.fr/kiosque/lettreinfo/specialRI/page1.htm :

La cigala i les formigues,
La cigala y las hormigas,
La cicala e le formiche,
A cigarra e as formigas,

on peut aisément imaginer que la construction simultanée de dictionnaires pour plusieurs langues gagne à être faite par le même groupe projet.

Un tel groupe projet pourrait utilement travailler en collaboration avec des projets de plates-formes d'enseignement des langues comme Galanet[3], qui est basé sur les principes de Galatea. Destiné à donner une capacité d'intercompréhension ou de dialogue à des personnes côtoyant une diversité de langues parentes, ce type de projet peut bénéficier des outils réalisés par le projet d'informatisation, et inversement, ce dernier peut utiliser les ressources réalisées par les didacticiens.

[1] http://www.u-grenoble3.fr/galatea/classic.htm.
[2] Nous avons nous-même pu vérifier cela lors de l'apprentissage simultané du laotien et du siamois (thaï), deux langues voisines l'une de l'autre.
[3] http://www.galanet.be/.

I.2.3 Appliquer une gestion adaptée

I.2.3.1 DÉTERMINER QUEL PRODUIT RÉALISER

Compte tenu des moyens réduits dont on dispose généralement pour informatiser une langue-π, il est important de cibler le produit que l'on veut réaliser.

- ⇨ S'agit-il d'un produit destiné au grand public ou d'un prototype de laboratoire ?
- ⇨ S'agit-il d'un traitement de texte, d'un service de traduction en ligne ou d'un outil de traitement de l'oral ?
- ⇨ S'agit-il d'un logiciel pour Palm ou pour micro-ordinateur sous Linux / Windows / Mac OS ?
- ⇨ La cible dispose-t-elle de ces moyens ?
- ⇨ S'agit-il d'un produit gratuit ou payant ?
- ⇨ Quand sera-t-il disponible ?
- ⇨ Comment sera-t-il diffusé ?

Ces questions de marketing sont essentielles pour que le projet produise un résultat utile.

I.2.3.2 DÉTERMINER QUI RÉALISE LES LOGICIELS ET RESSOURCES

L'informatisation d'une langue-π par les locuteurs eux-mêmes peut être rendue très difficile par le fait qu'ils ne sont pas suffisamment formés pour réaliser le travail. D'une manière plus générale, l'informatisation d'une langue nécessite des informaticiens et / ou linguistes qui peuvent être ou non des locuteurs de la langue. Plusieurs cas de figure s'offrent à nous.

L'informatisation peut être réalisée par :

- ⇨ un groupe local créé spécifiquement,
- ⇨ un groupe travaillant à un projet Open Source / logiciel libre,
- ⇨ des membres de la diaspora travaillant en réseau,
- ⇨ une société spécialisée,
- ⇨ un laboratoire universitaire scientifique,
- ⇨ un institut de langue et de linguistique,
- ⇨ une association entre plusieurs de ces possibilités.

Selon le financement prévu pour le projet, la participation d'un groupe local peut être impérative. C'est le cas de financements pour l'aide au développement (ICT4D) qui visent à l'acquisition d'une compétence informatique au sein des pays en développement[1]. Dans tous les cas, le besoin de financement devra être évalué au début du projet.

I.2.3.3 ÉTABLIR UN PLAN DE DÉVELOPPEMENT

L'informatisation d'une langue est un projet complexe dont la maîtrise passe par un découpage en tâches élémentaires qu'il faut organiser dans un **plan de développement**. À titre d'exemple, le plan de développement suivi, pour la langue basque, par le **groupe IXA**[2] de l'**université du Pays Basque** comprend schématiquement **cinq phases**[3] : 1) description de la langue, 2) morphologie, 3) syntaxe, 4) sémantique, 5) traduction automatique. En 2000, après douze années d'existence, il avait réalisé les trois premières phases et travaillait sur la quatrième. Parmi ses réalisations, citons une base de données lexicales, un analyseur et un générateur morphologiques, un lemmatiseur, un correcteur d'orthographe, un moteur de recherche.

[1] Le financement d'un projet d'informatisation de langue-π peut aller de l'auto-financement — par exemple dans le cas d'un groupe de bénévoles de la diaspora travaillant en groupe grâce à Internet — jusqu'au financement complet d'une société spécialisée, par exemple par un organisme public. Les salaires de chercheurs travaillant sur ces thèmes constituent des financements indirects.

[2] http://ixa.si.ehu.es.

[3] Voir une description détaillée dans [Agirre et al. 2001].

I.3 PLAN DE DÉVELOPPEMENT POUR UNE LANGUE DONNÉE, EXEMPLE DU LAOTIEN

I.3.1 Présentation

Ce chapitre présente la méthodologie et les étapes des développements que nous avons menés depuis fin 1996 sur la langue laotienne. Ces développements se sont déroulés en **deux phases** distinctes ayant nécessité chacune un **plan de développement** spécifique :

- la réalisation d'un traitement de texte bien adapté à l'écriture laotienne,
- le développement d'outils d'aide à la traduction et d'une base lexicale.

Les logiciels réalisés pendant ces deux phases sont résumés dans le tableau ci-dessous avec les dates auxquelles ils ont été disponibles.

Logiciel ou ressource	Date de réalisation	Description
Phase 1 (avant la thèse) — Premier plan de développement		
Tallao 1.0	Février 1997	Plate-forme d'expérimentation de fonctions pour un traitement de texte laotien
Tallao 2.0	Septembre 1997	
Tallao 3.2	Juin 1998	
LaoPad 1.0	Octobre 1998	Traitement de texte autonome
Site Lao Software	Juin 1999	Site Internet proposant logiciels et services
LaoWord 1.0	Septembre 1999	Complément laotien pour Microsoft Word (librairie dynamique)
LaoWord 2.0	Mars 2000	
LaoWord 3.0	Juillet 2000	
LaoWord 3.01	Novembre 2000	
Phase 2 — Deuxième plan de développement		
Police Times New Roman Lao	Mai 2001	Police Unicode laotienne
PapiLex	Mai 2001	Site Internet d'expérimentation de l'Unicode laotien en relation avec la lexicographie et le projet Papillon
Sylla 1.0	Juillet 2002	Outil linguiciel pour la mise au point de modèles syllabiques[1]
Sabaidi	Août 2002	Site Internet proposant des services de construction de dictionnaires et d'aide à la traduction (1er temps : construction de dictionnaires)
LaoUniKey 1.0	Août 2002	Clavier pour l'Unicode laotien
LaoUniKey 2.0	Mars 2003	
LaoWord 3.02	Mars 2003	Évolution de LaoWord (utilisation de COM)
LaoLex 1.0	Avril 2003	Fonction d'aide à la traduction et de construction de dictionnaire (sur le site Internet Sabaidi)
BanglaWord 1.0	Mai 2003	Complément bengali pour Word[1] (saisie Unicode)
LaoMonoKey 1.0	Mai 2003	Clavier pour polices laotiennes sur 8 bits
LaoDict 1.0	Juillet 2003	Dictionnaire construit avec LaoLex (présenté au séminaire *Papillon 2003*)
LaoWord 4.0	Septembre 2003	Évolution de LaoWord (Unicode)

Figure 10 : Logiciels et ressources réalisés pour les deux plans de développement présentés

[1] Liés à l'informatisation d'un groupe de langues, Sylla et BanglaWord ne seront présentés que dans la 3e partie.

I.3.2 Première phase (1996-2000) : réaliser un traitement de texte pour le laotien

Le logiciel Tallao, développé de décembre 1996 à juin 1998, écrit entièrement en C / C++ et utilisant uniquement le SDK (Software Development Kit) de Windows, visait les premiers niveaux du service de traitement du texte (voir I.1.1). Il inclut les fonctionnalités suivantes :

- saisie de textes laotiens indépendante de la police utilisée et utilisant un clavier intuitif,
- changement de police (donc transcodage),
- mise en forme canonique du texte sélectionné (standardisation, saisie non univoque)
- facilité de sélection à la souris et au clavier des syllabes et des mots laotiens,
- formatage de textes laotiens, pour les rendre utilisables par des traitements de texte commerciaux,
- export aux formats TeX et RTF,
- construction d'un lexique à partir de textes (ajouter, modifier, supprimer une entrée dans un lexique local),
- traduction en français de mots laotiens,
- transcription phonétique du texte sélectionné.

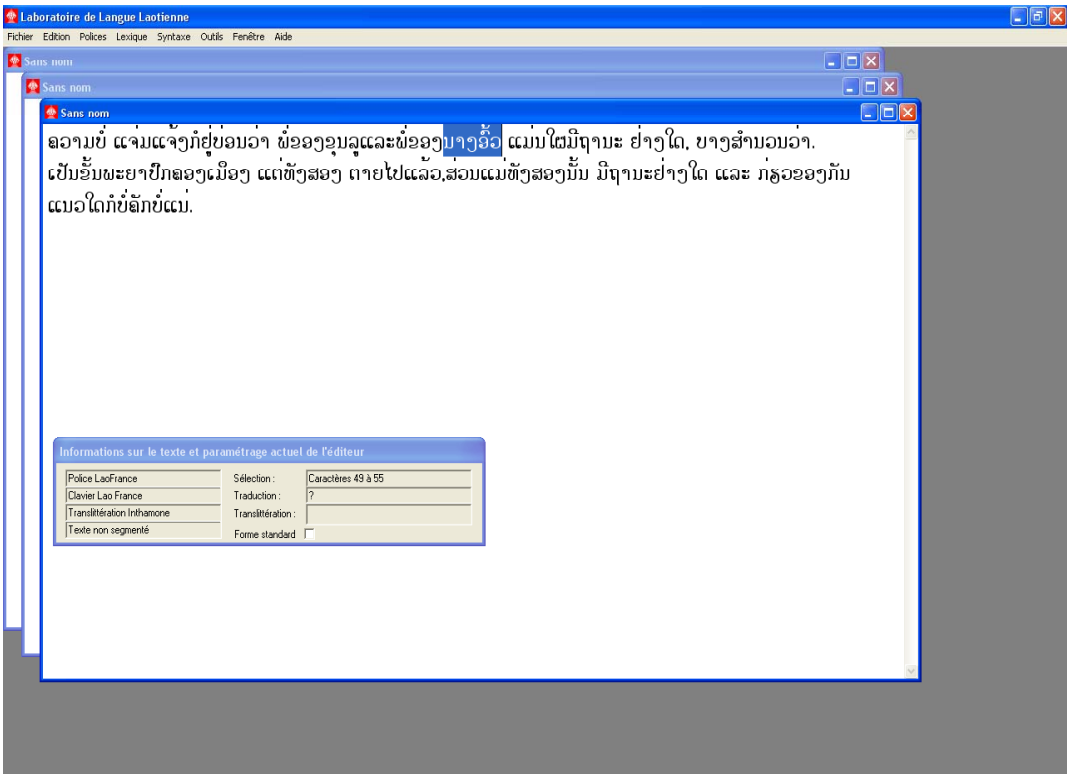


Figure 11 : Application Tallao 3.2

Suite à ce travail universitaire, **un premier plan de développement** s'est dessiné spontanément pendant l'été 1998. La plate-forme de recherche Tallao 3.2 nous avait permis de maîtriser les aspects technologiques des traitements au niveau graphotaxique (voir les difficultés liées à l'écriture laotienne au chapitre II.1.2) mais n'était pas assez ergonomique par rapport aux standards du marché des traitements de texte. En particulier, la fenêtre d'édition ne permettait d'avoir qu'une police à la fois et plusieurs fonctions, telles que l'impression, n'étaient pas développées. Il fallait aller un peu plus loin et **offrir un traitement de texte grand public de qualité offrant des fonctionnalités laotiennes**. Un emploi plus simple du système d'écriture laotien était le premier besoin qui s'imposait.

Ce premier plan de développement peut se résumer en quatre étapes :
1) développer avant fin 1998 un traitement de texte grand public :
- offrant les fonctionnalités courantes des traitements de texte (en particulier l'impression et une fenêtre d'édition capable de gérer le format *rich text*),
- offrant une saisie en laotien pratique et indépendante de la police,
- offrant des claviers virtuels[1] intégrés et adaptés aux habitudes des différentes communautés,
- résolvant les problèmes de non-standardisation et de saisie non univoque,
- évolutif,

2) le diffuser gratuitement via Internet,
3) l'enrichir d'une fonction de tri lexicographique,
4) l'adapter à Unicode lorsque ce standard sera opérationnel pour le laotien.

Nous nous sommes vite rendu compte qu'il serait long et difficile de dériver un tel produit de Tallao qui s'appuyait uniquement sur le SDK de Windows. Nous nous sommes tourné vers le traitement de texte WordPad qui pouvait servir de base à une version laotienne, ses sources étant disponibles dans Visual C++, l'environnement de développement C++ de Microsoft.

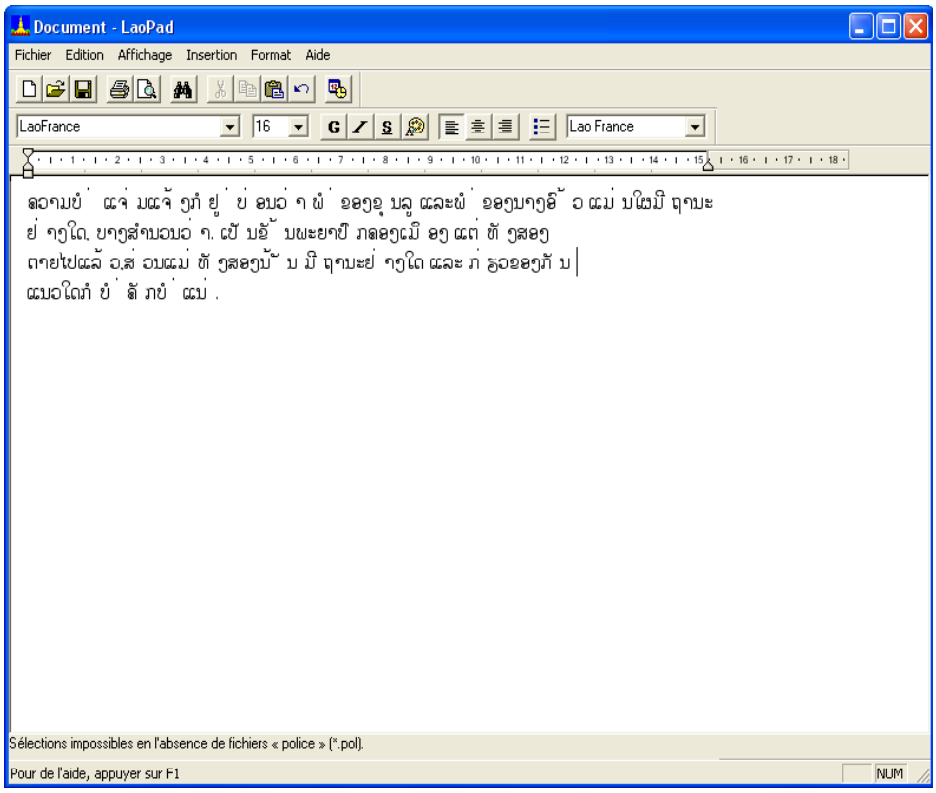


Figure 12 : Application LaoPad 1.0

WordPad est développé dans le langage objet C++ et s'appuie sur les classes de la bibliothèque MFC (*Microsoft Foundation Classes*). La classe de fenêtre d'édition *CRichEditCtrl* vue au chapitre I.1.4.2 y est utilisée et de nombreuses fonctions, dont la fonction d'impression et la gestion du format *rich text*, y sont intégrées.

LaoPad 1.0 est sorti en octobre 1998, incluant les « fonctions laotiennes » définies précédemment, en particulier la saisie indépendante de la police et les claviers virtuels. Nous sommes allé, en novembre, le présenter à l'ambassade de France au Laos, à la Bibliothèque Nationale du Laos et à des distributeurs laotiens. Lors de ce premier contact, M. Sysouk Keothammavong – directeur d'Alice Computer, l'une des principales entreprises d'informatique au Laos à cette époque – nous suggéra

[1] Voir le chapitre II.2.1 pour une définition des claviers virtuels.

d'intégrer les fonctions laotiennes de LaoPad dans Microsoft Word, d'en faire une version 16 bits pour Windows 3.1 — système d'exploitation largement répandu au Laos — et de faire une procédure d'installation sécurisée pour en faciliter la vente. Son avis se basant sur son expérience du marché laotien, nous décidâmes d'étudier une telle solution, LaoPad, basé sur WordPad ne pouvant pas fonctionner sous Windows 3.1. Nous ne retînmes cependant pas la réalisation de la procédure d'installation sécurisée que nous jugeâmes contraire à une diffusion efficace du logiciel.

La solution choisie fut d'utiliser le kit du développeur Word (*Word Developer's kit*), un petit ensemble de modules C qui permet de s'interfacer avec Microsoft Word. Le principe consiste à développer une librairie dynamique qui est chargée automatiquement par Word lors de son démarrage et qui appelle les fonctions de Word à travers une API appelée CAPI[1]. Bien que délicate, l'intégration des fonctions laotiennes à Word n'a pas posé de très gros problèmes à l'exception néanmoins des sélections syllabiques dont la réalisation a posé un sérieux problème de faisabilité et nécessité plusieurs « astuces », l'interface avec Word étant relativement limitée. **LaoWord 1.0** n'est ainsi sorti qu'en septembre 1999, les difficultés technologiques d'intégration à Word s'étant révélées longues à surmonter.

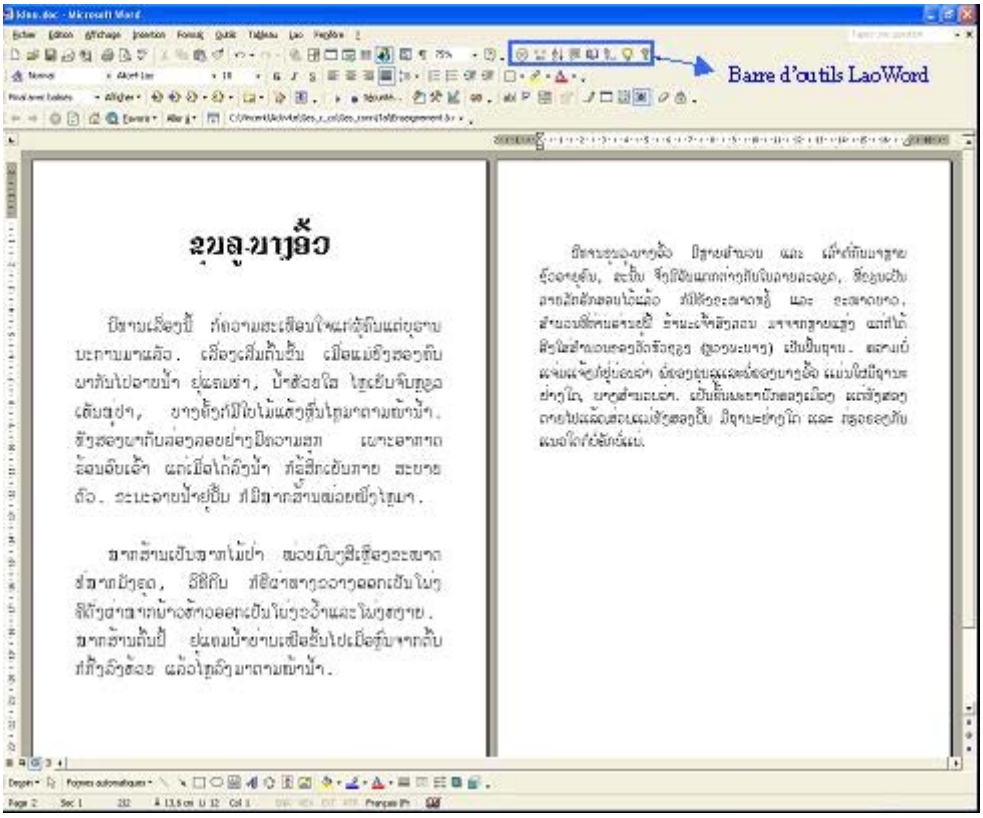


Figure 13 : Microsoft Word intégrant la librairie dynamique LaoWord

Le site **Lao Software** – www.LaoSoftware.com – créé peu de temps auparavant pour satisfaire l'objectif de diffusion gratuite sur Internet, n'est véritablement devenu opérationnel qu'avec l'arrivée de LaoWord. Sa promotion s'est faite grâce au référencement du site mais aussi et surtout par de nombreux liens placés sur des sites Internet dont plusieurs sites populaires auprès des laotiens.

[1] CAPI est distribué avec le livre [Microsoft Corporation 1995] « Microsoft Word Developer's Kit Third Edition », ISBN 1-55615-880-7.

Trois autres versions de LaoWord ont suivi. La **version 2**, mise au point pour la « fête de l'Internet » et le lancement du site Internet de l'ambassade de France au Laos en mars 2000, contenait un petit dictionnaire laotien-français. La **version 3**, diffusée en versions française et anglaise en novembre 2000, incluait en plus les fonctions de tri lexicographique (prévue par le plan de développement), de transcription phonétique et de mise en forme du texte. La **version 4**, qui incorpore la capacité Unicode, est sortie en septembre 2003 et se trouve ainsi chronologiquement rattachée à la phase deux. Chacune de ces versions a demandé une phase de test importante, en particulier à cause de l'exigence de compatibilité avec toutes les versions de Windows et de Word apparues depuis dix ans.

Contrairement aux logiciels Tallao et à LaoPad qui s'exploitent par l'intermédiaire de menus, LaoWord recourt à une barre d'outils pour le dialogue homme-machine. La description ci-dessous donne une vision concrète de l'exploitation de LaoWord. Elle fait référence à des fonctions qui sont détaillées au chapitre II.2.

Figure 14 : Barre d'outils LaoWord

De la gauche vers la droite, les fonctions des boutons sont :

- la configuration de LaoWord,
- le changement de polices laotiennes,
- le tri de tableaux en laotien,
- les transcriptions laotiennes,
- le dictionnaire électronique,
- la mise en forme du texte,
- le choix laotien-latin pour la saisie des textes,
- l'aide en ligne.

Configuration de LaoWord

Le premier bouton en partant de la gauche donne accès à la configuration de LaoWord :

- **clavier** : Lao France, Lao France New, Duang Jan, Lao US ou Sida Thong,
- **mode de sélection** : le mode standard de Word ou le mode LaoWord (par syllabes laotiennes),
- activation ou désactivation de quelques **raccourcis clavier**,
- paramétrage de la **transcription** (libre, phonétique),
- paramétrage du **tri** (ordre alphabétique, position des groupes consonantiques),
- paramétrage des **ligatures** en saisie et lors des changements de police,
- **chiffres** arabes ou laotiens (quand la police dispose des deux possibilités),
- activation ou désactivation de la **fenêtre d'accueil de LaoWord** au lancement.

Changement de polices laotiennes

En cliquant sur cette icône, on obtient la liste de toutes les polices laotiennes. Si on est en cours de saisie d'un texte avec une police latine, ce même bouton permet de passer à une saisie en police laotienne tout en conservant, approximativement, la taille de la police courante.

 **Tri de tableaux en laotien**

Le troisième bouton donne accès au tri des tableaux en laotien. Cette fonction permet d'ordonner les tableaux selon un ordre lexicographique à choisir entre les deux principaux ordres existants ainsi que définir la place des groupes consonantiques.

 **Transcriptions laotiennes**

Le quatrième bouton donne accès à des transcriptions de textes laotiens. Les différentes possibilités sont une transcription libre (simple à lire), une transcription phonétique (basée sur l'Alphabet Phonétique International) et une translittération réversible (qui retranscrit l'écriture et non pas les sons). Le texte sélectionné peut, au choix, être transcrit dans une fenêtre à part ou être remplacé par la transcription.

 **Dictionnaire électronique laotien-francais**

Le cinquième bouton donne accès à un dictionnaire laotien-français. Ce lexique est entièrement évolutif. Il est ainsi possible de modifier les traductions proposées et de créer de nouvelles entrées.

 **Fonctions de mise en forme des textes laotiens**

Le sixième bouton donne accès à des fonctions de finition des textes en laotien. Il permet :

- d'insérer automatiquement des caractères de césure entre les syllabes pour réaliser une bonne mise en page,
- de régler la hauteur des accents lorsque la police dispose de deux hauteurs,
- de régler la hauteur des voyelles ◌ᩮ et ◌ᩩ lorsque la police dispose de deux hauteurs,
- de ligaturer les voyelles suscrites et les accents lorsque la police le permet.

 **Basculement laotien-latin**

Ce bouton permet de passer d'une saisie en écriture latine (français, anglais, ...) à une saisie en laotien, et vice-versa. Cela n'est possible qu'avec les polices qui contiennent les deux types de caractères, ce qui est le cas en particulier des polices Unicode.

 **Aide en ligne**

Une aide en ligne est accessible via le huitième et dernier bouton.

Le premier plan de développement — hors capacité Unicode — a donc été mis en œuvre sur une période d'un peu plus de deux ans (de juillet 1998 à novembre 2000). Il a recouru à la méthode préconisée au chapitre I.2.2.2 — s'intégrer à des projets et environnements génériques existants — s'étant appuyé sur des traitements de texte existants et adaptables : WordPad et Word. Cette méthode s'est révélée très efficace, l'effort nécessaire ayant été très faible par rapport à celui qu'aurait nécessité le développement d'un traitement de textes complet.

Le tableau suivant donne une image synthétique et chronologique des fonctions de traitement du texte de plusieurs de ces logiciels. Les logiciels de la deuxième phase sont présentés au chapitre suivant et détaillés dans la deuxième partie. Les valeurs numériques données pour les claviers virtuels et les transcriptions indiquent le nombre de ces éléments inclus dans chaque logiciel.

		PHASE 1			PHASE 2				
	TalLao 3.2 06-98	LaoPad 1 10-98	LaoWord 1 09-99	LaoWord 3 11-00	PapiLex 1 05-01	LaoUni-Key 1 08-02	LaoLex 1 04-03	LaoUni-Web 1 04-03	LaoWord 4 07-03
Claviers virtuels	1	2	2	2	-	3	-	1	5
Saisie encodage 8 bits	x	x	x	x	-	-	-	-	x
Saisie encodage Unicode	-	-	-	-	-	x	-	x	x
Changement polices	x	x	x	x	-	-	-	-	x
Capacité « texte riche »	-	x	x	x	-	-	-	-	x
Sélections par syllabes	x	x	x	x	-	-	-	-	x
Sélections par mots	x	-	-	-	-	-	-	-	-
Césure calculée dynamiquement	x	-	-	-	-	-	-	-	-
Césure par insertion de CSL	x	-	-	x	-	-	-	-	x
Gestion d'un dictionnaire	x	-	x	x	x	-	x	-	x
Traduction laotien-français (mots)	x	-	x	x	-	-	x	-	x
Tri lexicographique	-	-	-	x	-	-	x	-	x
Transcriptions	2	-	-	3	-	-	x	-	3
Standardisation	x	-	-	-	-	-	-	-	-
Export TeX	x	-	-	-	-	-	-	-	-
Export RTF	x	x	x	x	-	-	-	-	x

Figure 15 : Fonctions incluses dans divers logiciels développés entre 1998 et 2003

I.3.3 Deuxième phase (2001-2003) : offrir une aide à la traduction et une base lexicale

Le commencement de la deuxième phase coïncide avec le début de cette thèse, en décembre 2000. Avec LaoWord 3.01, le plan de développement initial était réalisé, à l'exception du passage à Unicode. Le deuxième plan de développement consistait à étendre les services proposés à d'autres facettes que le « traitement du texte » et plus particulièrement à la traduction laotien-français.

Pour cela, nous avons cherché à intégrer la langue laotienne à un projet multilingue à pivot (voir le chapitre I.2.2.2), le projet Papillon, afin de développer un dictionnaire électronique laotien. Conçu en janvier 2000, ce projet Papillon a pour vocation d'être une base lexicale multilingue. Il contenait à l'origine le français et le japonais puis s'est étendu, en 2001 à l'anglais, au malais, au thaï et au vietnamien, en 2002 à l'allemand, et en 2003 au chinois. Dans Papillon, les articles sont en Unicode et formatées en XML. La décision de principe d'ajouter le laotien fut prise début 2001. Il n'existait alors ni police Unicode fonctionnant de manière satisfaisante[1] ni clavier virtuel pour la saisie, et la faisabilité d'un outil sur Internet pour gérer une base de données lexicales contenant du laotien était encore hypothétique[2]. La création de la police Unicode **Times New Roman Lao** et du site Internet **PapiLex** permirent de montrer, en mai 2001, la faisabilité d'une base lexicale en Unicode et au format XML pour le laotien avec une entrée de texte Unicode dans des champs HTML monoligne (*Input*) et multiligne (*TextArea*). Dans PapiLex, les accès à la base sont gérés via l'interface normalisée DOM (Document Object Model).

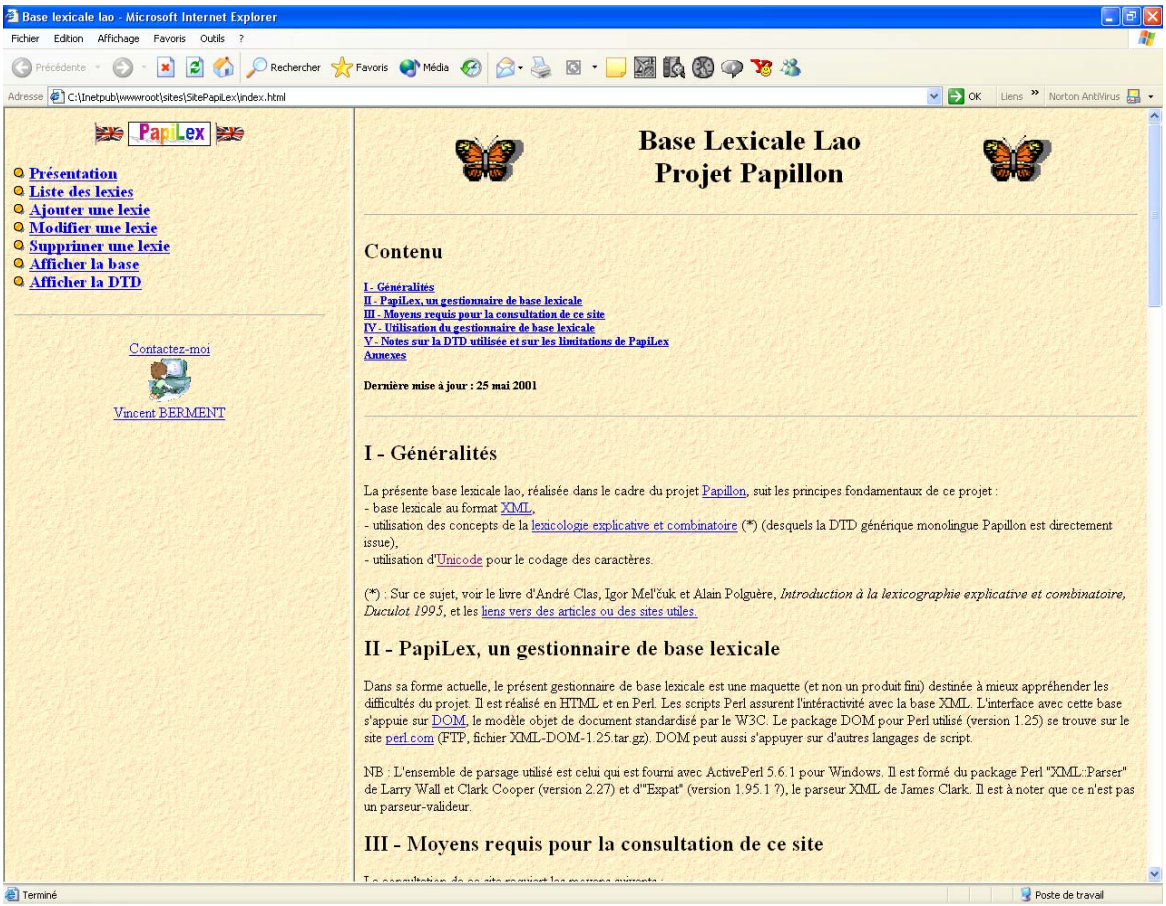


Figure 16 : Maquette de faisabilité PapiLex

[1] L'unique police Unicode incluant le lao à ce moment était la police *Arial Unicode MS* de Microsoft. Or, celle-ci donnait un rendu incorrect pour les accents ainsi que pour les voyelles suscrites et souscrites, ce qui la rendait inutilisable.

[2] Un site Papillon incluant des données en Unicode / XML et proposant une gestion de celles-ci ne verra d'ailleurs le jour qu'en juin 2002.

***Nota** : Bien que l'introduction du laotien dans le standard Unicode date de juin 1992 (version 1.01 d'Unicode, voir http://www.unicode.org/Public/UNIDATA/DerivedAge.txt), la première police Unicode contenant le laotien et fonctionnant correctement n'est donc apparue qu'en mai 2001. Encore aujourd'hui, l'utilisation pratique d'Unicode pour le laotien reste balbutiante à cause de l'absence de clavier virtuel laotien dans les systèmes d'exploitation et en particulier dans Windows, très utilisé au Laos. Les prochaines versions d'Office contiendront un tel clavier virtuel, ce qui devrait généraliser l'utilisation d'Unicode pour le laotien.*

La saisie au clavier de chaînes Unicode en laotien restait à réaliser. La technique palliative utilisée alors consistait à utiliser un éditeur de texte permettant l'insertion de texte codé en Unicode via un panneau de caractères puis à « copier-coller » de l'éditeur vers le navigateur. Pour permettre une saisie directement au clavier, nous avons développé deux logiciels. Chronologiquement, le premier fut, en août 2002, le logiciel de saisie **LaoUniKey**, un logiciel Open Source utilisable avec la plupart des logiciels compatibles Unicode sous Windows. Ses défauts étant de nécessiter une installation et de ne fonctionner que sous Windows, nous avons réalisé en avril 2003 **LaoUniWeb**, un module JavaScript intégré dans les pages Internet des sites l'utilisant. Ce module existe en deux versions, la version Unicode mentionnée précédemment et une version 8 bit, adaptée à la police Montaigne Lao utilisée dans LaoLex (voir ci-dessous) et appelée **LaoMonoWeb**. Une version 8 bit de LaoUniKey appelée **LaoMonoKey** a été réalisée en mai 2003 pour certains utilisateurs du site LaoLex.

Le site PapiLex avait prouvé la faisabilité technique d'une base de données lexicales du laotien. Sa réalisation effective a procédé à la fois du concept Montaigne d'un développement coopératif en réseau sur Internet (voir le chapitre I.2.2.3) et d'un couplage de la base avec un outil d'aide à la traduction humaine. Les fonctions d'aide à la traduction humaine et de construction de dictionnaires **LaoLex** ainsi que la base lexicale proprement dite, **LaoDict**, ont vu le jour en avril 2003. LaoDict a bénéficié du travail lexicographique demandé aux étudiants de licence de laotien à l'INALCO dans le cadre de leur projet de *traitement automatique du laotien*. La base lexicale **LaoDict**, ainsi que le **schéma XML** pour la langue laotienne décrivant le détail des articles avec les différentes parties du discours et niveaux de langue, ont été présentés au séminaire Papillon en juillet 2003. La base est en cours d'enrichissement, un contenu d'environ deux mille articles étant visé avant de procéder à un premier transfert vers la base Papillon.

Ce deuxième plan de développement, essentiellement centré sur l'aide à la traduction et les ressources lexicales, a été mis en œuvre sur une période de trois ans. Contrairement aux développements du premier plan de développement, ceux du deuxième ont recouru à plusieurs méthodes du chapitre I.2 :

- intégration à des projets d'informatisation génériques et à architecture pivot (Papillon),
- mutualisation et recours à la contribution linguistique généralisée avec LaoDict,
- recours à la diaspora et à Internet, en particulier pour LaoDict,
- réalisation de logiciels Open Source avec LaoUniKey,

Le logiciel Tallao et les techniques qu'il met en œuvre sont présentées dans [Berment 1998]. Les réalisations de la première phase postérieures à Tallao et celles de la deuxième phase sont présentées en détail dans la deuxième partie de ce document.

CONCLUSION DE LA PREMIÈRE PARTIE : DES DYNAMIQUES COMPLÉMENTAIRES

Née avec l'informatique, l'informatisation des langues a évolué, offrant de plus en plus de services pour de plus en plus de langues. C'est cependant un processus coûteux qui ne bénéficie actuellement qu'à une faible partie des langues du monde (moins de 1 %).

L'existence du standard Unicode a récemment permis la réalisation de systèmes d'exploitation et de logiciels couvrant de nombreux systèmes d'écriture tout en évitant la multiplication des incompatibilités entre plates-formes. Les langues-π — langues peu dotées informatiquement (indice σ inférieur à dix) — bénéficient ainsi d'outils d'édition performants. Par exemple, l'existence d'un mode d'écriture verticale pour le chinois ou le japonais favorise le développement de solutions pour les langues ayant le même mode d'écriture telle que les formes verticales du yi ou des langues altaïques : ouïgour, mongol, oirate, manchou, bouryate ([Daniels & Bright 1996], chapitres 19, 49 et page 439) ou encore du tai yo (voir le chapitre III.2.2.1), le codage et les principes de base étant communs aux différents systèmes d'écriture. La **dynamique Unicode** a ainsi conduit à la réalisation de logiciels performants et génériques couvrant en grande partie le premier niveau des services de traitements du texte, c'est à dire à la création de fenêtres d'édition fortement multilingues (services de saisie, de visualisation / impression, de recherche / remplacement et de sélection du texte). Il reste cependant plusieurs systèmes d'écriture non couverts par le standard (cham, môn, shan, tham laotien et siamois, lü, lanna…) ainsi que des services pris en compte incomplètement par les logiciels pour certains systèmes d'écriture contenus dans Unicode, par exemple la sélection du texte en laotien.

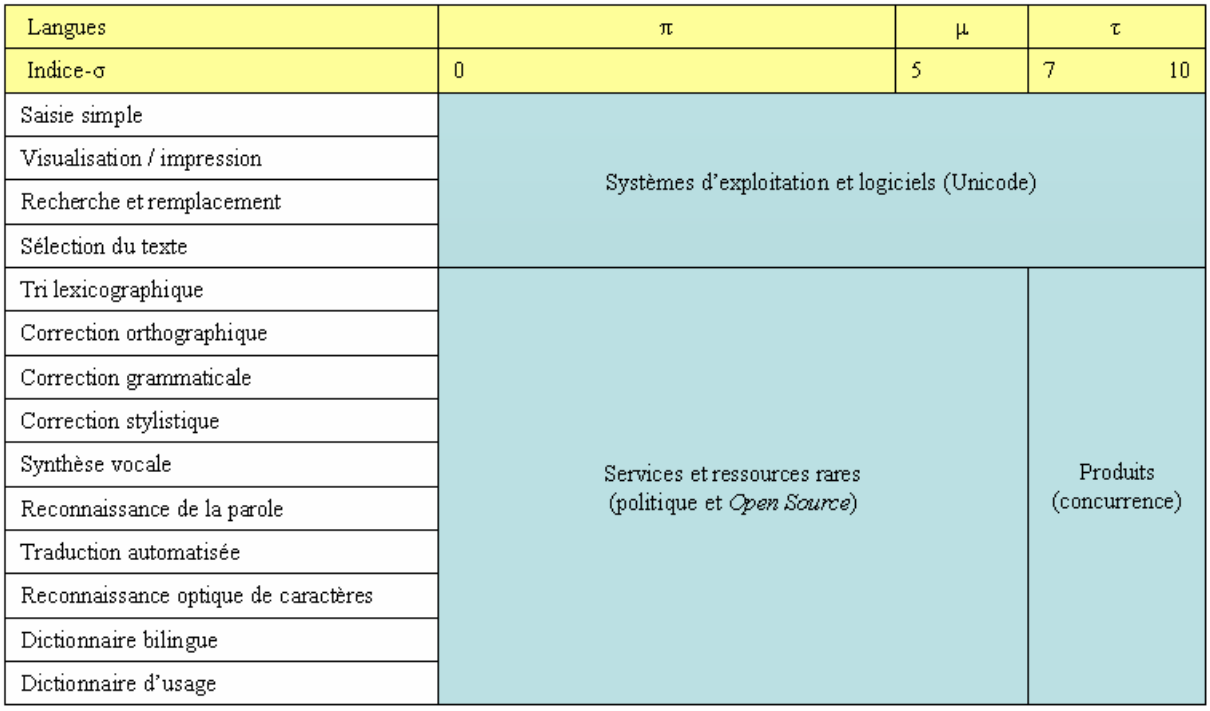


Figure 17 : État d'informatisation des ressources et services pour les langues

La **dynamique de concurrence** des produits bureautique et Internet, créée par les éditeurs privés et GNU/Linux, bénéficie à un nombre croissant de langues. Ainsi, ces éditeurs ont développé des outils linguistiques, tels que des correcteurs d'orthographe, pour quelques dizaines de langues qui en étaient dépourvues il y a encore dix ans.

Ainsi, ces deux dynamiques — Unicode et concurrence — contribuent significativement à satisfaire les besoins liés :

- ⇨ aux systèmes d'écriture (dynamique « Unicode ») : les quatre premiers services, correspondant à l'édition de texte, sont correctement couverts pour la grande majorité des langues, y compris pour les langues-π et μ,
- ⇨ aux autres aspects de l'informatisation des langues (dynamique « concurrence ») : les quelques dizaines de langues-τ bénéficient de ressources allant au-delà de l'édition de texte.

Les autres aspects de l'informatisation des langues restent rares pour les langues-π et μ, bien que celles-ci représentent plus de 99 % des langues du monde. Une **dynamique politique** émanant principalement des Nations Unies et destinée à sauvegarder le patrimoine linguistique de l'humanité voit progressivement le jour depuis la fin des années 1990. En particulier, l'UNESCO et le PNUD encouragent l'informatisation de ces langues[1] grâce à :

- ⇨ des recommandations (multilinguisme et accès universel au cyberespace),
- ⇨ des études (étude Linguapax sur les langues du monde, *Digital Opportunity Initiative*),
- ⇨ des projets (*International Open Source Network*, *Initiative B@bel*, polices de caractères).

La **dynamique « *Open Source* »** contribue aussi à doter de services et de ressources le rectangle « pauvrement » doté de la figure précédente. Cependant, il existe encore peu de tels projets pour les langues-π et μ, malgré l'encouragement des Nations Unies depuis 2001 pour cette philosophie de développement. Les locuteurs pouvant s'investir dans des projets d'informatisation sont en effet très rares, ces langues étant souvent peu parlées, et l'émulation reste limitée. Des initiatives individuelles ou universitaires viennent compléter le panorama de ces initiatives spontanées. Bien que peu coûteuses, elles sont souvent les plus efficaces et les plus utiles des contributions à l'informatisation des langues-π et μ.

En dépit de ces dynamiques politique et *Open Source*, l'informatisation des langues reste largement limitée aux langues-τ. L'informatisation d'une langue est, en effet, un projet requérant des moyens importants dont ne disposent que rarement les communautés linguistiques minoritaires. Pour réduire l'effort nécessaire, des méthodes sont possibles. Elles consistent à :

- ⇨ s'intégrer à des projets et environnements génériques, Open Source et à pivot,
- ⇨ recourir à la contribution linguistique généralisée : la mutualisation,
- ⇨ recourir aux diasporas et à Internet pour rapprocher les acteurs,
- ⇨ bénéficier de développements faits pour des langues liées,
- ⇨ recycler les dictionnaires existant,
- ⇨ informatiser par groupe de langues liées génétiquement.

Nous avons pu vérifier l'utilité des trois premières de ces méthodes d'optimisation dans le cadre de l'informatisation de la langue laotienne :

- ⇨ réalisation de traitements de texte adaptés basés sur les logiciels WordPad et Word,
- ⇨ développement d'outils d'aide à la traduction et d'une base lexicale, en partie grâce à la participation de la diaspora laotienne et à l'utilisation d'Internet.

En particulier, l'utilisation de solutions génériques et ouvertes se sont révélées extrêmement efficaces pour la réalisation de traitements de texte. Pour les deux autres méthodes expérimentées, nous avons constaté qu'il est nécessaire de susciter l'émergence d'un noyau de « personnes compétentes et minimalement intéressées ». Cela requiert d'entretenir la motivation par une animation du groupe et par des « retours » qui ne sont pas nécessairement financiers. Cette activité, essentiellement bénévole, pourrait être soutenue financièrement par des organismes comme l'UNESCO ou l'Union Européenne.

[1] Ils y participent parfois aussi directement.

DEUXIÈME PARTIE

MISE EN ŒUVRE CONCRÈTE SUR LA LANGUE LAOTIENNE

II. MISE EN ŒUVRE CONCRÈTE SUR LA LANGUE LAOTIENNE

II.1 PROBLÈMES SPÉCIFIQUES DE L'INFORMATISATION DU LAOTIEN

II.1.1 Situation de la langue et de l'écriture laotiennes

II.1.1.1 LE GROUPE DES LANGUES THAÏES

La langue laotienne est l'idiome de l'ethnie Thaï Lao, qui est l'ethnie majoritaire au Laos. La langue laotienne est parlée par environ 4 millions de personnes au Laos, auxquels on associe généralement les quelque 20 millions de laophones de la région Isan (plateau de Khorat au Nord-Est de la Thaïlande), tout à fait assimilables à des Laotiens du Laos, tant sur le plan historique et culturel que sur le plan linguistique bien qu'ils utilisent l'écriture thaïe (siamoise) et non l'écriture laotienne.

Les Thaïs Lao forment l'une des ethnies du groupe thaï. Ce dernier, relativement homogène, est réparti sur une aire géographique considérable qui va de l'Inde Orientale aux provinces de la Chine du Sud et à la Malaisie. Le découpage en ethnies ou sous-groupes proposé est le suivant[1] :

- ⇨ les Thaïs Lao ou Laotiens, dans la vallée du Mékong, au Laos et au nord-est de la Thaïlande,
- ⇨ les Thaïs Klang ou Siamois, dans la vallée de la Ménam au centre de la Thaïlande,
- ⇨ les Thaïs Yai ou Shans ou Khuns, sur le plateau shan au Myanmar (ex-Birmanie), et au sud du Yunnan,
- ⇨ les Thaïs Neua et les Thaïs Lü, en Chine méridionale, au nord de la Thaïlande et du Laos,
- ⇨ les Thaïs Yuon ou Laotiens de Chiang-Maï , au nord de la Thaïlande,
- ⇨ les Taïs, sous-groupe composite des régions montagneuses du Laos, du Tonkin et du sud de la Chine, peut-être rattachable à d'autres groupes,
  - o Thaï Dam, aux confins du Laos, du Tonkin et de l'Annam,
  - o Thaï Khao, sur les deux rives du Fleuve Rouge jusqu'à Lang-Son, et en Chine du sud,
  - o Thaï Deng, sur le Fleuve Rouge et la Rivière Noire,
  - o Tho, dans le bassin du Si-kiang et de la Rivière Claire, (Thô-Nhân signifie indigènes en vietnamien), parfois assimilés aux Thaï Khao,
  - o Nung au Tonkin (région de Cao-Bang) et au Guangxi,
  - o Nhang ou Nyang ou Giây, aux confins de la Chine et du Tonkin,
  - o Muongs du Fleuve Rouge au mont Hoang Son,
- ⇨ les Zhuang, dans les provinces méridionales de la Chine, essentiellement au Guangxi,
- ⇨ les Zhong-Jia ou Dioi ou Bu Yi, dans les provinces méridionales de la Chine, essentiellement au Yunnan, au Kouang Si et au Guizhou,
- ⇨ les Dong et les Shui, dans les provinces méridionales de la Chine, au Guizhou et au Hunan,
- ⇨ les Khamti, les Ngio et les Ahom (depuis le 18e siècle, ces derniers n'utilisent plus leur langue d'origine que dans un but religieux), en Assam et dans le nord du Myanmar,
- ⇨ les Lai ou Dai ou Li, de l'île de Hainan (peuplement ancien, influences indigènes),
- ⇨ les Vietnamiens du nord au sud du Vietnam (influences diverses).

Les langues vietnamienne et lai ont subi des influences telles que l'intercompréhension avec les autres langues du groupe thaï n'est pas possible. Pour cette raison, elles sont parfois traitées séparément du noyau homogène des langues thaïes, et souvent classées dans la catégorie des langues môn-khmères.

[1] Les sources suivantes nous ont servi pour écrire ce chapitre :

- ⇨ Henri Maspero, in Les langues du monde, A. Meillet et M. Cohen 1952, p. 571-588,
- ⇨ Henri Maspero, in « Un Empire colonial français, l'Indochine », G. Maspéro 1929, p. 63-80 (langues),
- ⇨ Teston et Percheron, « Indochine moderne » 1931,
- ⇨ S. Thierry, in Encyclopaedia Universalis 1980, Thaï p. 1031-1033,
- ⇨ Michel Malherbe, « Les langages de l'humanité », Seghers Paris 1983, p. 251-252,
- ⇨ Louis Armentier, « Orientalisme et linguistique », L'aurore Univers Montréal 1980, p. 122-134 et 162-181,
- ⇨ Louis Finot, BEFEO tome XVII n°5, Hanoï 1917,
- ⇨ Frémy et Frémy, Quid 96.

Si l'on exclut ces branches éloignées de la famille thaïe, dont le total représente un peu plus de 70 millions de personnes, l'ensemble des populations de langue thaïe est aujourd'hui de l'ordre de 100 millions de personnes. Les laophones du Laos et de la Thaïlande en représentent un peu moins du quart, les siamophones près du tiers. L'ensemble laophones-siamophones forme un groupe de près de 60 millions de personnes. Venant juste derrière en nombre, les locuteurs zhuang représentent environ un tiers du reste. Ce sont là des ordres de grandeur uniquement destinés à fixer les idées.

Figure 18 : Carte ethnolinguistique d'Asie du Sud-Est : les Thaïs

L'origine des Thaïs est encore incertaine. Il est cependant admis que ces différentes ethnies sont venues du nord (Chine), il y a au moins sept siècles, en descendant les grands fleuves (Ménam, Mékong, ...). C'est donc un peuplement relativement récent d'une terre occupée à l'époque par des populations déjà très composites (Indonésiens, Môn-Khmers, australoïdes) et souvent très civilisées.

L'ethnie Lao est arrivée dans la région qu'elle occupe actuellement au plus tard aux alentours du XIIIe siècle de notre ère. Venue du nord comme les autres Thaïs, elle a trouvé là l'empire khmer, qui faiblissait. La région occupée par l'empire khmer était dans la zone d'influence culturelle de l'Inde depuis le début du premier millénaire[1], et était ainsi pénétrée d'hindouisme et de bouddhisme. Une littérature d'origine sanskrite et pali y était présente. La langue laotienne a ainsi été influencée par les fonds khmer et indien. Plus récemment, la présence d'occidentaux (Français, Américains, Russes) a laissé une empreinte dans la langue laotienne.

En 1975, le régime en place a été renversé et une importante diaspora s'est créée en Europe, en Amérique et en Australie. En 2004, le Laos est l'un des pays les plus pauvres du monde.

[1] Georges Coedès, *Les Etats hindouisés d'Indochine et d'Indonésie*, 1964 [Coedès 1964].

II.1.1.2 LANGUE ET ÉCRITURE LAOTIENNES

La langue laotienne partage les caractéristiques des langues thaïes :

- ⇨ **tendance monosyllabique** (il est admis que les polysyllabes sont des emprunts),
- ⇨ **polytonie** (présence de plusieurs tons),
- ⇨ **polyvalence** ou **polycatégorie** (les mots ont plusieurs catégories grammaticales),
- ⇨ **invariabilité des mots** (les langues thaïes sont des langues isolantes),
- ⇨ usage de **spécificatifs** et de **classificateurs**,
- ⇨ le **déterminant** suit le **déterminé** (à l'inverse du chinois et des langues indiennes).
- ⇨ **durée pertinente** des sons vocaliques (on distingue les voyelles longues et courtes).

Le système d'écriture laotien moderne, très proche de l'alphabet thaï, est basé sur un alphabet original qui n'est utilisé qu'au Laos. Il comprend :

- ⇨ 28 signes ou graphèmes toujours simples, qui codent 27 consonnes simples (dont l'une, le « l », a deux graphies possibles : ລ et ×̲[1]) et 7 consonnes composées (dont l'une, le « 'ñ », a deux graphies possibles : ຫຍ et ຫ ງ, une autre, le « 'r », est théorique : ຫ ຣ, et deux autres, « 'n » et « 'm », peuvent être ligaturées : ໜ et ໝ). Ces lettres dérivent de l'écriture khmère,
- ⇨ 41 voyelles dont 3 de graphie archaïsante (ເ×ຽ [jaʔ], ເ×ັ ງ [jaː], et ×ັ ງ [ɔːj]), réalisées à partir de 21 graphèmes de base, parfois composés (par exemple ະ [aʔ], ແ [ɛː], et ◌ໍ ຽ [am]) et permettant de représenter jusqu'à 57 voyelles différentes si l'on prend en compte la durée et la position finale ou interconsonantique,
- ⇨ 4 accents[2] : ×́, ×̌, ×̀, ×̂,
- ⇨ 10 chiffres, de 0 à 9, le Laos ayant adopté depuis longtemps le système décimal : 0, ໑, ໒, ໓, ໔, ໕, ໖, ໗, ໘, ໙,
- ⇨ ainsi que les signes ໆ et ◌ ໆ servant à la ponctuation (les signes ◌ ◌ ◌, ◌ ◌ ◌ et ◌ ◌ ◌ ne sont plus utilisés).

Dans l'écriture laotienne, comme dans les autres écritures thaïes, il n'existe pas de distinction majuscule/minuscule pour souligner la place de la lettre dans la phrase ou l'initiale d'un nom propre. La position des voyelles (en finale de la syllabe ou en position interconsonantique) est cependant déterminante pour la forme écrite de la voyelle. L'écriture se fait de gauche à droite. Les lettres ne sont jamais attachées entre elles comme en écriture cursive occidentale.

Les **mots** sont constitués d'une (cas le plus fréquent) ou de plusieurs syllabes. Les syllabes sont généralement constituées sur le modèle (très simplifié) C[C]V[C], c'est à dire :

- ⇨ une ou deux consonnes et une voyelle (ou une diphtongue/triphtongue), ou
- ⇨ une ou deux consonnes, une voyelle (ou une diphtongue/triphtongue) et une consonne ou semi-consonne (ຍ, ວ) finale.

Un accent peut être ajouté, moyennant certaines limitations, pour préciser le ton à employer.

Toutes les consonnes existent à l'initiale des syllabes. Par contre, seules les consonnes ກ [k], ດ [d], ບ [b], ງ [ŋ], ນ [n], ມ [m], ວ [v] et ຍ [ɲ] existent en finale. Lorsqu'elles sont placées en finale de syllabes, les consonnes ກ, ງ, ນ et ມ gardent leur valeur phonologique, la consonne ດ ([d]) devient [t], ບ ([b]) devient [p], ວ ([v]) peut devenir [u] comme dans x̂ວ ([jːu]), [o] comme dans xັ ງວ ([jao]), [ɔ] comme dans xັ ວ ([aːɔ]) ou dans ແ xວ ([ɛːɔ]), et ຍ ([ɲ]) devient [j].

---

[1] Le signe x indique la position que prend la consonne prononcée juste avant la voyelle.

La lettre ◌̲ est la forme souscrite de ລ. Cette forme ne peut être employée que dans les pseudo-groupes consonantiques (voir le paragraphe consacré à ces pseudo-groupes), mais cela n'est pas systématique.

[2] On dénombre parfois cinq accents, en incluant le signe spécial ◌̄.

La description des systèmes consonantique et vocalique ci-dessous est due à Marc Reinhorn[1] et à Lamvieng Inthamone[2].

Système consonantique

Les consonnes laotiennes sont regroupées en trois séries : haute (H), moyenne (M) et basse (B). De la série à laquelle appartient la consonne dépend en partie le ton à attribuer à la syllabe. Les tableaux ci-dessous donnent la correspondance consonnes / phonèmes[3]. Les colonnes indiquent la consonne laotienne (1), une transcription phonétique basée sur l'Alphabet Phonétique International (2), une translittération réversible[4] (3) et le type de consonne : H=haute, M=moyenne ou B=basse (4).

lieu d'articulation	occlusives								aspirées								nasales							
	sourdes				sonores				normales				modulées[5]				normales				modulées			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
larynx (glotte)					ອ	ʔ	9	M	ຮ	h	h	B	ຫ	h	'h	H								
gorge (arrière du palais)	ກ	k	k	M					ຄ	k^{h6}	kh	B	ຂ	k^h	'kh	H	ງ	ŋ	ng	B	ຫງ[7]	ŋ	'ng	H
palais	ຈ	tj	c	M	ຢ	j	y	M									ຍ	ɲ	ñ	B	ຫຍ[8]	ɲ	'ñ	H
dents	ຕ	t	t	M	ດ	d	d	M	ທ	t^h	th	B	ຖ	t^h	'th	H	ນ	n	n	B	ຫນ	n	'n	H
lèvres	ປ	p	p	M	ບ	b	b	M	ຜ	p^h	ph	B	ຝ	p^h	'ph	H	ມ	m	m	B	ຫມ	m	'm	H

Figure 19 : Consonnes laotiennes à point d'articulation unique

lieu d'articulation	sourdes								sonores							
	normales				modulées				normales				modulées			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
lèvres + dents	ฝ	f	f	B	ฝ	f	'f	H	ว	v	v	B	หว	v	'v	H
langue (pointe) + alvéoles	ร	R	r	B	หร	R	'r	H								
langue + alvéoles	ล	l	l	B	หล[9]	l	'l	H								
langue/dents	ซ	s	s	B	ส	s	's	H								

Figure 20 : Consonnes laotiennes à point d'articulation double

[1] Professeur à l'Institut National des Langues et des Civilisations Orientales de 1948 à 1985, Marc Reinhorn consacra plus de quarante années de sa vie à l'étude de la langue laotienne. Parmi ses publications, citons sa grammaire de la langue laotienne [Reinhorn 1975] et son dictionnaire laotien-français [Reinhorn 1970].

[2] Enseignant à l'Institut National des Langues et des Civilisations Orientales depuis 1965 et maître de conférences depuis 1988, Lamvieng Inthamone a poursuivi certains travaux de Marc Reinhorn. Il est en particulier l'auteur d'une grammaire pratique [Inthamone à paraître].

[3] Ces phonèmes correspondent à des consonnes placées à l'initiale des syllabes, voir les changements apportés lorsqu'elles sont en finale en début du présent chapitre.

[4] Les transcriptions API ne sont pas réversibles, plusieurs consonnes ayant des prononciations identiques.

[5] Les consonnes modulées correspondent aux normales qui les précèdent. Seules les variations de la hauteur du son les différencient. En effet, dans les langues thaïes, les tons contiennent une information indispensable à la détermination du sens, comme en chinois ou dans les langues tibéto-birmanes.

[6] Les aspirées sont transcrites phonétiquement ici par x^h, conformément à la dernière version de l'API.

[7] Ces consonnes prennent la forme de « digrammes » composés d'un ຫ et d'une consonne basse. Ce sont des groupes consonantiques associatifs (GCA). Ils sont intégrés aux consonnes dans l'alphabet laotien.

[8] En graphie archaïque, ຫຍ s'écrit ຫ⁼.

[9] ຫລ peut aussi s'écrire de la manière condensée dite souscrite suivante : ຫ̳ (voir note 1 page précédente).

Système vocalique [1]

Le tableau ci-dessous donne les voyelles du laotien selon leur durée et leur position, en fin ou en milieu de syllabe (entre deux consonnes). Il donne aussi la correspondance voyelles laotiennes / phonèmes. Les colonnes indiquent la voyelle laotienne (1), une transcription phonétique basée sur l'Alphabet Phonétique International (2), une translittération (3).

COURTE						LONGUE					
FINALE			INTERCONSONANTIQUE			FINALE			INTERCONSONANTIQUE		
1	2	3	1	2	3	1	2	3	1	2	3
×ะ	aʔ	a	×ั×	a	a	×า	aː	ā	×า×	aː	ā
×ิ	iʔ	i	×ิ×	i	i	×ี	iː	ī	×ี×	iː	ī
×ึ	ɯʔ	ù	×ึ×	ɯ	ù	×ื	ɯː	ū	×ื×	ɯː	ū
×ุ	uʔ	u	×ุ×	u	u	×ู	uː	ū	×ู×	uː	ū
เ×ะ	eʔ	e	เ×ั×	e	e	เ×	eː	ē	เ××	eː	ē
แ×ะ	ɛʔ	ɛ	แ×ั×	ɛ	ɛ	แ×	ɛː	ɛ̄	แ××	ɛː	ɛ̄
โ×ะ	oʔ	o	×ົ×	o	o	โ×	oː	ō	โ××	oː	ō
เ×าะ	ɔʔ	ɔ	×อ็×	ɔ	ɔ	×ॅ	ɔː	ɔ̄	×อ×	ɔː	ɔ̄
เ×ิ	əʔ	œ	เ×ิ×	ə	œ	เ×ิ	əː	œ̄	เ×ิ×	əː	œ̄
เ×ียะ[2]	jaʔ	ia	×ีย×ั	ja	ia	เ×ีย	jaː	ia	×ีย×	jaː	ia
เ×ียะ	jaʔ	ia	×ีย×ั	ja	ia	เ×ย	jaː	ia	×ีย×	jaː	ia
เ×ือ	ɯaʔ	ùa	เ×ือ	ɯa	ùa	เ×ือ	ɯaː	ūa	เ×ือ	ɯaː	ūa
×ัวะ	waʔ	wa	×ว็×	wa	wa	×ัว	waː	wāː	×ว×	waː	wāː
ไ×	aj	ai				[3]	aːj	ai			
ใ×	aj/aə	ae				[3]	aːj	ae			
เ×า	ao	ao				[3]	aːo	ao			
×ำ	am	am				[3]	aːm	am			
[3]	ɔj	oy				×ย[2,3]	ɔːj	oy			

Figure 21 : Voyelles laotiennes

[1] Dans ce qui suit, le signe x indique la position que prend la consonne prononcée juste avant la voyelle.

[2] Cette ligne grisée (ancienne écriture) et la suivante (nouvelle écriture) contiennent deux variantes scripturales des mêmes voyelles. La forme « ancienne écriture » reste couramment rencontrée, en particulier dans les textes écrits par la diaspora laotienne. Les graphèmes constituant ces formes existent dans la nouvelle écriture.

[3] Les diphtongues longues aːj, aːo, aːm sont représentées par deux graphèmes, respectivement x า ย, x า ว et x า ม. La diphtongue aə (ใ) n'est jamais longue. La diphtongue longue ɔːj (×ย) peut aussi s'écrire ×อ ย en nouvelle écriture. Sa correspondante courte ɔj existe. Elle s'écrit (nouvelle écriture) ×อ ย.

Le signe ʔ (transcription des occlusions glottales suivant les voyelles finales brèves), est souvent supprimé dans la pratique par souci de simplification.

Notons ce que J.J. Hospitalier[1] appelle dans sa grammaire laotienne avec un certain humour *« la variété apportée par les places souvent capricieuses et toujours originales des voyelles »* [Hospitalier 1937, p. 3]. Il illustre ainsi le fait que, si du point de vue phonétique la voyelle suit la consonne initiale, il n'en est pas de même dans l'écriture où la voyelle peut apparaître :

- **après** la consonne initiale, (par exemple, ປາ=*pā*), mais aussi
- **avant** la consonne initiale (par exemple, ໄປ=*pai*),
- **au-dessus** (par exemple, ປີ=*pī*),
- **au-dessus et après** (par exemple, ຕໍ່ງ=*dam*),
- **au-dessous** (par exemple, ປູ=*pū*),
- **avant et au-dessus** (par exemple, ເປ=*pœ*), ou encore
- **avant et au-dessus et après** (par exemple, ເປົ=*pao*),

et donc, le « V » du modèle simplifié proposé plus haut (C[C]V[C]) peut à la fois ne pas être placé comme dans le modèle et être constitué de plusieurs graphèmes, qui peuvent être consécutifs ou non. Ce point sera à l'origine de difficultés, en particulier pour le tri automatique (voir II.1.2.4).

Accents

Il existe quatre accents, l'accent 1 (may ek) noté ×◌́ ou ×◌́, l'accent 2 (may tho) noté ×◌̂ ou ×◌̂, l'accent 3 (may tri) noté ×◌̌ ou ×◌̌, et l'accent 4 (may cattava) noté ×◌̀ ou ×◌̀. Nous voyons que chacun des quatre accents existe en deux hauteurs pour tenir compte de la présence de voyelles suscrites.

Les accents sont utilisés pour transcrire les tons. Cette transcription n'est pas directe ; elle dépend des accents, mais aussi de la consonne initiale et du caractère vivant ou mort de la syllabe[2]. Les tons résultants suivent approximativement le tableau suivant[3] :

consonne initiale	syllabe	sans accent	accent 1	accent 2	accent 3	accent 4
moyenne	vivante	–	ˋ	ˊ	ˊ	ˇ
	morte	˵		ˊ	ˊ	
haute	vivante	ˇ	ˋ	˵		
	morte	˵[4]				
basse	vivante	–	ˋ	ˊ		
	morte	ˊ				

Figure 22 : Accents et tons laotiens

[1] J. J. Hospitalier donna un cours libre de laotien à l'ENLOV de 1924 à 1940 (l'ENLOV — École Nationale des Langues Orientales Vivantes — est le nom qu'a porté l'INALCO entre 1914 et 1969).

[2] Une syllabe vivante est une syllabe terminée par l'une des consonnes ງ, ນ, ມ, ຍ, ວ ou par une voyelle longue.

Une syllabe morte est une syllabe terminée par l'une des consonnes ກ, ດ, ບ ou par une voyelle brève.

[3] Voir dans cette première partie, au paragraphe 2.1 « Phonologie laotienne », la page 15 « Tonèmes ».

[4] Cette notation tirée de l'API diffère de celle que Lamvieng Inthamone emploie dans *Je lis et j'écris lao*. Nous l'utilisons dans le but de nous limiter, ici, aux caractères de l'API.

Chiffres

Les chiffres s'écrivent de la manière suivante :

0	1	2	3	4	5	6	7	8	9
๐	๑	๒	๓	๔	๕	๖	๗	๘	๙

Figure 23 : Chiffres laotiens

Signes divers

Les signes de ponctuation occidentaux sont fréquemment utilisés dans les textes récents.
D'autres signes sont cependant spécifiques, principalement :

- la condamnation des fins de syllabes : ◌์ ; elle permet de conserver l'étymologie de l'écriture d'un mot et d'indiquer que la fin, ayant subi une apocope, ne doit pas être lue,
- le signe laotien équivalent à « etc. » : ๆลๆ et qui se prononce [kɔʔ laʔ] (ແກ່ຍລ້ະ),
- le signe de répétition d'une syllabe : ๆ ; il indique que la syllabe précédant le signe doit être répétée à la lecture,
- les signes, tombés en désuétude, de début de chapitre ou d'article ๏, de fin de paragraphe ฯ๎, et de fin de chapitre ou de livre ๛.

Ordre alphabétique

L'ordre alphabétique laotien distingue les consonnes et les voyelles. L'ordre alphabétique des consonnes simples est présenté ci-dessous.

ก ข ฅ ฆ จ ส ช ย ด ต ถ ท ม บ ป ผ ฝ พ ฟ ม ย ร ฺ ล ว ห อ ฮ.

Le système vocalique est généralement ordonné de la manière suivante (nouvelle écriture) [1, 2] :

◌ะ ◌ั ◌า ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู เ◌ะ เ◌ั แ◌ โ◌ะ ◌็ โ◌

◌าะ ◌ั ◌ำ ◌อ ◌ิ ◌ี ◌ัย ◌ั ◌ย ◌ัย ◌ิอ ◌ีอ ◌็อะ ◌ัว ◌็ว ◌อ

ไ◌ ใ◌ ◌ิ◌า ◌ำา

[1] Voir Lamvieng Inthamone *Je lis et j'écris lao* et Marc Reinhorn *Grammaire de la langue lao*.
[2] L'ordre alphabétique des voyelles présente quelques différences suivant les dictionnaires.

II.1.2 Difficultés dues à l'écriture et aux polices laotiennes

II.1.2.1 POLICES DE CARACTÈRES ET ABSENCE DE STANDARDISATION

Dès les années 1980 et surtout après l'apparition des polices TrueType en 1991 (MacOS 7) et 1992 (Windows 3.1), des informaticiens ont développé des polices de caractères laotiens. Pour cela, ils ont largement utilisé le logiciel Fontographer de Macromedia grâce auquel ils ont dessiné des caractères laotiens en lieu et place des caractères latins, la correspondance s'établissant selon la place souhaitée sur le clavier pour les caractères. Par exemple, certains ont voulu conserver approximativement la disposition des claviers de machines à écrire laotiennes et ont ainsi placé la lettre ກ, première lettre de l'alphabet laotien, sur la lettre d du clavier et ont donc dessiné ce caractère dans la case de code 0x64, et ainsi de suite. Ces polices ont été très employées, en particulier dans le traitement de texte Word de Microsoft, très répandu au Laos où Windows est pratiquement le seul système d'exploitation utilisé.

Pendant les cinq années qui suivirent l'apparition des premières polices TrueType, chacun a réalisé des polices avec une disposition clavier particulière, créant une confusion qui existe encore aujourd'hui. En 1997 sont arrivés Keyman, un logiciel permettant de réaliser assez facilement des claviers virtuels sous Windows, et les premières polices incorporant à la fois les caractères ASCII de l'anglais (codes 0x20 à 0x7F) et les caractères laotiens (codes 0x80 à 0xFF). Cela n'a pas suffi à établir un standard de fait.

Les polices TrueType laotiennes sont aujourd'hui assez nombreuses. Bien que l'on assiste à l'émergence de l'utilisation d'Unicode pour le laotien, les anciennes polices 8 bits non normalisées sont encore largement utilisées. Quelques polices sont listées ci-dessous, par groupes plus ou moins compatibles.

Type Alice 92

Apparues en 1992, ces premières polices TrueType n'offrent pas de caractères latins. Parmi elles, les polices appelées « Hongkad », du nom de leur auteur, offrent des possibilités très intéressantes de modèles stylisés.

EXEMPLES : alice_0, alice_1, alice_2, alice_3, alice_4, alice_5, alice_7, Alice_new, ANUNORM, anuvong A, ATYPO1, ATYPO2, HONGKAD2, HONGKAD4, HONGKAD5, HONGKAD6, HONGKAD7, HONGKAD8, HONGKAD9, HONGKAD10, HONGKAD12, HONGKAD14, HONGKAD15, HONGKAD16, HONGKAD18, Lao, LaoBanna, LaoCaligraphy, Laolium, LaoPatin, LaoSquare, Laos Standard, Lao Light, Lao Standard { SengChanh (C) }, LaoTangdeane, Sayawath, SE-LAO_0, SE-LAO_1, SE-LAO_2, SE-LAO_3, SE-LAO_4, Src_0, Src_1, Src_2, Src_3, Src_4, Src_5, Src_6, Src_7, Src_8, Xayavath.

Ces polices ont été conçues pour utiliser la disposition clavier dite Duang Jan avec un clavier QWERTY américain.

Figure 24 : Disposition Duang Jan (sur clavier QWERTY)

Type Alice 97 (Suffixes « Lao » Et « 95 »)

Apparues en 1997, ces polices offrent des caractères ASCII (mais pas de lettre accentuée, ni de « ç », par exemple). Leur structure s'inspire de celle des polices thaïes respectant la norme TIS 620-2533 du *Thai Industrial Standards Institute*. Ces polices permettent de saisir du texte en anglais ou en laotien avec ces polices de type Alice 97.

EXEMPLES : Alice0 Lao, Alice1 Lao, Alice2 Lao, Alice3 Lao, Alice4 Lao, Alice5 Lao, Bubble Lao, Chantabouli Lao, Hollow Lao, Saysettha Lao, Alice0 95, Alice1 95, Alice2 95, Alice3 95, Alice4 95, Alice5 95, Bubble 95, Chantabouli 95, Hollow 95, Saysettha 95.

Type Lao France

Cette police est une adaptation pour Windows des polices pour Macintosh mises au point à l'Institut National des Langues et des Civilisations Orientales par Lamvieng Inthamone. Elle permet de saisir du texte en laotien à partir du clavier intuitif Lao France, sans avoir recours à un logiciel particulier, et a servi à définir les dispositions claviers Lao France et Lao France New.

EXEMPLE : LaoFrance.

Figure 25 : Disposition Lao France (sur clavier AZERTY)

Type Unicode

Des polices respectant le standard Unicode et contenant les lettres laotiennes sont disponibles depuis le printemps 2001. Ces polices sont basées sur un jeu de caractères Unicode (les autres polices utilisent un codage des caractères sur 8 bits), le laotien disposant de la plage de 0E80 à 0EFF[1].

EXEMPLES : Saysettha OT, Saysettha Unicode, JG Basic Lao, Times New Roman Lao.

Notons que l'existence de plusieurs dispositions clavier ajoute aux difficultés dues à la multiplicité des encodages de polices. En effet, certains caractères ne sont pas disponibles dans toutes les dispositions de claviers (par exemple la ligature ໜ n'est pas directement accessible à partir du clavier *Lao France*).

Ainsi, les polices laotiennes sont généralement incompatibles, que ce soit parce qu'elles font partie de familles de polices complètement différentes, ou encore parce qu'elles présentent de petits écarts à l'intérieur d'une même famille.

[1] Voir http://www.unicode.org/.

Par exemple, demandons à Word de changer le texte suivant saisi en Saysettha Lao (famille Alice 97) dans la police Src_4 (famille Alice 92) :

ເລື່ອງຕາບອດຄຳຊ້າງ.

Cela donnera le résultat suivant (petits carrés) :

□□□□□□□□□□□□□□□□□.

La multiplicité des encodages existants et plus généralement l'absence de norme sont donc une première difficulté pratique pour le traitement de textes laotiens, les textes étant associés à des polices particulières. Aujourd'hui encore, lorsqu'un courrier électronique en laotien est envoyé, l'auteur indique quelle police utiliser et joint parfois la police à son envoi.

Viennent ensuite des problèmes plus intrinsèquement liés à l'écriture laotienne ou, dans certains cas, à la méthode de saisie utilisée.

II.1.2.2 SAISIE NON UNIVOQUE ET FORME CANONIQUE

L'existence de caractères sans avance[1] qui se superposent au caractère précédent[2] fait qu'il est possible de taper 1 fois, 10 fois comme 100 fois la voyelle ື, par exemple, sans que le résultat diffère à l'écran. Ainsi, si l'on cherche dans un dictionnaire électronique un mot saisi fautivement avec deux ື au lieu d'un seul, on ne le trouvera pas. De même, les mots tels que ເລື່ອງ, qui contiennent une voyelle sans avance et un accent, pourront s'écrire de deux manières selon que la voyelle sera frappée avant ou après l'accent. Les autres cas de saisie multiforme, décrits dans [Berment 1998], posent les mêmes problèmes d'unicité de la représentation.

La figure ci-dessous montre sur deux exemples qu'il peut y avoir plusieurs décompositions possibles pour un mot (chaque case correspond à un caractère saisi au clavier).

Mot	ເລື່ອງ							ແລ້ວ					
Forme canonique	ເ	ລ	່	ື	ອ	ງ		ແ	ລ	້	ວ		
Forme saisie	ເ	ລ	ື	່	່	ອ	ງ	ເ	ເ	ລ	້	ວ	

Figure 26 : Exemples de mots pouvant être saisis de plusieurs manières

Notons que cette relation non biunivoque entre le mot et sa représentation n'est pas possible en français, où des frappes successives produisent systématiquement autant de caractères distincts qui viennent se placer à la suite les uns des autres, sur une ligne. En laotien, les lettres qui doivent être suscrites ou souscrites sont, par nature, sans avance.

[1] Nous appelons « caractère sans avance » tout caractère dont la saisie ne provoque pas le déplacement du curseur d'édition. Lorsqu'il s'agit d'une espace, nous utiliserons la terminologie Unicode qui parle « d'espace sans largeur » (caractère *ZERO WIDTH SPACE*, de code 200B).
[2] Un peu comme le permettaient les machines à écrire françaises pour l'accent circonflexe ou le tréma dont la frappe, effectuée avant celle de la voyelle, ne faisait pas avancer le chariot.

II.1.2.3 ABSENCE D'ESPACES ENTRE MOTS

L'écriture du laotien ne recourt pas aux espaces pour séparer les mots entre eux. En français, un logiciel tel que Word coupe les lignes entre les mots (aux endroits où il y a des espaces) ou sur les traits d'union (code $0x2D_{16}$ du Latin 1). Il peut encore, selon l'option de césure choisie, couper les mots et insérer un trait d'union en suivant les règles de la typographie française.

En laotien, les traitements de texte ne couperont pas correctement les fins de ligne. Ils couperont là où de rares espaces seront présentes, ou en bout de ligne en l'absence d'espace, ou encore sur les caractères dont le caractère ASCII correspondant sera un trait d'union[1], entraînant une mise en page incorrecte et une lecture difficile. Par exemple, sur les 2e, 3e, 4e, et 5e lignes de l'extrait ci-dessous, la coupure de fin de ligne est incorrecte (absence d'espace).

ແຕ່ກ່ອນຍັງມີພໍ່ນາຜູ້ນຶ່ງໄປນາທຸກມື້ເພື່ອບົວຣະບັດເຂົ້ານາ.ລາວສັງເກດເຫັນຮູໃນນາໃຮ່ນຶ່ງຈຶ່ງຖົມມັນ.ມື້ຕໍ່ມາລາວ
ຄືນໄປເລາະລຽບນານໍ້ຍັງເຫັນຮູຢູ່ຄືເກົ່າຈຶ່ງປ້ານຄັນນາຄືນ.ເປັນດັ່ງນີ້ຢູ່ຫຼາຍເທື່ອລາວຈຶ່ງຕັດສິນໃຈຂຸດຕາມຮູນັ້ນເບິ່
ງເລີຍພົບອ່ຽນດ່ອນໂຕນຶ່ງ;ລາວເງືອດພ້າຈະຟັນມັນແຕ່ພໍດີໄດ້ຍິນສຽງອ້ອນວອນວ່າ:ຢ່າຂ້າຂະນ້ອຍທ້ອນ!ຂະນ້ອ
ຍຂໍອີນຊີຊີວິດໄວ້.ເມື່ອໄດ້ຍິນດັ່ງນັ້ນພໍ່ນານໍາສົງສານຈຶ່ງເອົາມັນໄປລ້ຽງໄວ້.ມື້ນຶ່ງອ່ຽນດ່ອນເວົ້າຕໍ່ພໍ່ນາວ່າ:ຢູ່ເຕ່ເຮືອ
ນກໍເຫງົາໃຈ;ຢາກຂໍໄປຫາປານໍາພໍ່.ພໍ່ນາກໍເຫັນດີເອົາອ່ຽນດ່ອນໄປນໍາ;ລາວໄດ້ເອົາໜ້າດວງນຶ່ງໄປພ້ອມຕາມຄໍາສັ່
ງຂອງມັນ.

Cette absence d'espace entre mots cause encore trois inconvénients :

- la sélection des mots est difficile ; les sélections réalisées à la souris renvoyant les chaînes de caractères alphanumériques contigus, cela donne un résultat aléatoire en laotien, parfois tout un paragraphe,
- la fonction de correction orthographique ne peut pas se faire simplement à partir d'une liste de mots, une segmentation en mots devant être réalisée au préalable,
- la justification à droite et à gauche des textes donne des résultats aléatoires.

II.1.2.4 PROBLÈMES LIÉS AU TRI LEXICOGRAPHIQUE

Si le classement lexicographique des mots ne rencontre pas, dans la pratique, un accord unanime de la part des lettrés laotiens, il existe cependant un quasi-consensus sur l'ordre des consonnes (seule la place des consonnes digrammes — les six dernières consonnes de la liste ci-dessous — varie d'un auteur à l'autre) :

- ກ ຂ ຄ ງ ຈ ສ ຊ ຍ ດ ຕ ຖ ທ ນ ບ ປ ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ວ ຫ ອ ຮ ຫງ ຫຽ ໜ ໝ ຫຼ ຫວ,

et sur celui des voyelles :

- ะ ັ า ິ ี ึ ื ุ ู เะ แะ เัะ เ แะ แัะ แ โะ ิะ โ เาะ ็อ ำ ็อ เิ เี เ ียะ[2] เ ือ ัย เ ีย[2] เ ย ัย เื อ เื อ ิอะ ็อ ิอ อ ใ ไ ิ า ำ ย[2].

[1] Par exemple, avec la police Laos Standard, la lettre laotienne ຊ a pour code hexadécimal $0x2D_{16}$, qui est le code du trait d'union. Word, à qui rien ne dit qu'il traite du laotien, verra un trait d'union (il ne voit que le code, à l'exclusion du dessin que forme le caractère) et coupera la ligne — si nécessaire — à cet endroit. L'utilisation d'Unicode résout ce problème.

[2] Ancienne écriture.

Il en va malheureusement tout autrement de la manière de classer les mots dans un lexique. Cela provient, ici encore, de l'absence de norme ou d'usage accepté comme en français ou en khmer. Nous pouvons cependant constater l'existence des principes généraux suivants :

- le classement est réalisé par syllabe ; ainsi, dans une suite classée, tous les mots commençant par la même syllabe se suivent, et deux mots polysyllabiques commençant par la même syllabe sont classés en fonction de leur deuxième syllabe, et ainsi de suite,
- au niveau syllabe, le classement est phonologique et non graphotaxique[1], c'est à dire qu'il dépend de la manière dont la syllabe se prononce et non de celle dont elle s'écrit (le problème de la position irrégulière des voyelles dans la syllabe est présenté au II.1.1.2).

Ce principe de classement phonologique n'est cependant pas toujours suivi. En particulier, les dictionnaires réalisés au Laos suivent, pour la plupart, un ordre établi par Maha Sila Viravong, grand lettré laotien du 20e siècle, qui consiste à considérer la priorité des voyelles comme étant du second ordre par rapport à celle de la consonne finale (en français, un tel ordre conduirait à classer *sud* avant *sas*, parce que *d* est avant *s*), ce qui donne un résultat qui nous paraît peu pratique. Une explication avancée serait la plus grande facilité pour trouver des mots pour lesquels la voyelle n'est pas connue avec certitude (durée, en particulier). Nous trouvons donc les priorités de tri suivantes[2] :

CI-V-CF-A[3] (Reinhorn), par exemple les syllabes seront dans l'ordre ກະ ກັກ ກັງ ກັ່ງ ກາ ກາກ ... ກິ ກີ,

CI-CF-V-A (Maha Sila Viravong, Soukhavong, Kerr), par exemple l'ordre ກະ ກາ ກິ ກີ ກຸ ກູ ... ກັງ ກັ່ງ.

Au niveau des groupes consonantiques[4], il n'y a pas unanimité non plus, certains auteurs de lexique reportant en début de liste les mots commençant par un groupe consonantique, d'autres en fin de liste, ou encore un panachage de ces possibilités, quand ce n'est pas moins rigoureux encore. D'une manière générale, nous n'avons pas trouvé de règle écrite concernant la mise en œuvre du tri lexicographique.

Il existe aussi des **divergences morphologiques dans l'orthographe des mots**, entre l'orthographe officielle de la République Démocratique Populaire du Laos (RDPL, Nouveau Régime) et celle du Comité Littéraire Lao (Ancien Régime, comité créé en 1950). Ces deux orthographes se réfèrent à l'ordonnance royale n°10 du 27 janvier 1949, qui stipule « la règle de l'écriture phonétique ». Les divergences concernent l'orthographe des groupes consonantiques englobant et guidant et celle des consonnes finales dans les polysyllabes constituées d'au moins trois syllabes[5], l'écriture en vigueur depuis 1975, se réclamant d'une écriture phonétique *stricto sensu*, et l'écriture antérieure à 1975 conservant une part d'étymologie dans l'écriture.

---

[1] Notons que le tri du laotien est, de ce fait, plus complexe à réaliser que le tri du thaï qui est davantage lié à la forme écrite (voir l'algorithme de tri du thaï à l'adresse http://www.linux.thai.net/thep/tsort.html). Ainsi, en thaï, *phanuak* (ผนวก, ajouter, combiner) sera placé après *phong* (ผง, poussière) parce que น (*n*) est après ง (*ng*) dans l'alphabet thaï, bien que, phonologiquement, *pha* soit avant *phong*.

[2] **Notations** : CI pour Consonne Initiale, V pour Voyelle, A pour Accent, et CF pour consonne finale.

[3] Cette notation signifie que pour une consonne initiale donnée (ก dans l'exemple), on prendra, dans l'ordre, toutes les voyelles (ะ,ั, า, ...), puis toutes les consonnes finales (en commençant par l'absence de consonne finale), puis tous les accents.

[4] Nous avons vu les groupes consonantiques constitués d'un ຫ et d'une consonne qui sont appelés digrammes ou groupes associatifs. Il existe deux autres types de groupes consonantiques : les groupes englobants constitués d'une consonne simple suivie ວ, ຣ ou ລ (les deux derniers cas n'existent plus dans la nouvelle écriture) et les groupes guidants constitués de deux consonnes, la première se prononçant avec un 'a' très bref (n'existent plus dans la nouvelle écriture). Voir [Inthamone1987], p. 158 à 167.

[5] Voir Lamvieng Inthamone, *Je lis et j'écris lao*, p. 158 à 167 ([Inthamone 1987]).

II.1.2.5 SYNTHÈSE DES DIFFICULTÉS LIÉES À L'ÉCRITURE ET AUX POLICES LAOTIENNES

En résumé, les difficultés liées à l'écriture et aux polices laotiennes sont :

- absence d'espace entre mots (II.1.2.3),
- saisie non univoque (II.1.2.2),
- ordre alphabétique et tri (II.1.2.4),
- absence de normalisation pour le clavier (II.1.2.1),
- absence de normalisation pour le codage des caractères laotiens (II.1.2.1).

Les conséquences pratiques sont résumées dans le tableau ci-dessous.

	Saisie	Césure (segmentation)	Sélection (segmentation)	Recherche de texte	Tri lexicographique	Réalisation de corpus	Correction d'orthographe	Changement de police
Absence d'espace entre mots	x	x	x		x		x	
Saisie non univoque				x	x	x	x	
Ordre lexicographique et tri					x[1]		x	
Normalisation d'une disposition clavier	x							
Normalisation d'un codage des caractères laotiens	x			x	x		x	x

Figure 27 : Conséquences pratiques des difficultés dues à l'écriture et aux polices laotiennes

[1] En particulier pour la mise en ordre des dictionnaires.

II.2 PREMIÈRE PHASE : TRAITEMENT DU TEXTE

II.2.1 Claviers virtuels

II.2.1.1 CLASSIFICATION DES CLAVIERS VIRTUELS

Le terme de « clavier virtuel » est généralement utilisé pour désigner un clavier dessiné à l'écran et permettant de saisir du texte grâce à des clics souris[1]. Nous l'emploierons dans un sens plus général qui désignera tout logiciel permettant la saisie de texte dans un certain système d'écriture à partir d'un clavier non prévu pour cela. Formellement, c'est donc une correspondance (tableau) entre les touches d'un clavier et des chaînes de caractères.

Les claviers virtuels réalisés peuvent être classés en fonction de leur portée :

- des claviers dont la portée est limitée à un logiciel (Tallao, LaoPad, LaoWord) ou un type de logiciel (LaoMonoWeb, LaoUniWeb),
- des claviers dont la portée s'étend à tous les logiciels (LaoMonoKey, LaoUniKey),

des systèmes d'exploitation dans lesquels ils fonctionnent :

- Windows (Tallao, LaoPad, LaoWord, LaoMonoKey, LaoUniKey),
- Tous systèmes d'exploitation (LaoMonoWeb, LaoUniWeb),

de leur mode de fonctionnement :

- en local (Tallao, LaoPad, LaoWord, LaoMonoKey, LaoUniKey),
- en réseau (LaoMonoWeb, LaoUniWeb).

Sous Windows[2], les applications reçoivent les informations de leur environnement, en particulier celles venant du clavier, sous forme de messages appelés *messages Windows*. Elles ont ainsi une boucle, appelée boucle de messages, qui tourne en permanence en attendant ces messages Windows. Lors de l'appui sur une touche du clavier (ou lors d'un mouvement de la souris...), le système d'exploitation génère un message incorporant le caractère frappé (ou la nouvelle position de la souris...) que la boucle pourra alors interpréter.

Les technologies de clavier virtuel différeront donc aussi selon l'API utilisée, qui diffère selon le logiciel considéré :

- Tallao, qui s'appuie sur l'interface de programmation SDK (*Software Development Kit*) de Windows, a directement accès à sa boucle de messages,
- LaoPad, qui s'appuie sur les classes MFC (*Microsoft Foundation Classes*), a accès à sa boucle de messages à travers les méthodes de la classe *CRichEditCtrl*,
- LaoWord n'a pas accès à la boucle de messages Windows de Word et doit faire appel à la technologie dite des *hooks*,
- LaoMonoKey et LaoUniKey, qui n'ont pas accès aux boucles de messages des applications, font aussi appel à la technologie des *hooks*,
- LaoMonoWeb et LaoUniWeb ont accès à la boucle de messages du navigateur utilisé à travers son interpréteur JavaScript.

[1] Voir par exemple les sites http://perso.wanadoo.fr/michel.staelens/clavier/index.htm, http://www.thailande-guide.com/fr/langue_clavier.php.

[2] Rappelons que l'objectif de ce premier plan de développement est la réalisation d'un traitement de texte sous Windows adapté au laotien, Windows étant pratiquement le seul système d'exploitation utilisé au Laos.

II.2.1.2 LES DIFFÉRENTES TECHNOLOGIES DE CLAVIERS VIRTUELS UTILISÉES

II.2.1.2.1 Technologie utilisant l'interface de programmation SDK de Windows

Dans **Tallao**, c'est la boucle de messages de l'application qui traite directement les messages caractère, comme cela est décrit dans le code ci-dessous.

```
while (GetMessage(&msg, NULL, 0, 0)) {
  ...
  if ((msg.message==WM_CHAR)&&(GetFocus()==hEdit)) {
    if (0==lstrcmp(Clavier,"Clavier Lao France")) {
      code_pivot=clavier_LF_donne_code_pivot[(unsigned
          int)(BYTE)msg.wParam] ;
      msg.wParam=code_pivot_donne_code_police[(unsigned
          int)code_pivot] ;
      SendMessage (hEdit,WM_CHAR,msg.wParam,msg.lParam);  } }
  ... }
```

Explication : À chaque fois qu'un caractère est frappé au clavier, un message de type « caractère » (*WM_CHAR*) est envoyé à l'application. Celle-ci remplace le code initialement contenu dans le paramètre *msg.wParam* par un autre code, en fonction de la disposition choisie pour le clavier et de la police de caractères utilisée. Cela se fait donc en deux étapes : 1) calcul d'un code caractère abstrait (code_pivot) grâce au tableau « clavier » : clavier_LF_donne_code_pivot, 2) calcul du code caractère réel lui correspondant dans la police utilisée grâce au tableau « police » : code_pivot_donne_code_police (voir la Figure 31 page 96). Le code obtenu est renvoyé à la fenêtre fille (*hEdit*) de l'application via l'API *SendMessage*. La fenêtre fille est une classe EDIT, fenêtre d'édition simple ne permettant d'afficher qu'une police à la fois (voir le chapitre I.3.2).

II.2.1.2.2 Technologie utilisant l'interface de programmation MFC de Windows

Dans **LaoPad**, les messages clavier sont accessibles via le membre *OnChar* de la classe *CWordPadView*, qui hérite de la classe *CRichEditView*, classe « vue » du contrôle d'édition RTF *CRichEditCtrl*. Nous avons surchargé ce membre pour qu'il modifie le caractère reçu en fonction de la disposition clavier souhaitée.

```
void CWordPadView::OnChar(UINT nChar, UINT nRepCnt, UINT nFlags) {
  ...
  if (0==lstrcmp(Clavier,"Clavier Lao France"))
    code_pivot=clavier_LF_donne_code_pivot[(unsigned
          int)(BYTE)msg.wParam] ;
  if (0==lstrcmp(Clavier,"Clavier Duang Jan"))
    code_pivot=clavier_DC_donne_code_pivot[(unsigned
          int)(BYTE)msg.wParam] ;
  nChar=(UINT)theApp.code_pivot_donne_code_police[(unsigned
        int)code_pivot];
  SendMessage(WM_CHAR,nChar,0);
  ... }
```

Explication : Comme le caractère modifié est envoyé par *SendMessage* au même membre *OnChar*, un test en début de fonction (non présenté ci-dessus) permet de n'exécuter ce code qu'une fois sur deux et d'éviter le bouclage.

II.2.1.2.3 Technologie des *hooks*

Une technique spécifique appelée *hook* peut être utilisée pour intercepter les messages et pour les modifier avant qu'ils ne soient traités par les applications.

La fonction *SetWindowsHookEx* contenue dans Visual C++ — l'environnement de développement de Microsoft — permet d'installer une « procédure de *hook* ». Ces procédures de *hook* sont des fonctions dites *callback*[1] qui contiennent un code modifiant les messages Windows. Elles sont généralement placées dans une librairie dynamique (DLL) et chaînées entre elles lorsque plusieurs sont installées. Dès qu'un *hook* est installé, tous les messages sont traités par la procédure vers laquelle il pointe. L'un des paramètres de la fonction *SetWindowsHookEx* détermine si la portée du *hook* s'étend à toutes les fenêtres ou à une seule.

La procédure de *hook* reçoit tous les paramètres contenus dans le message. Ils incluent le type de message (dans notre cas, il s'agit du message « caractère » *WM_CHAR*), le caractère lui-même ainsi que plusieurs informations associées. Cependant, comme le fait remarquer Wright [Wright et al. 2001][2], Unicode est encore peu intégré aux versions 95 et 98 de Windows[3] (versions encore largement utilisées au Laos). Pour contourner ces limitations, nous avons développé une version spécifique pour Microsoft Word du logiciel de saisie en Unicode pour le laotien. Cette version s'appuie sur la capacité COM (Component Object Model) de Word qui lui permet d'être accédé à partir de processus externes[4].

Nous avons aussi développé une version plus générale utilisant l'astuce proposée par Wright, basée sur un principe de « copier-coller » et qui fonctionne dès Windows 95 dans tout éditeur acceptant le raccourci de collage *Ctrl V*. Ce logiciel, appelé **LaoUniKey**, est un logiciel **Open Source** disponible en ligne sous **licence GPL**[5] (voir www.laosoftware.com/luk.zip). Il peut être utilisé pour saisir du texte en laotien en conformité avec le standard Unicode et peut être adapté à d'autres systèmes d'écriture inclus dans Unicode mais manquant d'un clavier virtuel. LaoUniKey a été développé pour permettre la saisie du laotien en Unicode. Il intègre le *hook* et le procédé de « copier-coller » présenté précédemment mais pas d'optimisation des glyphes ni de vérification de la cohérence de la saisie. Le logiciel est constitué d'un exécutable (LaoUniKey.exe) et d'une librairie dynamique (LaoUniKey.dll).

La tâche de l'exécutable est principalement :

- de fournir une interface de dialogue pour :
  - activer et désactiver la saisie en laotien,
  - choisir le clavier,
- de charger la librairie dynamique grâce à l'API *LoadLibrary*,
- d'installer le *hook*.

La librairie dynamique contient :

- le tableau de correspondance entre les claviers physique et virtuel,
- la procédure de *hook*,
- la gestion du « copier-coller ».

[1] Les fonctions *callback* sont des fonctions appelées par Windows.
[2] Leur article est disponible sur le site http://www.punchdown.org/rvb/papers/EriPaper3C.html.
[3] Ils suggèrent que la meilleure solution est de copier par programme le caractère Unicode dans le presse-papier et de simuler le raccourci standard de collage — *Ctrl V* — pour copier le caractère dans l'éditeur utilisé.
[4] COM permet à des objets (comme des documents Word 97, XP et 2000) d'offrir leurs fonctionnalités à d'autres objets ou à des applications externes.
[5] GPL : GNU General Public Licence. Voir http://www.gnu.org/licenses/licenses.html et l'annexe A.13.

II.2.1.2.4 Technologie utilisant JavaScript

Dans les navigateurs récents, un texte Unicode peut être saisi directement dans des champs HTML monoligne (*input*) et multiligne (*textarea*). Pour de nombreuses langues, les systèmes d'exploitation gèrent la saisie de texte en Unicode. Pour d'autres cependant, un tel service de saisie n'est pas fourni. Ce chapitre présente une technique pour générer un outil de saisie en Unicode pour ces dernières[1,2].

Depuis sa version 1.2, JavaScript offre un service de gestion d'événements[3]. Ce service incluant la gestion des événements clavier, un « *script* » peut être utilisé pour modifier les caractères saisis avant qu'ils ne soient délivrés au navigateur. Par exemple, pour transformer un clavier qwerty en clavier azerty, des transformations sur les caractères modifieront, entre autres, les lettres q et w respectivement en a et z.

Malheureusement, la syntaxe de gestion d'événements dépend de la famille du navigateur et de sa version. Par exemple, dans Internet Explorer, le nom de la fonction de substitution à appeler lors d'un événement clavier est directement copié dans le membre *onkeypress* de l'objet *document* (principe de « bouillonnement ») alors que dans Netscape, l'événement doit d'abord être « capturé » par une méthode telle que *captureEvents* (Netscape 4.0) ou *addEventListener* (Netscape 6.0)[4]. Ensuite, l'événement doit être obtenu, soit de la méthode *event* de l'objet *document* (Internet Explorer), soit directement sous la forme du paramètre *evt* dans la fonction JavaScript qui fait la substitution (Netscape). Le caractère frappé lui-même, bien qu'étant un membre de l'événement dans les deux cas, a deux noms différents : *which* dans Netscape 4.0 et *keyCode* dans Internet Explorer et Netscape 6.0.

Ces différences entre navigateurs conduisent à élaborer un script d'interception des caractères qui tient compte des navigateurs. La syntaxe suivante a été implémentée.

```
if ( document.addEventListener ) {  // Netscape 6.0
   document.addEventListener
   ( "keypress", handlePress, true ) ; }
else {   // Internet Explorer ou Netscape 4.0
   if ( document.captureEvents ) // Netscape 4.0
         document.captureEvents
               ( Event.KEYPRESS )  ;
   document.onkeypress = handlePress ; }

var whichASC = "" ;
evt = ( evt ) ? evt : ( document.event ) ?
   document.event : ""
if ( evt.which ) { whichASC = evt.which }
else { whichASC = event.keyCode }
var whichKey = String.fromCharCode (whichASC) ;
```

[1] L'affichage des caractères peut aussi avoir recours à une applet Java comme dans le site de traduction thaï-japonais / japonais-thaï Saikam (http://saikam.nii.ac.jp/).
[2] Cette description ne tient pas compte des règles complexes qui sont nécessaires pour certains systèmes d'écriture comme l'arabe ou les écritures indiennes.
[3] Une très bonne description de JavaScript est disponible dans [Goodman 2001].
[4] Netscape 6.0 inclut aussi le principe de « bouillonnement » de IE4+.

Ainsi, si l'on appelle *whichKey* le caractère frappé, le nouveau contenu du champ de saisie sera obtenu par concaténation de son contenu courant avec le caractère frappé et transformé. Dans le code proposé ci-dessous, le caractère *d* est ajouté au champ *field* suite à la frappe de la touche *k*. Ce caractère correspond à la lettre laotienne ດ dans la police Montaigne Lao.

```
switch ( whichKey ) {
   case "k" : field =  field + "d" ;
   return false ; break;
   ...
   default:
   return true ; break ; }
```

Notons que le programme JavaScript a besoin de savoir quel champ de saisie il doit modifier lorsque plusieurs champs de saisie sont présents dans un formulaire. Pour cela, nous avons adopté une solution dans laquelle les champs *input* et *textarea* appellent leurs gestionnaires d'événements *onBlur* et *onFocus* de façon à mettre à jour une variable contenant l'identité du champ en y entrant (*onFocus*) et en le quittant (*onBlur*).

```
<input   onBlur = PerteFocus ()
         onFocus = GainFocus ( 'Lexie' )
         name = Lexie>
```

Ainsi, le champ peut être accédé via *document.forms*[*FormName*].*elements*[*InputText*].*value;* où *FormName* est le nom du formulaire et *InputText* celui du champ. Ces paramètres doivent être mis à jour par des fonctions (*PerteFocus* et *GainFocus* dans notre exemple) sur la base du paramètre reçu (*'Lexie'* dans notre exemple).

II.2.1.2.5 Technologie utilisant Microsoft Keyboard Layout Creator

Notons la récente introduction par Microsoft (juin 2003) de l'outil de création de claviers virtuels *Microsoft Keyboard Layout Creator*[1]. Plusieurs articles de Microsoft relatifs à la réalisation de claviers virtuels compatibles avec Unicode ont été présentés à la 23e conférence internationale Unicode (24-26 mars 2003)[2].

II.2.2 Sélections à la souris et au clavier

En français ou dans une autre langue à écriture segmentée, un double-clic souris sélectionnera généralement un mot. En laotien, aucun des objets d'édition que nous avons utilisés (classes EDIT pour Tallao, CRichEditView pour LaoPad et OpusWwd/_WwG pour Word) n'offre de sélection correcte du texte pour le laotien. Un double-clic sélectionnera donc une zone de texte inappropriée. De même, dans beaucoup de classes d'édition, la sélection d'une zone de texte par « clic » puis mouvement de la souris est prévue pour sélectionner des mots entiers et a un comportement peu pratique en laotien. En particulier, lorsque la zone que l'on souhaite sélectionner se termine par une voyelle suscrite ou souscrite ou encore par un accent, il est impossible de savoir si cette dernière lettre est incluse ou non dans la sélection sans recourir au clavier qui permet une sélection lettre à lettre (par exemple dans ມີ).

[1] Voir la page Internet http://microsoft.com/downloads/details.aspx?FamilyId=FB7B3DCD-D4C1-4943-9C74-D8DF57EF19D7&displaylang=en. Cet outil est exclusivement destiné à Windows (Windows 2000, Windows XP et Windows Server 2003).

[2] Voir les pages http://www.microsoft.com/globaldev/reference/presentations/23rd_Unicode_Conf.mspx et http://www.unicode.org/iuc/iuc23/.

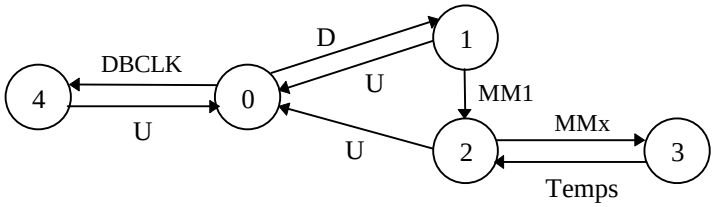
Afin de rendre les sélections plus pratiques, nous avons développé une logique adaptée au laotien :

- sur **double-clic**, une syllabe est sélectionnée.
- sur **mouvement souris** lorsque le clic gauche est enfoncé, une sélection cumulative par syllabe est réalisée, à l'image de ce qui est réalisé dans Word pour la sélection des mots.

Cela a été mis en œuvre sous Windows dans les trois logiciels Tallao, LaoPad et LaoWord.

Comme nous l'avons rappelé au chapitre II.2.1.1, les applications Windows reçoivent les informations de leur environnement sous forme de messages. Le principe donc est d'intercepter les messages souris[1] et de réaliser cette logique de sélection.

Pour cela, nous gérons l'automate ci-dessous.



0 : état initial (Inits.cpp) et après réception d'un WM_LBUTTONUP (U)
1 : après réception d'un WM_LBUTTONDOWN (D)
2 : après réception d'un premier WM_MOUSEMOVE (MM1) et après calcul de la syllabe (Temps)
3 : après réception des WM_MOUSEMOVE suivants (MMx)
4 : après réception d'un WM_LBUTTONDBLCLK (DBCLK)

Figure 28 : Automate des événements souris

Lorsque l'application reçoit un message souris provoquant l'une des transitions 0-4, 1-2 ou 2-3, elle calcule et affiche la nouvelle sélection (la transition 1-2 sert aussi à initialiser des variables de la sélection).

Dans **Tallao**, c'est la fenêtre mère qui traite directement les messages souris et renvoie la nouvelle sélection à afficher vers la fenêtre fille active[2] sous la forme d'un message Windows *EM_SETSEL* accompagné des indices de début et de fin de la nouvelle sélection. Le message Windows *EM_GETSEL* lui permet auparavant d'acquérir l'indice du caractère sous la souris et l'API GetWindowText le texte contenu dans la fenêtre fille.

Dans **LaoPad**, les messages souris sont accessibles via les membres *OnLButtonDown*, *OnLButtonUp*, *OnLButtonDblClk* et *OnMouseMove* de la classe *CWordPadView*, qui hérite de la classe d'édition RTF *CRichEditView*. Nous avons surchargé ces membres pour qu'ils mettent à jour la sélection courante du texte. Cela est réalisé en utilisant la méthode *SetSel* du contrôle d'édition *GetRichEditCtrl* associé à la classe *CWordPadView*. De manière similaire à Tallao, l'indice du caractère et le texte situé sous la souris sont obtenus respectivement via les membres *GetCharPos* et *GetWindowText* de *GetRichEditCtrl*.

[1] Ces messages souris sont WM_LBUTTONDOWN, WM_LBUTTONUP, WM_LBUTTONDBLCLK et WM_MOUSEMOVE, respectivement envoyés par le système aux applications lors d'un appui et d'un relâchement du bouton gauche de la souris, d'un double-clic et d'un mouvement de la souris.

[2] Tallao est une application MDI (Multiple Document Interface), c'est à dire qu'elle est constituée d'une fenêtre dite fenêtre mère qui contient une ou plusieurs fenêtres d'édition pouvant être ouvertes en même temps. De ce fait, elle doit déterminer la fenêtre fille active pour calculer la sélection associée.

Dans **LaoWord**, un *hook* est créé pour intercepter les messages Windows. La fonction interceptant ces messages est contenue dans une librairie dynamique chargée par Word lors de son lancement. Cela permet au code contenu dans cette fonction d'accéder à la fenêtre d'édition de Word via l'interface CAPI (voir I.3.2). CAPI est utilisée à la fois pour connaître la position courante du curseur, la position du pointeur de la souris dans le texte, pour accéder au contenu de la fenêtre et pour renvoyer la nouvelle sélection à afficher. Depuis la version 97 de Word, il est possible de réaliser cette communication via l'interface *Common Object Model* (COM) puisque Word est devenu alors un objet COM. Depuis Word 2000, les différentes versions de cette interface simplifient considérablement les traitements et offrent des API permettant d'améliorer le fonctionnement, en particulier *Word.ActiveWindow.RangeFromPoint*, une API qui fournit l'indice du caractère sous la souris et qui manquait précédemment, contraignant à des contournements complexes.

Le calcul de la nouvelle sélection est réalisé de la même façon dans les trois logiciels. Un automate analyse le texte situé dans le voisinage du pointeur souris. Il est fait de deux algorithmes : un premier algorithme qui détermine toutes les chaînes de ce voisinage pouvant être une syllabe et un deuxième déterminant si la chaîne est ou non une syllabe. Une syllabe contenant au moins deux caractères et au plus sept, le premier algorithme présente au deuxième toutes les chaînes contenant entre deux et sept caractères et situés sous le pointeur souris. Le deuxième algorithme est réalisé par un code obtenu en compilant une grammaire décrivant les syllabes laotiennes. Il détermine ainsi si la chaîne présentée appartient au langage décrit par cette grammaire ou non [Berment 1997 ; Berment 1998]. Cette méthode est rappelée au chapitre III.2.7.

II.2.3 Tri lexicographique

La fonction de tri s'appuie sur le fait que l'ordre lexicographique laotien dérive directement de celui des syllabes (voir II.1.2.4) qui, dans cette langue à tendance monosyllabique, sont souvent aussi des mots. Le tri d'un tableau en laotien sera donc basé sur une fonction de comparaison de syllabes. Afin de bénéficier du tri natif de Word[1], nous calculons des chaînes numériques à partir des syllabes constituant les mots à comparer, de telle manière que l'ordre sur les chaînes formées des nombres obtenus corresponde à l'ordre lexicographique sur les mots. Nous préfixons alors ces mots par les chaînes calculées, le tri de Word s'effectuant alors sur ces chaînes. Cette méthode permet aussi, de manière simple, de paramétrer le tri en modifiant le calcul du préfixe numérique.

La forme graphotaxique générale de la syllabe laotienne est CI-[A]-V-[CF][2], à ceci près que la place de la voyelle est variable. Elle peut être placée avant, après, voire répartie autour de la consonne initiale et peut être constituée de plusieurs caractères (voir II.1.1.2). L'algorithme qui calcule la chaîne numérique commence par segmenter le ou les mots en syllabes puis renvoie la concaténation, avec insertion d'espace entre chaque chaîne, des chaînes numériques calculées pour la suite de syllabes obtenues. La chaîne numérique calculée pour une syllabe est le résultat d'une analyse pragmatique de la syllabe de la gauche vers la droite :

- lorsque la syllabe commence par une consonne :
  - identification de la ou des consonnes initiales,
  - identification de l'éventuel accent,
  - identification de la voyelle,
  - identification de l'éventuelle consonne finale,

[1] Nous n'avons implémenté le tri lexicographique que dans LaoWord.

[2] **Notations** : CI pour Consonne Initiale, V pour Voyelle, A pour Accent, et CF pour consonne finale, les [x] désignant un élément facultatif. Notons que la consonne initiale peut être un groupe consonantique constitué de deux consonnes.

- lorsque la syllabe commence par une voyelle :
    - hypothèses sur la voyelle (la fin de la voyelle peut se trouver à droite de la consonne),
    - identification de la ou des consonnes initiales,
    - identification de l'éventuel accent,
    - fin d'identification de la voyelle,
    - identification de l'éventuelle consonne finale.

Cette analyse renseigne une structure contenant des identifiants pour la première et l'éventuelle deuxième consonne initiale, l'éventuel accent, la voyelle et l'éventuelle consonne finale. Ces identifiants sont utilisés comme indice de tables numériques :

- CI1 : première consonne initiale (101 à 141),
- CI2 : deuxième consonne initiale (100 à 124),
- V1 : voyelle (valeur vocalique) (101 à 117),
- V2 : durée de la voyelle (1 à 2),
- CF : consonne finale (100 à 124),
- A : accent (0 à 5).

Chaque syllabe est donc représentée par une chaîne numérique de quatorze caractères. L'ordre de concaténation de ces différentes composantes dépend du système choisi[1] :

- CI-V-CF-A pour le système de Marc Reinhorn,
- CI-CF-V-A pour le système de Sila Viravong.

<u>Exemples</u> ແຕ່ [tè] et ຕາ [ta] (noter la voyelle antéposée ແ dans ແຕ່) :

→ ແຕ່ [tè] (signifiant « mais ») donnera 11010010621100 : 110 pour la consonne ຕ (t), 100 pour l'absence de deuxième consonne initiale (consonne simple), 106 pour la voyelle ແ (è), 2 pour indiquer une voyelle longue, 1 pour l'accent ◌່ (accent 1) et 100 pour l'absence de consonne finale.

→ ຕາ [ta] (signifiant « œil ») donnera 11010010120100 : 110 pour la consonne ຕ (t), 100 pour l'absence de deuxième consonne initiale (consonne simple), 101 pour la voyelle າ (a), 2 pour indiquer une voyelle longue, 0 pour l'absence d'accent et 100 pour l'absence de consonne finale.

► ແຕ່ sera donc classé après ຕາ car 11010010621100 est supérieur à 11010010120100.

Pour trier un texte laotien avec LaoWord, il faut le placer dans un tableau comme dans l'exemple ci-dessous, puis sélectionner les lignes et les colonnes à trier (éventuellement tout le tableau) et cliquer sur l'icône de tri de la barre d'outils de LaoWord (voir le chapitre I.3.2).

ravaudeur, -euse nm., -f.	ຜູ້ທສິບແສ່ວດາບເກົ່າ
rave nf.	ຫົວຜັກກະລ່ຳປີ
ravenala nm.	ຕົ້ນກ້ວສກາບຊ້ອນ
ravenelle nf.	ຕົ້ນດອກແດງ
ravi, -e am., -f.	ຊື່ນເບີກບານ ດີໃຈ
ravier nm.	ແສ່ງນ້ອຍ
ravigotant, -e am., -f.	ພາໃຫ້ມີແຮງ
ravigoter vt.	ໃຫ້ກຳລັງ, ພາໃຫ້ມີແຮງ

Figure 29 : Extrait d'un tableau à trier

[1] Avec CI = CI1:CI2 et V = V1:V2, les deux points représentant ici la concaténation.

On obtient une fenêtre telle que dans l'exemple suivant.

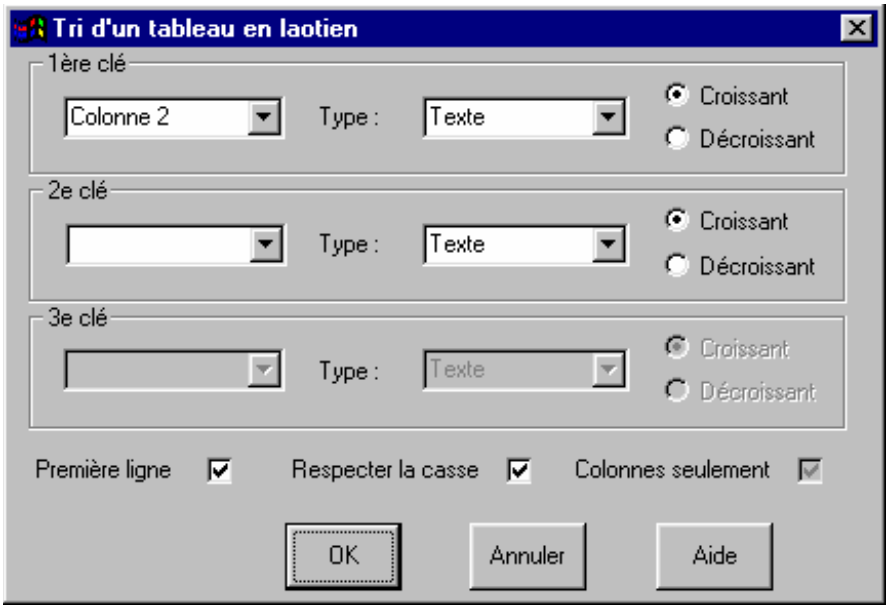


Figure 30 : Fenêtre du tri LaoWord

Comme nous le voyons, cette fenêtre suit la logique du tri de Word. Plusieurs clés sont possibles, le tri se faisant sur la base de la première clé puis, en cas de stricte identité de cellules, de la deuxième et de la troisième clé.

LaoWord facilite le paramétrage de sa fenêtre de tri en proposant les paramètres les plus probables pour la sélection à trier. En particulier, il localise une colonne du tableau écrite avec une police laotienne et la propose comme première clé dès l'ouverture de la fenêtre de tri comme dans l'exemple ci-dessus[1] où la colonne 2 est présélectionnée avec le type texte et l'ordre croissant.

Les paramètres proposés dans la fenêtre sont les paramètres classiques de Word :

- choix Texte/Nombre/Date (« Texte » est toujours sélectionné pour les colonnes en laotien),
- ordre croissant ou décroissant,
- inclure ou non la première ligne dans le tri (parfois, c'est un titre que l'on veut garder en haut),
- prise en compte ou non des majuscules/minuscules (casse), ce paramètre n'est pas pris en compte pour le tri des parties en laotien,
- tri sur toutes les colonnes ou uniquement sur les colonnes sélectionnées (paramètre accessible quand une partie seulement des colonnes est sélectionnée, par exemple 2 colonnes sur 3).

[1] Dans cet exemple, la 3e clé n'est pas autorisée parce que le tableau n'a que 2 colonnes.

II.2.4 Fonctions diverses

II.2.4.1 GESTION DE LA NON-NORMALISATION

II.2.4.1.1 Forme canonique du texte

Nous avons vu, aux chapitres II.1.2.1 et II.1.2.2, que l'absence de norme pour l'encodage des caractères laotiens et la possibilité d'obtenir le même résultat visuel à partir de saisies différentes posent le problème de l'unicité de la représentation du texte. Deux des fonctions précédemment présentées — la sélection du texte et le tri lexicographique — sont concernées et doivent donc être précédées d'une mise en forme canonique du texte pour n'avoir à opérer que sur une forme unique.

La mise en forme canonique s'effectue en deux étapes, à partir de la représentation pivot (voir le chapitre II.2.1.2) :

- transcodage dans une représentation de base (un encodage de référence) pour pallier l'absence de norme d'encodage,
- standardisation dans cet encodage de référence pour éliminer les cas de saisie non univoque.

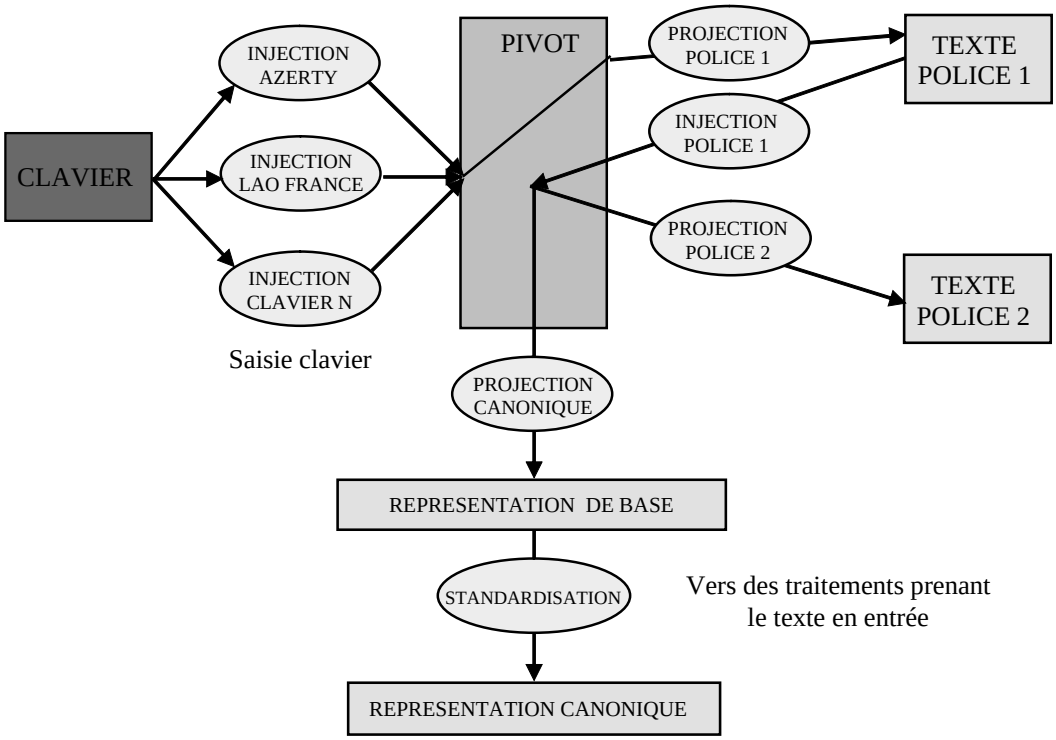


Figure 31 : Logique des traitements de mise en forme canonique

Contrairement aux fonctions précédentes, la mise en forme canonique ne fait pas directement partie des services listés comme éléments de l'informatisation d'une langue. Elle est néanmoins indispensable pour simplifier l'utilisation d'un traitement de texte en laotien.

***Nota** : Dans le cas de la saisie (voir les extraits de programme donnés aux chapitres II.2.1.2.1 et II.2.1.2.2) ou des changements de polices, le traitement se fait aussi en deux étapes (la figure ci-dessus montre que la standardisation n'est, dans ce cas, pas nécessaire) :*

- *un transcodage touche-pivot (saisie) ou texte-pivot (changement de polices),*
- *un transcodage pivot-texte dans l'encodage de destination.*

II.2.4.1.2 Fonction de transcodage

La fonction de transcodage s'appuie sur un ensemble contenant tous les caractères laotiens existants dans les différentes polices existantes soit au total 108 caractères ainsi que les 256 caractères du Latin 1. L'accès à cet ensemble pivot se fait par une injection, par exemple lorsque l'on frappe une touche au clavier ou lorsque l'on demande un changement de police, tous les caractères ne figurant pas sur le clavier et n'existant pas forcément dans une police donnée. Nous avons appelé projection l'opération inverse qui transforme un code pivot en un ou deux caractères. Voyons dans la figure ci-dessus les différents cas de figure.

Saisie au clavier
La saisie d'un caractère au clavier provoque l'appel d'une fonction qui transforme le caractère tapé en code pivot. Selon le clavier utilisé, cela peut être l'identité (clavier AZERTY) ou une autre fonction définie par une table de correspondance (une table par clavier). De là, le code pivot obtenu est transformé en fonction de la police utilisée, une deuxième table étant utilisée pour cela (une table par police).

Changement de police
Dans ce cas, le texte est transformé en représentation pivot puis projeté dans l'encodage de la nouvelle police.

Traitement utilisant le texte en entrée (algorithme de sélection, tri lexicographique)
Dans ce cas, le texte est transformé en représentation pivot puis projeté dans la représentation utilisée pour les traitements qui est un encodage particulier (« représentation de base »). Le texte obtenu subit alors une standardisation, comme cela est décrit dans le chapitre suivant.

II.2.4.1.3 Fonction de standardisation

Les problèmes de saisie éliminés par la standardisation sont les suivants :

- répétition de caractères sans avance (voyelles suscrites ou souscrites, accents), par exemple :
  - ◌́ + ◌́ + ◌́ + ... = ◌́ (accent *may ek*),
  - ◌ ◌ + ◌ ◌ + ◌ ◌ + ... = ◌ ◌ (voyelle *o ouvert*),
- inversion V/A, où V est une voyelle suscrite ou souscrite, et A un accent, par exemple :
  - ◌ ◌́ = ◌́ + ◌ ◌ = ◌ ◌́ (voyelle *o ouvert* + accent *may ek*),
- caractères doubles, caractères dédoublés et ligatures, par exemple :
  - ◌ำ + ◌ ◌ = ◌ ◌ + ◌ำ = ◌ำ (voyelle *am*),
  - ◌ ◌ = ◌ ◌ + ◌ ◌ (voyelle *è*),
  - ◌ ◌ = ◌ ◌ + ◌ ◌ (ligature *ho-no*),
- les accents ont parfois deux hauteurs possibles,
- les « ◌ู » (voyelle *ou*) ont parfois deux hauteurs possibles.

À ces types de « cas non standard » correspond le traitement suivant :

- suppression des répétitions de voyelles sans avance et d'accents,
- mise en ordre « canonique » des voyelles sans avance et des accents,
- mise en forme « canonique » des caractères doubles,
- transformation des accents bas en accents hauts,
- transformation des « ◌ู » hauts en « ◌ู » bas.

Cette fonction de standardisation est utilisée dans d'autres services que les deux présentés précédemment. En particulier, LaoWord l'utilise pour des fonctions de recherche dans un lexique, de transcription phonétique et de mise en forme du texte. Ces deux dernières fonctions sont décrites dans les chapitres suivants.

II.2.4.2 MISE EN FORME DU TEXTE

Les quatre fonctions ci-dessous, complémentaires des précédentes, ont été développées pour améliorer le rendu visuel et la mise en page.

Réglage de la hauteur des accents (LaoWord)

Certaines polices laotiennes disposent de deux hauteurs d'accent, en particulier les accents 1 et 2, pour s'adapter au mieux à la voyelle associée. Il arrive cependant que le texte saisi ait des accents trop hauts par rapport à la voyelle située au dessous ou, ce qui est plus gênant, que l'accent soit trop bas et se mêle à une voyelle suscrite. Cette fonction corrige la hauteur des accents d'une sélection.

Réglage de la hauteur des voyelles « u » (◌ᳪ) (LaoWord)

Comme pour les accents, certaines polices offrent deux hauteurs pour les « sala u » courts et longs (◌ᳩ et ◌ᳪ) et certains cas de figure peuvent entraîner un positionnement impartait des « sala u ». Cette fonction corrige la hauteur de ces voyelles dans une sélection.

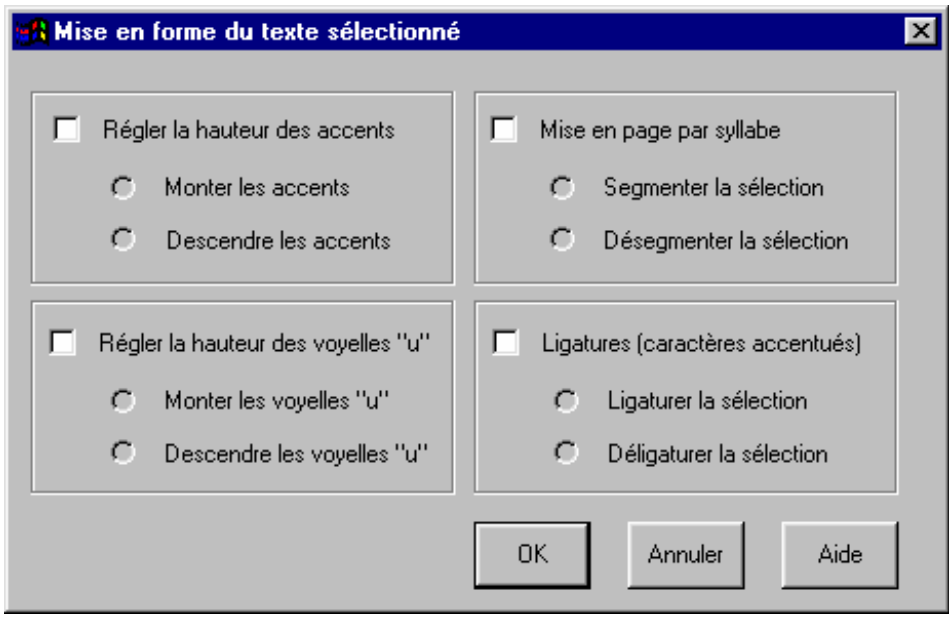


Figure 32 : Fenêtre de mise en forme du texte dans LaoWord

Mise en page par syllabe (Tallao, LaoPad et LaoWord)

Parmi les inconvénients causés par l'absence d'espace entre mots et rappelés au chapitre II.1.2.3, celui de la coupure des fins de lignes peut être résolu de deux façons :

- en calculant dynamiquement les points où la césure doit être réalisée,
- en insérant des caractères de césure invisibles entre les syllabes du texte sélectionné.

La première méthode a été mise en œuvre dans Tallao, la seconde dans LaoWord.

Dans **Tallao**, une fonction, dite de *callback* parce qu'elle est appelée par Windows, calcule le prochain endroit où la classe EDIT de la fenêtre d'édition (voir le chapitre II.2.1.2.1) doit couper les fins de ligne. Cette classe connaît le point d'entrée de la fonction *callback* grâce au message EM_SETWORDBREAKPROC qui lui est envoyé lors des initialisations. La fonction est appelée chaque fois que la fenêtre d'édition a besoin de couper une ligne. Cette fenêtre passe en paramètre l'indice du caractère à partir duquel la fonction doit chercher un point de coupure correct à droite ou à gauche ou encore un indice de caractère afin de déterminer s'il est un caractère de césure. La fonction renvoie vrai ou faux dans ce dernier cas et l'indice demandé dans les deux premiers.

Dans **LaoWord**, les différentes API disponibles n'offrent pas cette possibilité de fonction *callback*. La solution de remplacement consiste à insérer des caractères de césure invisibles entre les syllabes du texte sélectionné. Ce caractère dépend de la police utilisée et peut être :

- une espace sans largeur prévue dans la police et placée au code correspondant au trait d'union (code 45 en décimal ou 2D en hexadécimal),
- le caractère Unicode appelé ZERO WIDTH SPACE (ZWSP, code 200B) géré uniquement par les classes d'édition compatible Unicode,
- le caractère de code 31 en décimal ou 1F en hexadécimal, qui représente le caractère de contrôle US (Unit Separator) et est utilisé comme caractère de césure par Word.

Word coupe alors les lignes au niveau de ces caractères de césure et contribue ainsi à une mise en page correcte en évitant que les lignes ne soient coupées au milieu d'une syllabe. Cette technique introduit des caractères qui ne font pas partie du texte et qu'il pourra être nécessaire de supprimer pour réaliser certains traitements.

Ligature des voyelles accentuées (LaoWord)

La plupart des polices de caractères laotiens contiennent des ligatures faites d'une voyelle suscrite et d'un accent (par exemple ◌ັ◌ ◌ ◌ັ◌ à la place de ◌ ◌ັ◌). Cela est dû au clavier traditionnel « Duang Jan » qui contient de telles ligatures. Ces mêmes polices contiennent généralement aussi les voyelles suscrites et les accents correspondants, pris séparément.

Il peut arriver, par exemple après plusieurs changements de polices successifs, qu'une ligature se retrouve transformée en une voyelle suivie d'un accent, cela même si la police finale contient cette ligature[1].

La fonction de ligature des voyelles accentuées permet de transformer les couples voyelle-accent d'une sélection en ligatures et inversement, afin d'atteindre le meilleur rendu esthétique.

NOTA : Lorsque l'on utilise des polices Unicode, les fonctions d'optimisation graphique (réglages de hauteur et ligatures), lorsqu'elles existent, sont gérées par les polices elles-mêmes (en particulier, les polices OpenType). Ces fonctions ne sont donc pas utilisées avec les polices Unicode.

[1] Pour éviter cela, on peut choisir de ligaturer systématiquement les couples « accent-voyelle » lors des changements de police.

II.2.4.3 TRANSCRIPTIONS PHONÉTIQUES

La fonction de transcription phonétique que nous avons réalisée dans LaoWord est très proche du code développé pour le tri lexicographique (voir chapitre II.2.3). Cette fonction de tri réalise en effet :

- une segmentation en syllabes,
- une analyse de chaque syllabe en six identifiants :
  - CI1 : première consonne initiale,
  - CI2 : deuxième consonne initiale,
  - V1 : voyelle (valeur vocalique),
  - V2 : durée de la voyelle,
  - CF : consonne finale,
  - A : accent.

L'algorithme de transcription phonétique réalise les mêmes étapes que le tri puis utilise les six identifiants calculés comme indice de tables phonétiques. Trois tables coexistent et fournissent trois transcriptions différentes :

- une transcription libre n'utilisant que des caractères latins,
- une transcription phonétique utilisant l'Alphabet Phonétique International,
- une transcription du texte laotien, entièrement réversible.

Pour obtenir la transcription d'un passage dans LaoWord, il faut le sélectionner comme montré ci-dessous :

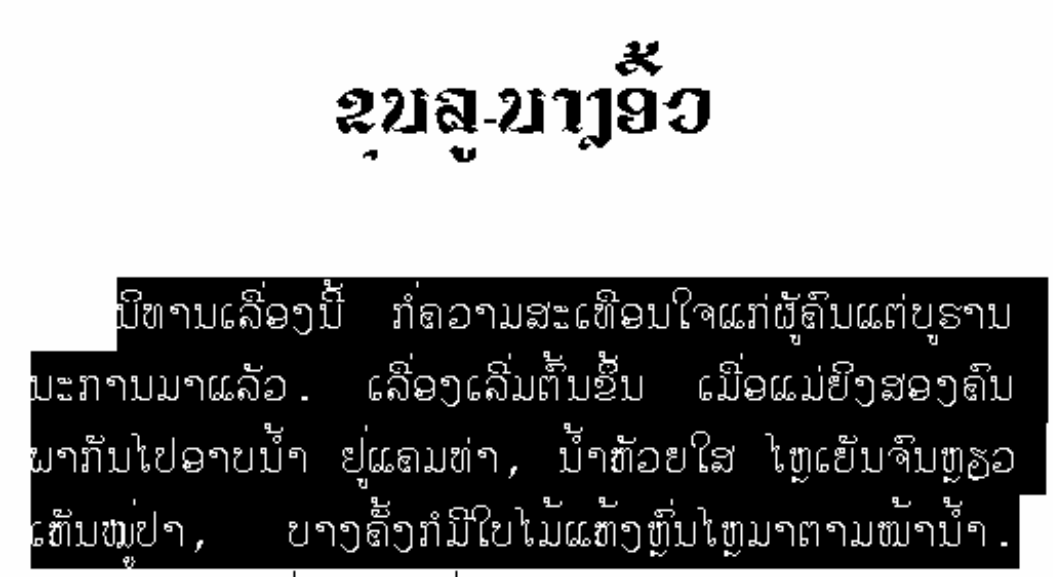


Figure 33 : Passage de texte sélectionné

puis cliquer sur l'icône de transcription phonétique dans la barre d'outils de LaoWord (voir le chapitre I.3.2).

Une fenêtre contenant la transcription s'ouvre alors comme cela est montré dans l'exemple en transcription libre ci-dessous.

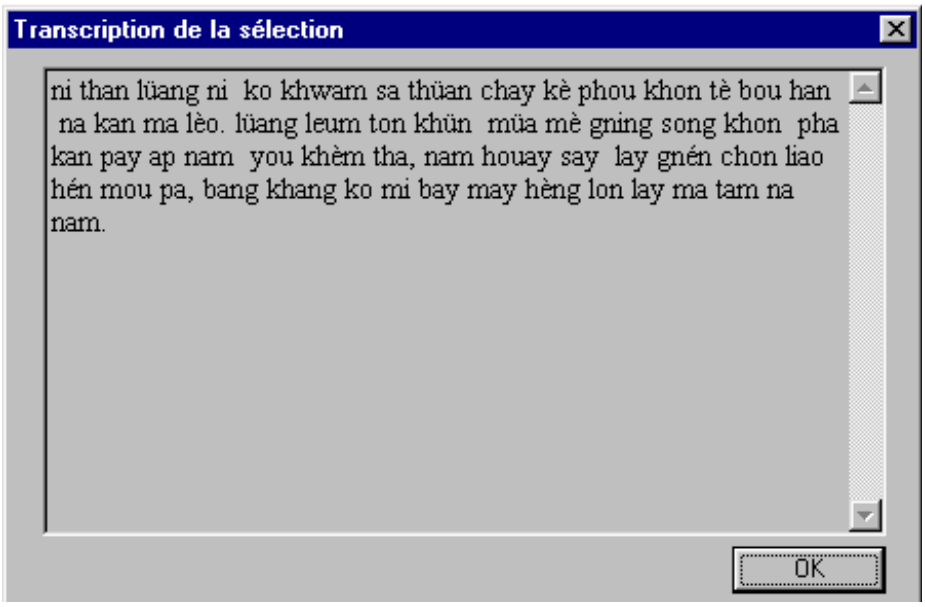


Figure 34 : Affichage de la transcription libre dans une fenêtre

En fonction du type de transcription sélectionné, le texte ຂຸນລู ນาๆอ๊ือ pris précédemment comme exemple donnera :

- en transcription « libre » :
  ni than lüang ni ko khwam sa thüan chay kè phou khon tè bou han na kan ma lèo. lüang leum ton khün müa mè gning song khon pha kan pay ap nam you khèm tha, nam houay say lay gnén chon liao hén mou pa, bang khang ko mi bay may hèng lon lay ma tam na nam.

- en transcription « phonétique » :
  ni tʰa:n lɯ̀:aŋ nî: kɔ̀: kʰwa:m sa tʰɯ:an cay kè: pʰù: kʰon tè: bu: ha:n na ka:n ma: lê:w. lɯ̀:aŋ là:m tôn kʰɯ̀n mɯ̀:a mè: ɲiŋ sɔ̌:ŋ kʰon pʰa: kan pay ʔà:p nâm yù: kʰɛ:m tʰà:, nâm hù:ay sǎy lǎy ɲen con ǐ:aw hěn mù: pa:, ba:ŋ kʰâŋ kɔ: mi: bay mây hɛ̀:ŋ lòn lǎy ma: ta:m nà: nâm.

- en transcription « réversible »[1] :
  ni tha:n lùa:ŋ1 ni:2 kɔ:1 khwa:m 'sa thùa:n cae kɛ:1 'phu:2 khon tɛ:1 bu: ha:n na ka:n ma: lɛ:w2. lùa:ŋ1 lœ:m1 ton2 'khùn2 mùa:1 mɛ:1 ñiŋ 'sɔ:ŋ khon pha: kan pai ʔa:b nam2 yu:1 khɛ:m tha:1, nam2 'hwa:y2 'sae lai ñen con lia:w 'hen mu:1 pa:, ba:ŋ khaŋ2 kɔ: mi: bae mai2 'hɛ:ŋ2 lon1 lai ma: ta:m na:2 nam2.

[1] Les chiffres ne sont pas traités dans cette transcription, ce qui permet d'utiliser les chiffres 1 à 4 pour transcrire les accents.

II.3 DEUXIÈME PHASE : AIDE À LA TRADUCTION, DICTIONNAIRES, OPEN SOURCE, UNICODE

II.3.1 Aide à la traduction : *LaoLex*

II.3.1.1 AIDE À LA TRADUCTION ET CRÉATION COOPÉRATIVE D'UN DICTIONNAIRE

Repartant de l'idée de contribution linguistique généralisée développée au chapitre I.2.2.3 mais aussi de celle du recours aux diasporas du chapitre I.2.2.5, nous avons développé un site Internet[1] destiné :

- à fournir un service d'aide à la lecture active : **LaoLex**,
- à développer un dictionnaire à la fois d'usage et bilingue laotien-français : **LaoDict**.

Ce site propose à toutes les personnes le souhaitant de développer des ressources linguistiques de manière coopérative.

Les trois tâches ci-dessous constituent la première étape du **projet LaoLex** :

a) un groupe de personnes formées crée des entrées lexicales, utilisant pour partie des dictionnaires papier existants,
b) un groupe de linguistes dérive de ces articles un dictionnaire de référence disponible en ligne,
c) plusieurs outils et applications sont développés pour utiliser ces ressources.

En parallèle de ce travail de spécialistes, une contribution lexicale provenant d'un large public d'internautes est aussi compilée par le groupe de linguistes. Cette contribution complémentaire s'appuie sur un **environnement d'aide à la traduction** disponible sur le site Internet http://sabaidi.imag.fr/. En particulier, un **éditeur bilingue** laotien-français est proposé aux visiteurs associé à un **service de traduction mot à mot**. Ces services utilisent une police particulière non Unicode[2] de la famille Alice 92 qui a été adaptée pour pouvoir fonctionner à la fois sous Windows et sous MacOS (même nom de police — « Montaigne Lao » — mais version spécifique pour chacun des deux systèmes d'exploitation).

[1] http://sabaidi.imag.fr/.

[2] Le choix d'une police non Unicode provient du principe de création lui-même. En effet, parmi les utilisateurs de la première heure, aucun ne disposait d'environnements logiciels capables d'Unicode. Ainsi, malgré la maquette PapiLex de laquelle nous aurions pu repartir (voir I.3.3), nous avons donc opté pour cette solution technique non Unicode afin d'assurer une ouverture et une compréhension les plus larges possibles. Ce choix demeure provisoire, un portage au format Unicode pouvant être réalisé dès que les moyens des utilisateurs le permettront.

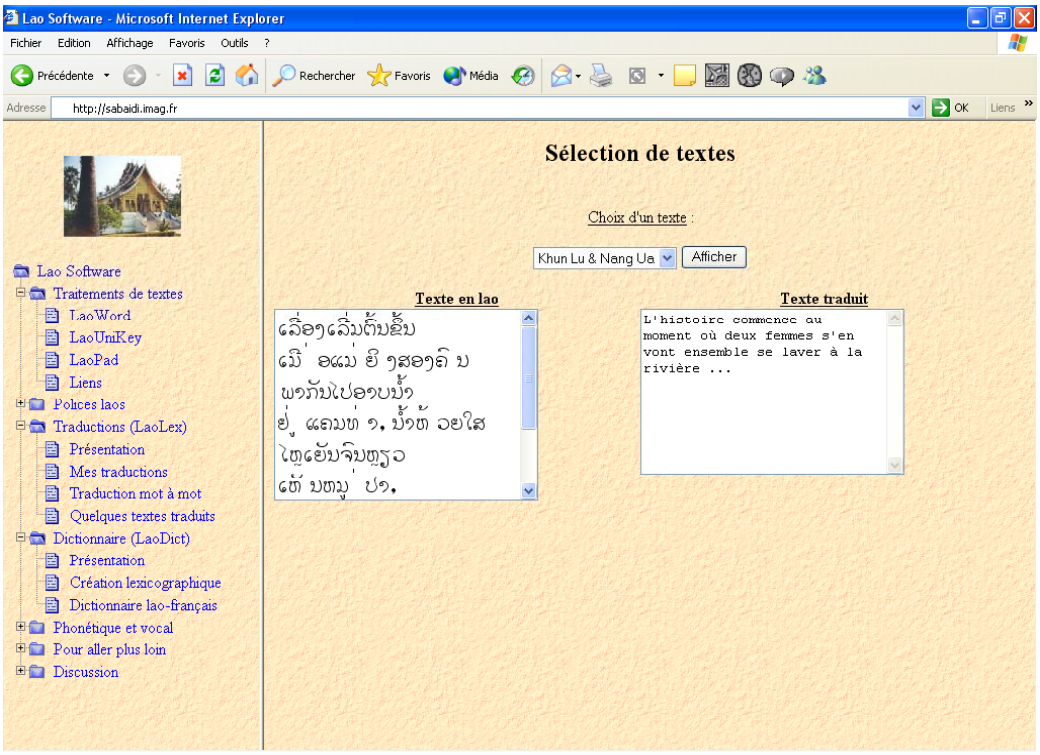


Figure 35 : Editeur bilingue laotien-français

Figure 36 : Traduction mot à mot

S'ils veulent être aidés pour réaliser leurs traductions personnelles, les utilisateurs peuvent demander une traduction mot à mot de leurs textes en laotien. Dans le cas où un mot ne serait pas présent dans le dictionnaire ou s'il est incorrectement traduit, ils ont la possibilité d'ajouter ou d'améliorer l'entrée correspondante dans un **dictionnaire personnel** qui sera utilisé prioritairement lors de leurs prochaines demandes de traduction. Ces dictionnaires personnels, ainsi que les textes bilingues eux-mêmes, ne peuvent être vus que par leurs auteurs et sont stockés sur le serveur. Ainsi, pour améliorer le dictionnaire de référence en ligne — et donc le service de traduction mot à mot — les utilisateurs peuvent proposer que les articles de leurs dictionnaires personnels **soient intégrés au dictionnaire général**. Cette intégration est réalisée sous le contrôle du groupe de linguistes.

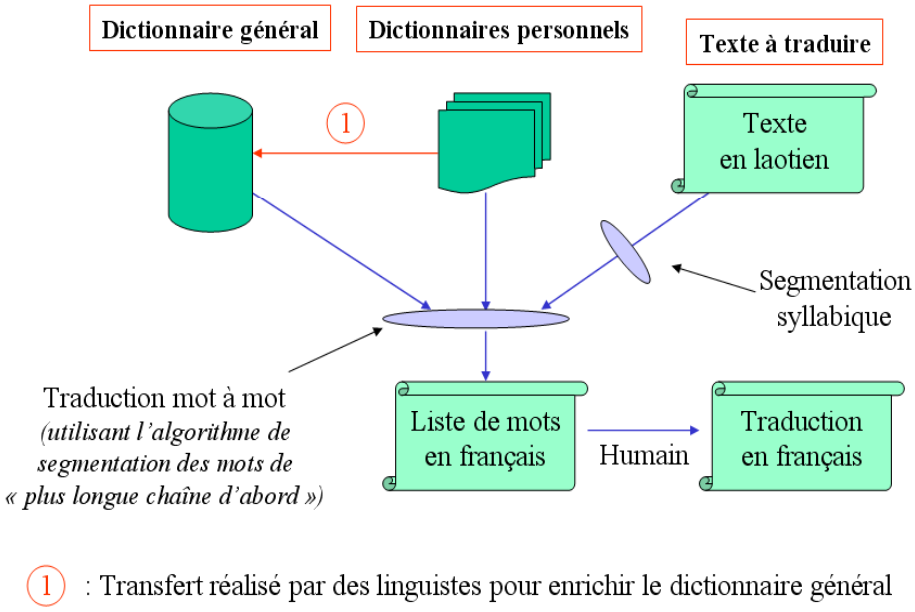


Figure 37 : Architecture fonctionnelle de LaoLex

II.3.1.2 ARCHITECTURE ET ALGORITHMES MIS EN ŒUVRE

II.3.1.2.1 Présentation

Les dictionnaires sont stockés sous forme de tables dans une base de données MySQL, de même que les informations relatives aux contributeurs. Des programmes en C++ sont utilisés pour segmenter les textes laotiens en mots, pour le tri lexicographique des dictionnaires et pour la transcription phonétique. Ils utilisent la technique de reconnaissance de syllabes [Berment 1998] rappelée au chapitre III.2.3 (le code de Tallao est réutilisé) ainsi qu'un algorithme de segmentation « plus longue chaîne d'abord » sans retour arrière (*longest match*, voir chapitre III.2.2.2 et par exemple [Meknavin et al. 1997]).

II.3.1.2.2 Saisie du texte en laotien

La saisie du texte peut se faire via les deux dispositions clavier les plus répandues (Duang Jan et Lao France) grâce au logiciel LaoMonoKey ou au programme JavaScript LaoMonoWeb et à des champs de formulaires HTML *TextArea* et *Input* (voir les chapitres I.3.3 et II.2.1). Le programme de saisie en Unicode LaoUniKey ou le script LaoUniWeb (voir le chapitre II.3.3) pourront être utilisés lors d'un portage futur de LaoLex / LaoDict au format Unicode.

II.3.1.2.3 Algorithme utilisé pour la traduction mot à mot

Lors d'une traduction mot à mot d'un texte laotien en français, le texte à traduire est envoyé au serveur via un formulaire HTML. Le traitement de la requête est réalisé par un programme en langage PHP qui procède de la manière suivante :

1. appel d'un CGI, écrit en C++, qui prend en entrée le texte laotien et qui renvoie le texte segmenté en syllabes, avec pour chaque syllabe :
    - sa forme d'origine dans le texte, utilisée pour l'affichage,
    - sa forme canonique, utilisée pour les requêtes MySQL,
    - sa transcription phonétique,
2. recherche, dans le dictionnaire personnel[1] et dans le dictionnaire général, de la plus longue suite de syllabes formant un mot[2] ; cette recherche, qui est faite de la gauche vers la droite du texte, produit un tableau de mots ; en cas d'échec dans la recherche d'un mot, on se décale d'une syllabe vers la droite,
3. affichage des mots (forme d'origine) et de leurs transcriptions, suivis des différentes traductions trouvées (tous les sens d'un mot sont affichés) ; une ligne est attribuée à chaque mot.

II.3.1.2.4 Accès aux lexies, mise en forme canonique et tri alphabétique

Lors de l'ajout d'un mot dans un dictionnaire personnel, la forme canonique (voir II.2.4.1) du mot est calculée par un CGI, puis stockée dans le champ « LexieLS » (**Lexie** sous la forme **L**aotienne **S**tandard) de la base LaoDict, avec sa forme d'origine (pour de futurs affichages) et les divers champs saisis (traduction, partie du discours…).

En plus de ces champs renseignés par le lexicographe, un champ « Tri » est renseigné automatiquement par le résultat du traitement d'un CGI. Ce traitement prend en entrée la forme canonique du mot laotien de n syllabes et calcule une chaîne numérique de n fois quatorze caractères caractérisant la place du mot dans le dictionnaire (voir II.2.3). Cette chaîne est de la forme $(CCCXXXVVVDFFFA)^+$:

- CCC = code de la consonne initiale,
- XXX = code de l'éventuelle deuxième consonne initiale (valeur minimale si absente),
- VVV = code de la voyelle,
- D = code de la durée de la voyelle,
- FFF = code de la consonne finale,
- A= code de l'éventuel accent.

Les différents codes utilisés respectant l'ordre de leur élément associé — consonne, voyelle... — et étant concaténés dans l'ordre CI-V-CF-A (voir II.2.3), il sera possible de réaliser un affichage dans l'ordre lexicographique de Marc Reinhorn en introduisant simplement la précision « order by 'Tri' » dans la requête SQL. Par exemple, la requête suivante permettra de lister toutes les lexies de $Auteur dans l'ordre lexicographique laotien défini par Marc Reinhorn, les entrées étant renvoyées ordonnées dans la variable $resultat.

```
$resultat=mysql_query(select * from $dico where Auteur='$Auteur' order by 'Tri' );
```

Lors de la consultation du dictionnaire, la forme canonique du mot demandé est calculée par un CGI. Cette forme est ensuite recherchée dans le champ LexieLS de la base.

[1] Dans le cas d'un visiteur n'ayant pas de compte personnel, seul le dictionnaire général est utilisé.
[2] Les cas d'ambiguïté étant rares en laotien, ils n'ont pas été traités dans cette première version de LaoLex. C'est actuellement la plus longue syllabe possible qui est utilisée.

II.3.2 Ressources linguistiques laotiennes : *LaoDict*

II.3.2.1 LES ARTICLES DE LAODICT

D'un point de vue lexicographique, la base de données LaoDict suit les concepts de la lexicographie explicative et combinatoire ([Mel'čuk et al. 1995]). Les articles contiennent le mot lui-même (ou la locution), sa partie du discours, sa traduction en français et, facultativement, des exemples et des expressions idiomatiques. La formule sémantique, le régime et les fonctions lexicales de la lexicographie explicative et combinatoire ne font pas encore partie des informations attachées aux articles.

L'entrée élémentaire est appelée lexie ou unité lexicale. Cette notion de lexie est définie dans [Mel'čuk et al. 1995]. Il s'agit d'un lexème ou d'un phrasème, c'est à dire d'un mot ou d'une locution pris dans une acception bien spécifique. Par exemple, le mot ຢູ່ a plusieurs sens qui sont : « vivre », « être »... Ces sens correspondront donc à plusieurs articles (lexies) dans LaoDict.

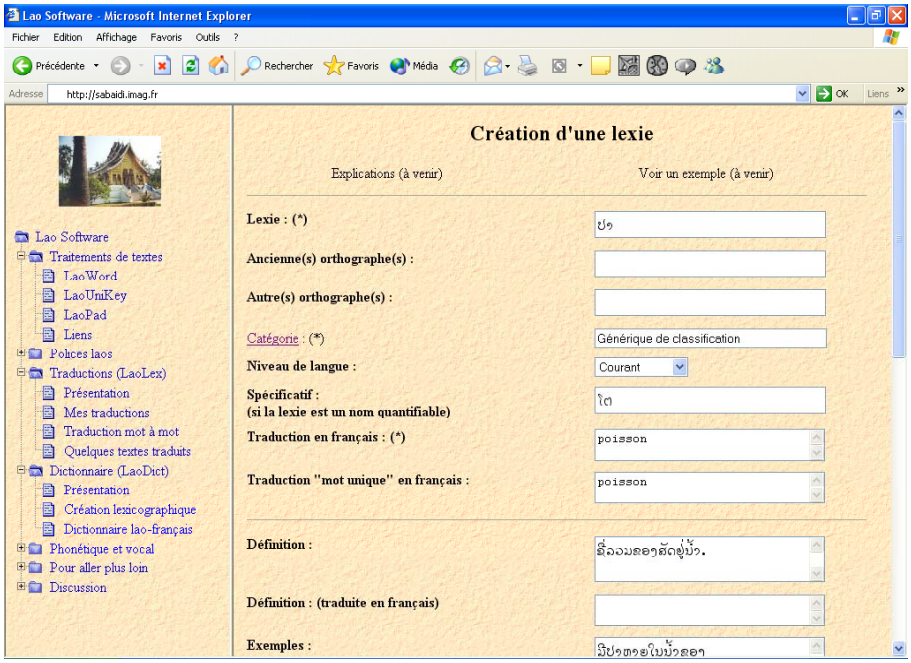


Figure 38 : Page de création d'une entrée

Plusieurs champs viennent compléter ces informations, de manière à offrir un champ d'application plus large de la base de données, en particulier pour réaliser des dictionnaires papier (exemples courts) et des outils d'aide à la traduction (traductions « mot unique », destinées à donner quelques éléments de traduction français-laotien en inversant le sens du dictionnaire). Quelques exemples de lexies sont proposés en annexe A.10.

II.3.2.2 PARTIES DU DISCOURS DANS LAODICT

II.3.2.2.1 Trois niveaux de parties du discours

Un important travail de réflexion et de synthèse a été nécessaire pour établir une liste cohérente et détaillée de parties du discours pour le laotien[1]*. Cette liste, qui n'existait pas, était nécessaire pour intégrer le laotien au projet Papillon en définissant son schéma XML spécifique (voir annexe A.9). Nous noterons le déplacement, dans notre approche, de certaines catégories par rapport à la grammaire traditionnelle. En particulier :*

- *les **spécificatifs** (ລັກສະນະນາມ / ລັກສນະນາມ) sont classés parmi les pronoms du fait de leurs rôles : compter (ລົດສາມຄັນ : trois voitures), désigner (ລົດຄັນນີ້ : cette voiture) ainsi que comme pronom (ລົດລາວແມ່ນຄັນສີດຳ : sa voiture est la noire), bien que se comportant comme des noms avec les adjectifs,*
- *les **collectifs** et les **qualificatifs**, généralement classés avec les noms, et que nous avons rattaché aux prédicatifs parce qu'ils marquent le genre et le nombre.*

À chaque entrée du dictionnaire LaoDict[2] est associée une catégorie (ປະເພດ). Cette catégorie — ou partie du discours — est attribuée en fonction de critères principalement syntaxiques mais aussi sémantiques et morphologiques.

Nous proposons une description en trois niveaux :

- un niveau groupe,
- un niveau catégorie,
- un niveau sous-catégorie.

Le **premier niveau**, appelé groupe parce qu'il regroupe plusieurs catégories, coïncide avec les « sept parties du discours » de la grammaire « traditionnelle » du laotien ([Comité Littéraire 1962]) :

- noms (ຄຳນາມ),
- pronoms (ຄຳສັບພະນາມ / ຄຳສັພນາມ),
- verbes (ຄຳກິຣິຍາ),
- predicatifs[3] (ຄຳວິເສດ),
- prépositions (ຄຳບຸບພະບົດ / ຄຳບຸພບົດ),
- conjonctions (ຄຳສັນທານ),
- interjections (ຄຳອຸທານ).

Les **deuxième et troisième niveaux** dérivent principalement des travaux de Marc Reinhorn complétés par ceux de Lamvieng Inthamone. Le lien entre ces deux niveaux est parfois artificiel mais permet de conserver une continuité avec les dictionnaires existants.

[1] Les travaux originaux que nous présentons ici n'ont la prétention ni d'être définitifs ni d'être universels. Cette liste de parties du discours évolue au fil des remarques et du travail lexicographique.
[2] LaoDict ne contient que des mots, excluant ainsi, par exemple, la ponctuation et les chiffres laotiens.
[3] Terme emprunté à Marc Reinhorn.

Le deuxième niveau contient 57 catégories réparties quantitativement de la manière suivante (voir le détail à l'annexe A.7) :

- noms : 8 catégories,
- pronoms : 11 catégories,
- verbes : 10 catégories,
- prédicatifs : 22 catégories,
- prépositions : 1 catégorie,
- conjonctions : 2 catégories,
- interjections : 3 catégories.

Ces 57 parties du discours sont celles utilisées dans LaoDict (niveau catégorie).

Le troisième niveau (niveau sous-catégorie) est donné à l'annexe A.8 sous forme de tableaux structurés et illustrés par des exemples. La liste de ces parties du discours, affinée au fil du travail lexicographique, est encore fluctuante.

II.3.2.2.2 Implémentation dans LaoDict

Dans LaoDict, la partie du discours est sélectionnée dans une liste hiérarchique correspondant aux 57 catégories. L'avantage de la liste hiérarchique par rapport à une liste simple (« à plat ») est que la sélection se fait en deux temps, chacun proposant un nombre de choix limité. Au premier niveau (groupe), les sept parties du discours traditionnelles sont présentées. Alors, le lexicographe en ouvre un (par exemple *nom* : huit catégories) et sélectionne une catégorie (par exemple *Nom commun*). Les sous-catégories, encore instables, ne sont pas proposées dans la version actuelle. Notons que des raccourcis pourraient être implémentés dans le futur (par exemple *nc* pour *nom commun*).

D'un point de vue technique, les listes hiérarchiques sont réalisées en *JavaScript* grâce à un outil gratuit de Mysoft/Martins, appelé *Visual Folder Tree Builder*[1], qui génère du code JavaScript à partir d'une interface graphique ergonomique.

II.3.2.3 LES NIVEAUX DE LANGUE DANS LAODICT

Les niveaux de langue sont socialement importants en laotien. LaoDict utilise les onze niveaux de langue suivants :

- courant,
- respectueux,
- familier,
- argotique,
- spécialisé,
- recherché,
- bonze,
- royal,
- littéraire,
- parlé,
- archaïque.

Un travail de comparaison entre ces niveaux de langue et ceux choisis pour le thaï par les chercheurs de l'université Kasetsart à Bangkok est actuellement en cours. Le thaï et le laotien sont, en effet, des langues très voisines.

[1] http://mysoft.s5.com/.

II.3.2.4 DE LA CRÉATION DE LA BASE À L'EXPORT VERS PAPILLON

II.3.2.4.1 Expérimentation et mise au point du travail coopératif autour de LaoLex

Après la publication sur Internet du logiciel LaoLex et l'établissement de la liste des parties du discours, nous avons demandé aux étudiants de la licence de laotien aux Langues O' (promotions 2002-2003 et 2003-2004[1]) d'enrichir le dictionnaire LaoDict, dans le cadre de leur projet de traitement automatique du laotien. Nous avons aussi mis à contribution une enseignante du cursus, Madame Thakkhinh Jacqmin, qui a tenu le rôle du linguiste et géré l'intégration des lexies créées par les lexicographes au dictionnaire général.

Les étudiants ont reçu chacun quelques dizaines de mots courants de la langue laotienne à entrer dans le dictionnaire. Ils avaient pour cela à leur disposition des dictionnaires laotiens (Maha Sila Viravong, Onemanisone) et laotien-français (Reinhorn). Pour que le résultat puisse s'intégrer facilement dans Papillon, les étudiants ont séparé les différents sens possibles des mots et ont ainsi entré des lexies, au sens du chapitre II.3.2.1. Cela a fréquemment été nécessaire, les entrées de ces dictionnaires mêlant souvent les différents sens possibles des mots dans une même définition. Les étudiants 2003-2004 de la licence de laotien ajoutent, au moment de la rédaction de cette thèse, une centaine de lexies dans LaoDict.

II.3.2.4.2 Essai de travail coopératif autour de LaoLex

Nous avons tenté d'élargir l'audience de LaoLex/LaoDict d'un réseau d'universitaires à un public entièrement ouvert. Au printemps 2002, nous avons donc proposé, avec le concours d'Houmphanh Thongvilu, un créateur de forums laotiens très connu des internautes laotiens, le développement, par la communauté internaute, d'un dictionnaire laotien en ligne basé sur LaoLex/LaoDict.

Bien qu'ayant permis la création de quelques articles dans LaoDict, cette expérience fut globalement décevante. En effet, très rapidement, un projet concurrent s'est déclaré, trouvant l'idée intéressante mais voulant créer ses propres outils. Afin d'éviter le développement de deux dictionnaires différents, nous avons rejoint ce groupe, intitulé *Pak Lao*. Mais malgré les nombreuses heures passées en discussions, le groupe s'est étiolé sans qu'aucun outil ni aucune entrée de dictionnaire n'ait été créés.

De nombreuses personnes ayant été très motivées et prêtes à donner de leur temps à ce projet, nous renouvellerons cette expérience avant de conclure à l'incapacité d'un tel groupe à créer un dictionnaire de qualité. Il est cependant probable que des communautés plus restreintes et mues par un intérêt, professionnel ou autres, lié au service d'aide à la traduction — traducteurs professionnels, étudiants, chercheurs... — donneront une bien meilleure efficacité.

II.3.2.4.3 Schéma XML spécifique au laotien

Le contenu d'une entrée du dictionnaire Papillon est proche de celui d'une entrée LaoDict. Sa structure technique dérive du schéma DML défini par Gilles Sérasset dans [Sérasset 1994]. Cette structure est un schéma XML, commun à toutes les bases de données monolingues de Papillon. Pour adapter ce schéma à une langue particulière comme le laotien, la tâche principale consiste à définir ce qui est spécifique dans la langue, comme les parties du discours et le contexte social (niveau de politesse...). Les deux chapitres précédents conduisent donc au schéma XML suivant pour le laotien (le schéma complet avec les parties du discours est donné en annexe A.9).

[1] Il s'agit de Blandine de Chanet, Ya-Ambhan Manaschun, Keomany Senesoukdara (promotion 2002-2003), Nicolas Cordonnier, Fabrice Mignot, Patricia Polgar et Douanglattana Souphanthong (promotion 2003-2004).

Schéma XML pour le laotien

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!-- XML Schema for the Papillon Lao lexies
Includes the language specific elements (parts-of-speech, etc.)
Namespace: http://www-clips.imag.fr/geta/services/dml
Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd
$Author: Vincent $ Vincent BERMENT Vincent.Berment@imag.fr
$Date: 2003/03/15 06:25:32 $
$Revision: 1.0 $ -->
<schema
  xmlns:d='http://www-clips.imag.fr/geta/services/dml'
  xmlns='http://www.w3.org/2001/XMLSchema'
  xmlns:xlink='http://www.w3.org/1999/xlink'
  targetNamespace='http://www-clips.imag.fr/geta/services/dml'>
<annotation>
  <documentation xml:lang="en">
    XML Schema for the Papillon Lao lexies.
    Includes the language specific elements (parts-of-speech, etc.).
    Namespace: http://www-clips.imag.fr/geta/services/dml.
    Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd.
    Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd.
  </documentation>
</annotation>
<!-------- Elements of the Papillon common schema redefinition =======-->
<redefine
 schemaLocation="http://www-clips.imag.fr/geta/services/dml/papillon.xsd">
 <!— ** The parts-of-speech we will use for the Lao language ** -->
 <simpleType name='posType'>
  <restriction base='d:posType'>
    <!-- Nom commun -->
        <enumeration value="nc"/>
    <!-- Locution nominale -->
        <enumeration value="locn"/>
    <!-- Nom propre -->
        <enumeration value="np"/>
    <!-- ... (the 57 POS described in the appendix 1) -->
   </restriction>
  </simpleType>
 <!— ** The levels of language we will use for the Lao language ** -->
 <simpleType name="leveloflanguageType">
   <restriction base="d:leveloflanguageType">
    <enumeration value="courant" />
    <enumeration value="respectueux" />
    <enumeration value="familier" />
    <enumeration value="argotique" />
    <enumeration value="spécialisé" />
    <enumeration value="recherché" />
    <enumeration value="bonze" />
    <enumeration value="royal" />
    <enumeration value="littéraire" />
    <enumeration value="parlé" />
    <enumeration value="archaïque" />
   </restriction>
  </simpleType>
</redefine>
</schema>
```

II.3.2.4.4 Portage vers Papillon

Pour porter la base LaoDict vers la base Papillon, les articles doivent être transformés depuis leur format d'origine — encodage « Montaigne Lao »[1] des caractères et stockage avec MySQL — vers le format XML / Unicode de Papillon. Ce portage est prévu après la fin de l'année universitaire 2003-2004, attendant en effet qu'un nombre suffisamment significatif d'articles ait été saisi. La procédure de portage d'un dictionnaire XML dans la base Papillon est décrite dans [Mangeot 2002].

Pour pouvoir interroger la base Papillon de façon satisfaisante, la fonction de mise en forme canonique existant actuellement dans LaoLex devra être intégrée au site Papillon. En effet, la saisie des mots laotiens ne se fait pas forcément de manière unique (voir II.1.2.2). Par exemple, /năŋsŭ:/, le mot laotien pour *lettre*, peut s'écrire soit ໜັງສື soit ຫນັງສື si bien que l'une des formes doit être considérée comme canonique, par exemple ຫນັງສື, et l'autre, ໜັງສື, devra être transformée en ຫນັງສື qui sera la forme stockée dans la base de données Papillon (voir le chapitre II.2.4.1). Ce logiciel de mise en forme canonique peut être dérivé de celui qui est utilisé pour les requêtes dans LaoLex (voir II.3.1.2.4), le passage de Windows à Linux nécessitant principalement de retravailler la gestion mémoire (type HANDLE sous Windows) et les fonctions appelées par le système (fonctions CALLBACK).

De même, si l'on souhaite pouvoir ajouter de nouveaux mots dans la base directement via le site Papillon, les techniques décrites au chapitre II.3.1.2.4 devront être mises en œuvre pour que la forme canonique des mots soit stockée dans la base et que l'affichage puisse ultérieurement se faire dans l'ordre lexicographique laotien.

II.3.2.5 RESSOURCES LINGUISTIQUES EN PROJET

II.3.2.5.1 Informatisation du dictionnaire français-laotien de Marc Reinhorn

Les travaux publiés par Marc Reinhorn, professeur de laotien à l'INALCO de 1948 à 1985, incluent un dictionnaire laotien-français [Reinhorn 1970] et une grammaire [Reinhorn 1975], tous deux saisis à l'aide d'une machine à écrire. Il existe cependant un autre ouvrage cher au cœur de cet auteur et sur lequel il a travaillé jusqu'à ses derniers jours, son dictionnaire français-laotien ; certes resté inachevé, il contient déjà plus de mille deux cents pages et est rédigé sur un ordinateur. Ce dictionnaire contient, en particulier la traduction de nombreux exemples et expressions du français, l'auteur s'étant basé sur le *Petit Robert* pour établir la liste de ses articles.

La transformation de ce dictionnaire — qui est complet jusqu'à la lettre 't' — en base de données lexicale serait un apport considérable pour un système d'aide à la traduction français-laotien. Elle est envisagée en 2004, de même que la saisie des derniers articles du dictionnaire (de 't' à 'z') à partir des fiches papier laissées par Marc Reinhorn.

II.3.2.5.2 Corpus en laotien et textes alignés laotien-français

Les corpus en laotien ou bilingues — laotien + français, laotien + russe ou laotien + anglais — sont difficiles à trouver en dehors du Laos. Nous avons conservé quelques centaines de courriers électroniques qui fournissent principalement un corpus en laotien et présentent plusieurs inconvénients :

- ⇨ ils sont saisis avec des encodages multiples,
- ⇨ ils sont, le plus souvent, dans l'ancienne orthographe (diaspora),
- ⇨ ils comportent souvent des fautes de frappe.

Ces documents pourront être standardisés pour être utilisés comme corpus de test mais une mise à disposition de documents professionnels réalisés au Laos (documents journalistiques, douaniers, de chancellerie, législatifs, administratifs...) serait un apport plus efficace.

[1] Encodage non standardisé défini par la police du même nom.

II.3.3 Utilisation d'Unicode : *LaoUniKey, LaoWord 4 et navigateurs Internet*

II.3.3.1 LAOUNIKEY

LaoUnikey utilise la technologie des *hooks* mentionnée au chapitre II.2.1.2.3 pour intercepter les messages Windows. Une technique de « copier-coller » permet alors de transférer à l'application le caractère intercepté par le *hook*. Cela se fait en copiant un tampon mémoire dans le presse-papiers en précisant qu'il contient de l'Unicode.

```
SetClipboardData(CF_UNICODETEXT,hMem) ;
```

Seule la copie nécessite de préciser que le texte est en Unicode, le collage étant réalisé en simulant un « Contrôle V », raccourci du collage dérivé du Macintosh et extrêmement répandu sous Windows.

Notons qu'avec une technologie non-Unicode sur huit bits, il n'est pas nécessaire de recourir à cette méthode de « copier-coller » pour transférer un caractère à l'application. Il suffit de substituer le caractère souhaité au caractère reçu par le hook dans le message WM_CHAR et de laisser ce message atteindre l'application, avec son nouveau paramètre. Mais cette technique ne fonctionne pas correctement en Unicode avec les premières versions Unicode de Windows (voir la remarque de [Wright et al. 2001] dans le chapitre II.2.1.2.3).

II.3.3.2 LAOWORD 4

LaoWord 4 utilise aussi la technique des *hooks*. Contrairement à LaoUniKey, le transfert des caractères Unicode vers l'application s'appuie sur une interface spécifique à Word reposant sur le modèle COM (Component Object Model) et non sur ce principe du « copier-coller ». Cette interface de programmation offre un accès direct aux objets de Word. L'insertion d'un caractère utilise la méthode *TypeText* de l'objet *Selection* de Word. Son prototype prenant en paramètre une chaîne dans le format BSTR spécifique à COM (chaîne de caractères Unicode préfixée par sa longueur), il suffit de convertir le caractère Unicode grâce à l'appel suivant puis d'appeler la méthode *TypeText* pour que Word insère le caractère :

```
bstrCarUnicode=::SysAllocString((wchar_t*)CarUnicode);
```

Comme pour LaoUniKey, la modification du code caractère des messages WM_CHAR suffit pour transférer un caractère à Word lorsque l'on n'est pas en Unicode et il n'est alors pas nécessaire de recourir à COM pour cela. D'ailleurs, les versions de LaoWord antérieures à LaoWord 4 n'utilisaient pas COM. Ce modèle objet n'étant arrivé qu'avec Word 97 et LaoWord fonctionnant avec les versions 6 et 95 de Word, la méthode de substitution des codes WM_CHAR était utilisée.

II.3.3.3 LAOUNIWEB ET NAVIGATEURS INTERNET

La technique présentée au chapitre II.2.1.2.4 pour créer un clavier virtuel indépendant de la plate-forme et utilisable dans un navigateur quelconque s'applique aussi en Unicode : c'est le logiciel LaoUniWeb (voir I.3.3). Il suffit de mettre la représentation UTF-8 du caractère Unicode souhaité à la place du caractère substitué.

```
case "k" : field =  field + "à°  " ;
```

Dans le code proposé ci-dessus, les trois caractères *à°* (dont l'un n'est pas imprimable) correspondent à l'encodage de la lettre laotienne ໂ en UTF-8 ([k], code 0E80), lettre qui est accessible via la touche 'k' du clavier (*case "k"*).

UTF-8 est un encodage proposé par le standard Unicode pour transformer les caractères Unicode (généralement sur seize bits : UTF-16) en caractères sur huit bits pouvant être transmis par des moyens de transmission utilisant des caractères ou via des interfaces avec les *devices* Linux. Le nombre des caractères nécessaires en UTF-8 pour chaque caractère Unicode varie de un à quatre selon le système d'écriture, par exemple un pour les caractères ASCII (première ligne du tableau ci-dessous) et trois pour les caractères laotiens (troisième ligne).

Valeur	UTF-16	1er octet	2e octet	3e octet	4e octet
00000000 0xxxxxxx	00000000 0xxxxxxx	0xxxxxxx			
00000yyy yyxxxxxx	00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
zzzzyyyy yyxxxxxx	zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
000uuuuu zzzzyyyy yyxxxxxx	110110ww wwzzzzyy 110111yy yyxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

Nota : uuuuu = wwww + 1.

Figure 39 : Relation entre UTF-16 et UTF-8

Exemple : Nous avons ໂ = 0E80 = 0000 1110 1000 0000.

Cela donne :

- caractère 1 = 1110 0000 = E0 = 'à' en Latin 1,
- caractère 2 = 1011 1010 = BA = '°' en Latin 1,
- caractère 3 = 1000 0000 = 80 = ' ' en Latin 1.

II.3.4 Application simple : *aide à la traduction thaï-laotien*

Une application simple nous a été demandée par un responsable laotien pendant la conférence *PAN All Partner's* organisée à Vientiane par l'IDRC en mars 2003. Cette application illustre le fait que des traitements très simples et rapides à réaliser peuvent être très utiles. Le besoin vient de ce que la population laotienne comprend assez facilement le thaï mais ne peut pas toujours le lire. Les systèmes d'écriture et les langues étant voisins, il est possible de transformer un texte thaï écrit en écriture thaïe en un texte thaï écrit en écriture laotienne, texte que les laotiens comprennent même si la transcription ne donne pas des résultats parfaits. La correspondance entre les caractères thaïs (encodage TIS 620 2529/2533) et les caractères laotiens codés dans une police laotienne donnée (la police Montaigne Lao, en l'occurrence) est donnée ci-dessous. Cette aide à la traduction est disponible en ligne à l'adresse http://sabaidi.imag.fr/transcription/. Elle est réalisée avec trois petits fichiers PHP et un fichier HTML pour la page d'accueil. Les caractères thaïs sont à gauche, les laotiens à droite.

ก	ກ	ฎ	ດ	ป	ປ	ส	ສ	ู	ູ
ข	ຂ	ฏ	ຕ	ผ	ຜ	ห	ຫ	เ	ເ
ฃ	ຂ	ฐ	ຖ	ฝ	ຝ	พ	ພ	็	ັ
ค	ຄ	ฑ	ທ	พ	ພ	อ	ອ	แ	ແ
ฅ	ຄ	ฒ	ທ	ฟ	ຟ	ฮ	ຮ	โ	ໂ
ฆ	ຄ	ณ	ນ	ภ	ພ	ะ	ะ	ไ	ໄ
ง	ງ	ด	ດ	ม	ມ	ั	ັ	ใ	ໃ
จ	ຈ	ต	ຕ	ย	ຍ	า	າ	ำ	ໍາ
ฉ	ສ	ถ	ຖ	ร	ຣ	ิ	ິ	ฤ	ຣຶ
ช	ຊ	ท	ທ	ล	ລ	ี	ີ	ฤๅ	ຣຶໆ
ซ	ຊ	ธ	ທ	ว	ວ	ึ	ຶ	ฦ	ລຶ
ฌ	ຊ	น	ນ	ศ	ສ	ื	ື	ฦๅ	ລຶໆ
ญ	ຍ	บ	ບ	ษ	ສ	ุ	ຸ

Figure 40 : Tableau de correspondance thaï-laotien pour une aide à la traduction

CONCLUSION DE LA DEUXIÈME PARTIE : BILAN DE L'EXPÉRIENCE DU LAOTIEN

Au début des années 1990, il n'existait guère, pour utiliser la langue laotienne sur un ordinateur, que quelques polices de caractères peu pratiques pour Windows et Macintosh, ainsi qu'un traitement de texte pour DOS dérivé du traitement de texte thaï CU-Writer de l'université Chulalongkorn. L'arrivée de Windows 3.1 et de ses polices TrueType au printemps 1992 a rapidement entraîné l'éclosion de dizaines puis de centaines de polices laotiennes. Outre leur qualité esthétique, meilleure que celle des polices *bitmaps* et vectorielles, les polices TrueType offrent la possibilité de superposer des caractères grâce à leur procédé d'espacement de caractère ABC ([Bür et Bauder 1992]), ce qui permettait enfin d'écrire en laotien dans les traitements de texte du commerce. La saisie suivait cependant la logique d'encodage choisie pour la police, et variait donc d'une police à l'autre.

Des logiciels ont alors été développés[1] pour aider l'utilisation de ces polices et apporter de nouveaux services : claviers virtuels, sélection syllabique, tri lexicographique, compatibilité entre encodages, fonctions de mise en forme du texte, transcriptions phonétiques. Dans cette première phase, nous avons travaillé seul, recherchant une base performante dans laquelle implanter les fonctions laotiennes. Ce fut WordPad pour LaoPad, puis Word pour LaoWord.

Dans une deuxième phase, nous avons cherché à utiliser davantage les idées et méthodes présentées dans la première partie. Ainsi, nous avons développé LaoLex, un site Internet d'aide à la traduction et de construction de dictionnaire laotien-français permettant un travail coopératif auquel pouvait participer, en particulier, la diaspora laotienne. Ce site a permis la réalisation d'une vingtaine de dictionnaires personnels. Ils sont en grande partie la réalisation des étudiants en laotien aux Langues O'. Quelques autres personnes ont néanmoins apporté une contribution. Un premier dictionnaire commun d'une centaine d'articles (lexies) a été dérivé des dictionnaires personnels. Il est utilisé pour les traductions mot à mot de LaoLex.

Nous avons essayé d'élargir la participation à tous les visiteurs du site LaoLex grâce au forum de discussion *FrancoLao*, entretenu par Houmphanh Thongvilu depuis plus de dix ans, parmi plusieurs autres forums laotiens. Un forum spécifique — ປ໋ກລາວ (*Pak Lao*) — a été créé, mais le succès n'a pas été au rendez-vous. Un temps insuffisant consacré à l'animation, un concept de lexie trop complexe pour des visiteurs non spécialistes et pourtant intéressés, un dictionnaire de départ trop petit : peu de lexies ont été ajoutées par les participants à ce forum.

Plusieurs évolutions du site permettraient d'améliorer la participation, en qualité et en quantité :

- ⇨ faciliter la compréhension des principes (lexies, parties du discours...) grâce à des exemples,
- ⇨ permettre d'ajouter facilement un mot trouvé inconnu lors d'une traduction,
- ⇨ utiliser les textes bilingues du site comme base pour des mémoires de traduction,
- ⇨ rendre le site plus agréable et intuitif (ergonomie).

Le format des dictionnaires créés par LaoLex a été prévu pour pouvoir facilement les exporter vers la base de données lexicales Papillon. La participation à ce projet multilingue à pivot permettra, à terme, d'obtenir des traductions de mots laotiens dans plusieurs langues et inversement, sans avoir à réaliser tous les couples de langues correspondants. L'intégration complète du laotien dans Papillon (fonction permettant le tri et l'interrogation de la base...) et son interfaçage avec LaoLex permettront ainsi d'offrir un service complémentaire de traduction multilingue aux visiteurs du site.

La diffusion sous licence GPL[2] du logiciel LaoUniKey (saisie du laotien en Unicode) a aussi été expérimentée. Bien qu'aucune version modifiée de LaoUniKey n'ait encore vu le jour, plusieurs personnes nous ont déjà donné des idées d'amélioration et l'une d'elle nous a transmis du code à intégrer dans le logiciel.

[1] Essentiellement par deux non Laotiens : John Durdin, un Australien vivant au Laos et nous-même. Plus récemment, Anousak Souphavanh, un Américain d'origine laotienne, a entrepris la localisation de Linux/KDE.
[2] http://www.gnu.org/licenses/licenses.fr.html#GPL.

Nous avons établi le tableau d'évaluation du niveau d'informatisation de la langue laotienne en prenant des valeurs de criticité évaluée à partir de multiples contacts avec des utilisateurs depuis cinq ans. Il indique que le laotien, qui avait un indice-σ de l'ordre de 4/20 au début des années 1990, est encore assez faiblement doté mais néanmoins proche de la catégorie des langues-μ, catégorie des langues moyennement dotées. Dans la pratique, le laotien est maintenant couramment utilisé pour la saisie de texte en local et dans les échanges par courrier électronique.

	Services / ressources	Criticité (0 à 10)	Note (/20)	Note pondérée
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	16	160
	Recherche et remplacement	6	12	72
	Sélection du texte	6	16	96
	Tri lexicographique	8	16	128
	Correction orthographique	4	0	0
	Correction grammaticale	2	0	0
	Correction stylistique	1	0	0
Traitement de l'oral				
	Synthèse vocale	4	0	0
	Reconnaissance de la parole	4	0	0
Traduction				
	Traduction automatisée	8	8	64
ROC				
	Reconnaissance optique de caractères	6	0	0
Ressources				
	Dictionnaire bilingue	8	4	32
	Dictionnaire d'usage	5	0	0
Total		**82**		**712**
Moyenne (/20)				**712 / 82 = 8,68**

Figure 41 : Tableau d'évaluation du niveau d'informatisation pour le laotien (début 2004)

Outre les améliorations de LaoLex citées précédemment, de nombreuses idées restent à mettre en œuvre et à expérimenter. En particulier, la récupération du dictionnaire français-laotien de Marc Reinhorn ou la participation à cet autre grand projet multilingue qu'est UNL permettraient d'ajouter des services de traduction mot à mot français-laotien et de traduction automatique depuis toutes des langues d'UNL vers le laotien, cela ne nécessitant que la réalisation d'un déconvertisseur UNL-laotien. La réalisation d'un tel déconvertisseur pourrait bénéficier du prototype anglais-thaï développé en Ariane, le système de traduction automatique du GETA.

Enfin, nous souhaitons réaliser un générateur générique de sites du type LaoLex pour des langues à écriture non segmentée (comme le laotien) avec divers services incluant un clavier virtuel *Open Source*, un service d'aide à la traduction et un service de construction de base lexicale. Les principales difficultés sont la standardisation, la segmentation, le tri et la phonétisation.

TROISIÈME PARTIE

INFORMATISATION D’UN GROUPE DE LANGUES

III. MÉTHODES POUR L'INFORMATISATION DE GROUPES DE LANGUES PEU DOTÉES

III.1 RÉFLEXION SUR L'EXPÉRIENCE DU LAOTIEN : RÉUTILISABILITÉ ET ÉCONOMIES D'ÉCHELLE

III.1.1 Technique de réutilisation de code appliquée et bilan quantitatif

III.1.1.1 ANALYSE DU TEMPS DE DÉVELOPPEMENT DES OUTILS POUR LA LANGUE LAOTIENNE

Le tableau ci-dessous présente une répartition analytique du temps consacré au développement des logiciels Tallao, LaoPad, LaoWord et LaoUniKey (conception, codage, validation). Nous y avons distingué les heures passées sur les développements liés au traitement des langues — développements « linguistiques » — de celles passées sur les aspects plus généraux — développements « généraux » — de manière à mettre en évidence la réutilisation du code[1]. Le temps consacré à la documentation est donné à titre indicatif, ce service étant une partie essentielle pour la diffusion dans le grand public.

		Tallao 3.2	LaoPad 1.0	LaoWord 3.01	LaoUniKey 1.0	LaoWord 4.0
Développements "généraux"		**9%**	**60%**	**67%**	**86%**	**89%**
	SDK MDI (gestion complète : menus, fenêtres...)	100				
	MFC (interfaçage / héritage, menu, liste déroulante des claviers)		300			
	Word (interfaçage, barre d'outils, fenêtres pour configuration, polices, tri, transcription, mise en forme...)			800	R	R+400
	Technologie LaoUniKey (copier-coller)				60	
Développements "linguistiques"		**91%**	**0%**	**8%**	**14%**	**0%**
G	Algorithmes de sélection	500	R	R		R
G	Algorithmes de transcriptions et tri lexicographique	100		R		R
G	Algorithmes divers (réglage de hauteurs de caractères, ligatures, insertion d'espaces de largeur nulle)			100		R
G	Gestion pivot (saisie & cht polices)	50	R	R	R	R
G	Gestion du dictionnaire	50		R		R
S	Modèle syllabique	150	R	R		R
S	Tableaux police-pivot	100	R	R	R+10	R
S	Dictionnaire	20		R		R
S	Transcription phonétique	30		R		R
Documentation utilisateur		**0%**	**40%**	**25%**	**0%**	**11%**
	LaoPad		200			
	LaoWord			300		50
Total		1100	500	1200	70	450

Figure 42 : Temps consacré à des développements pour la langue laotienne (heures)

III.1.1.2 RÉUTILISATION DU CODE « LINGUISTIQUE » RÉALISÉ POUR TALLAO

Le langage de programmation C++ a été utilisé pour les quatre logiciels, ce qui a facilité la réutilisation du **code « linguistique »** réalisé pour Tallao. Les rares développements complémentaires dans ce domaine sont, soit des fonctions nouvelles (fonctions de mise en forme de LaoWord 3.01), soit des tables de transcodage pour de nouveaux encodages de police, en particulier Unicode (LaoUniKey). Cela montre combien il est intéressant de réaliser les fonctions « linguistiques » indépendamment de l'environnement qui les utilise et dans un langage de programmation portable.

L'économie[2] réalisée sur les développements linguistiques de Tallao grâce à la réutilisation de code approche les 100 % sans les atteindre, des améliorations difficilement mesurables et non indiquées ci-dessus ayant été réalisées après Tallao. Nous estimons que l'économie réelle est d'environ **98 %** (environ 20 heures, hors nouvelles fonctions).

[1] La réutilisation d'un code source antérieur est indiquée par la lettre R (le tableau classe les logiciels dans l'ordre chronologique). Les lettres G et S indiquent les développements applicables à plusieurs langues (**g**énériques) ou **s**pécifiques à une langue particulière. Elles seront expliquées plus en détail au chapitre III.1.3.2.

[2] L'économie est définie ici par la formule : $(T_S - T_A) / T_S$, avec T_S = temps de développement sans réutilisation et T_A = temps de développement avec réutilisation. S'agissant d'un module isolé, l'économie est égale au taux de réutilisation.

III.1.1.3 RÉUTILISATION DU CODE « GÉNÉRAL » POUR PLUSIEURS LANGUES

Le temps nécessaire pour réaliser le **code « général »** mérite aussi réflexion. Nous estimons que ce temps est en moyenne de cent à deux cents heures. Quelques centaines d'heures supplémentaires peuvent être nécessaires pour maîtriser l'environnement lorsqu'il est nouveau, ce qui a été le cas des MFC (*Microsoft Foundation Classes*) avec LaoPad, alors qu'avec Tallao nous repartions d'un environnement connu. L'effort plus important que nous voyons apparaître dans le tableau pour LaoWord 3.01 et 4.0 est dû à la plus grande complexité de l'interfaçage avec Word, l'interaction avec ce logiciel se faisant à travers deux API : CAPI qui fonctionne avec toutes les versions existantes de Word mais qui est assez limitée, en particulier pour la gestion des événements souris (API utilisée dans LaoWord 3.01), et COM qui n'est disponible qu'à partir de Word 8 (API utilisée en complément de CAPI dans LaoWord 4.0).

LaoUniKey est le logiciel Open Source présenté au chapitre II.3.3 et permettant de saisir du texte en laotien dans une fenêtre d'édition quelconque pourvu qu'elle soit compatible avec Unicode. La partie « générale » de ce logiciel a bénéficié de l'acquis de la technologie des *hooks* développée pour LaoWord (voir II.2.1.2.3). L'adaptation de cet acquis à une fenêtre d'édition quelconque n'a nécessité que soixante heures, dont :

- quarante ont été consacrées à implémenter le mécanisme de « copier-coller » (non utilisé précédemment),
- vingt ont été consacrées à l'interface homme-machine, elle aussi originale.

Nous estimons que le développement de la partie « générale » de LaoUniKey à partir de zéro aurait nécessité environ deux cents heures, ce qui représente une économie de **70 %** (60 heures, soit 140 heures d'économie sur 200) obtenue grâce à la réutilisation de LaoWord.

En mars 2003, lors d'une conférence organisée par l'IDRC, Mustafa Jabbar — un chef d'entreprise bangladeshi présent à la conférence — nous a demandé d'adapter LaoUniKey à l'écriture bengali et au clavier Bijoy, le clavier le plus répandu au Bangladesh. Nous avons appelé ce logiciel BanglaUniKey.

Le coût en heures de ce développement a été de **quinze heures** :

- ⇨ **cinq heures** pour rendre générique la **partie « générale »** de LaoUniKey,
- ⇨ cinq heures de travail à deux (Mustafa Jabbar et moi-même) pour définir la table de correspondance clavier-glyphe, soit **dix heures** au total pour la **partie « linguistique »**.

L'économie réalisée sur la partie générale est de 91,7 % (5 heures, soit 55 heures d'économie sur 60) par rapport à LaoUniKey qui lui-même avait bénéficié de l'acquis sur LaoWord. L'économie globale est donc de **97,5 %** (5 heures, soit 195 heures d'économie par rapport aux 200 estimées pour un développement de la partie générale de BanglaUniKey en partant de zéro).

À la suite de cette expérience, nous avons porté dans LaoWord la table de correspondance clavier-code obtenue dans BanglaUniKey, et créé un nouveau logiciel : BanglaWord. L'unique fonction de BanglaWord est la saisie du bengali en Unicode avec un clavier compatible Bijoy. **Trois heures** ont été suffisantes pour cette opération dont la plus grosse partie a consisté à enlever de LaoWord les fonctions inutiles.

L'économie réalisée est ici de **98 %** (3 heures, soit 147 heures d'économie par rapport au temps nécessaire pour réaliser la partie générale de la saisie Unicode dans Word, estimé à 150 heures).

III.1.2 Économies d'échelle réalisables

Nous avons vu dans le chapitre précédent que des gains de temps importants pouvaient être réalisés grâce à la réutilisation de code. En particulier, cela peut être le cas :

- ⇨ lorsque du code linguistique développé pour une application donnée est intégré dans une plate-forme d'accueil différente (par exemple : Tallao → LaoWord),
- ⇨ lorsqu'une plate-forme dotée d'un complément adapté reçoit des compléments linguistiques pour d'autres langues ou écritures (par exemple : LaoUniKey → BanglaUniKey),
- ⇨ lorsque l'on développe une nouvelle plate-forme d'accueil et que l'on récupère le code général réalisé pour une autre plate-forme (par exemple : LaoWord → LaoUniKey).

Le tableau ci-dessous récapitule les économies constatées dans nos développements (CL = Code Linguistique et CG = Code Général).

	Code de LaoPad et de LaoWord		**Code de LaoUniKey**		**Code de BanglaUniKey**		**Code de BanglaWord**	
Code réutilisé	Code linguistique de Tallao		Codes général et linguistique de LaoWord		Codes général et linguistique de LaoUniKey		Code général de LaoWord + code linguistique de LaoUniKey	
Heures	CL	CG	CL	CG	CL	CG	CL	CG
- Temps passé	~20	300/800	10	60	10	5	0	3
- Sur total estimé	1000	300/800	150	200	150	200	150	150
- Économie	980	0	140	140	140	195	150	147
Économie (%)	98 %	0 %	93,3 %	70 %	93,3 %	97,5 %	100 %	98 %

Figure 43 : Économies réalisées grâce à la réutilisation de code

Supposons maintenant que l'on veuille informatiser **nL langues** dans **nE environnements**. Si $\mathbf{tL}_i$ est le temps nécessaire pour réaliser les développements « linguistiques » de la langue i et si $\mathbf{tE}_j$ est le temps nécessaire pour réaliser les développements « généraux » dans l'environnement j, le temps de développement **sans réutilisation** sera de $T_S = \Sigma_{i,j} (tL_i + tE_j)$, soit :

$$\mathbf{T_S = \Sigma_i\, tL_i * nE + \Sigma_j\, tE_j * nL}.$$

Si $\forall$ i, $tL_i = tL$ et si $\forall$ j, $tE_j = tE$, cette formule devient :

$$\mathbf{T_S = nL*tL*nE + nE*tE*nL = nL*nE*(tL+tE)}.$$

Ce temps correspond, par exemple, à des travaux faits sans coordination, chaque groupe de développement réalisant son logiciel pour sa langue, isolément. Le regroupement de ces efforts dispersés, par exemple au sein d'un projet de type GNU, permettrait de réaliser des gains de temps importants.

Avec un taux de réutilisation des modules de rL % pour le code linguistique et de rE % pour le code général, le temps de développement T_A ne sera que de[1] :

$$T_A = \Sigma_i tL_i+(nE-1)*(1-rL)*\Sigma_i tL_i + \Sigma_j tE_j+(nL-1)*(1-rE)*\Sigma_j tE_j.$$

En réarrangeant les termes, on obtient :

$$\mathbf{T_A = \Sigma_i tL_i*(nE-rL*(nE-1))+\Sigma_j tE_j*(nL-rE*(nL-1))}.$$

Si $\forall$ i, tL_i = tL et si $\forall$ j, tE_j = tE, la formule précédente devient :

$$\mathbf{T_A = nL*tL*(nE-rL*(nE-1))+nE*tE*(nL-rE*(nL-1))}.$$

Le **gain de réutilisation** vaut ainsi :

$$\mathbf{T_S - T_A = nL*tL*rL*(nE-1)+nE*tE*rE*(nL-1)},$$

et **l'économie relative**[2] :

$$\mathbf{(T_S - T_A) / T_S = (nL*tL*rL*(nE-1)+nE*tE*rE*(nL-1)) / (nL*nE*(tL+tE))}.$$

Exemple

En donnant aux différents paramètres des ordres de grandeur tirés de notre expérience :

tL_i = 1000 heures, tE_j = 500 heures, rE = rL = 95 %,

et pour, par exemple, quatre environnements différents, on obtient une économie globale de :

$$\mathbf{Économie\ globale_{95\%} = 0,95*(5000*nL - 2000) / (6000*nL)}.$$

Cela donne les résultats suivants :

Nombre de langues	Temps sans réutilisation	Temps avec réutilisation	Économie
1	6 000 h	3 150 h	47,50 %
2	12 000 h	4 400 h	63,33 %
3	18 000 h	5 650 h	68,61 %
4	24 000 h	6 900 h	71,25 %
5	30 000 h	8 150 h	72,83 %
10	60 000 h	14 400 h	76,00 %
100	600 000 h	126 900 h	78,85 %
1000	6 000 000 h	1 251 900 h	79,14 %

Avec une réutilisation totale des modules[3] (rE = rL = 100 %), on obtient une économie globale de :

$$\mathbf{Économie\ globale_{100\%} = (5000*nL - 2000) / (6000*nL)}.$$

Cela donne les résultats suivants :

Nombre de langues	Temps sans réutilisation	Temps avec réutilisation	Économie
1	6 000 h	3 000 h	50,00 %
2	12 000 h	4 000 h	66,67 %
3	18 000 h	5 000 h	72,22 %
4	24 000 h	6 000 h	75,00 %
5	30 000 h	7 000 h	76,67 %
10	60 000 h	12 000 h	80,00 %
100	600 000 h	102 000 h	83,00 %
1000	6 000 000 h	1 002 000 h	83,30 %

[1] On développe chaque langue dans un environnement particulier et chaque environnement avec une langue particulière. On adapte ensuite chaque langue aux *nE–1* environnements restants (réutilisation de rL %) et chaque environnement aux *nL–1* langues restantes (réutilisation de rE %).

[2] Formule *économie* = $(T_S - T_A) / T_S$ vue précédemment.

[3] C'est ce que l'on obtiendrait en définissant une interface standard entre les modules linguistique et général.

III.1.3 Méthodologie d'architecture issue de l'expérience sur le laotien

III.1.3.1 SÉPARATION DES PARTIES « LINGUISTIQUE » ET « GÉNÉRALE » DES DÉVELOPPEMENTS

L'analyse précédente[1] nous suggère d'orienter l'effort vers le développement de logiciels, ouverts ou non, mais constitués d'une **partie « linguistique »** et d'une **partie « générale »** distinctes. Un autre point très important est **l'emploi de logiciels de base réalisés par les grands éditeurs de logiciels**, commerciaux ou libres, et offrant une possibilité d'adaptation. La réalisation et la maintenance de ces logiciels (traitements de texte, navigateurs...) étant en effet très coûteuses, leur utilisation permet de maintenir une offre pour les langues-π et μ au niveau des standards de qualité des langues-τ. Elle présente l'inconvénient d'être dépendante des évolutions, non maîtrisables, du logiciel de base. Cet inconvénient est cependant mineur si plusieurs logiciels de base sont utilisés concurremment.

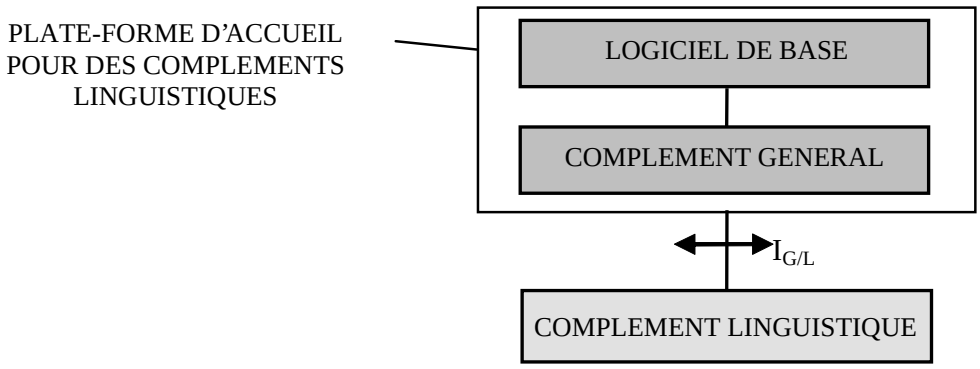


Figure 44 : Séparation des parties « linguistique » et « générale » des développements

La séparation des parties « linguistique » et « générale » entraîne le besoin de définir formellement une interface entre elles. Cette interface permet, en principe, d'atteindre une réutilisation totale (rL = rE = 100 %). En prenant l'exemple d'un traitement de texte, cette interface $I_{G/L}$ entre le complément général et le complément linguistique peut être définie par des règles de compatibilité incluant, en particulier, le langage de programmation et des zones pour insérer du code, par exemple :

- ⇨ une zone de code paramétrable pour :
  - o menus,
  - o barres d'outils,

Exemple tiré de LaoWord (logiciel de base = Microsoft Word) :
Après la création de la barre d'outils de LaoWord (voir le chapitre I.3.2) :

```
CAPIAddToolbar(0,(unsigned char*)"LaoWord 4.0");
AjouterBouton(DocID,(unsigned char*)"LaoWord 4.0",
      1,(unsigned char*)"Options",IconeLaoWordInactif);
AjouterBouton(DocID,(unsigned char*)"LaoWord 4.0",
      2,(unsigned char*)"ChangerLaPolice",IconeChangerPolice);
AjouterBouton(DocID,(unsigned char*)"LaoWord 4.0",
      3,(unsigned char*)"TrierLeTableau",IconeTrier);
...
AjouterBouton(DocID,(unsigned char*)"LaoWord 4.0",
      8,(unsigned char*)"AideLaoWord",IconeAideLaoWord);
```

[1] Les exemples LaoUniKey et LaoWord donnés aux chapitres précédents ont montré, en particulier, que la modification d'un fichier d'en-tête (fichier .h contenant le tableau de correspondance touche-code) suffisait pour créer un clavier virtuel pour le bengali. Noter qu'il faut aussi changer le nom du clavier dans l'interface de sélection du clavier et la documentation associée.

L'appui sur le bouton 1 provoque l'appel du code utilisateur :

```
extern "C" __declspec(dllexport) void Options() {
      code utilisateur pour l'appui sur le bouton 1 }
```

⇨ une zone d'accueil pour le code à exécuter sur événements :
- souris (plusieurs),
- clavier (plusieurs),
- horloge...

Exemple (indépendant du logiciel de base) :
Après le lancement d'un *hook* (initialisation) :

```
ThreadID=GetCurrentThreadId();
SetWindowsHookEx(WH_GETMESSAGE,
      (HOOKPROC)GetProcAddress(hInstance,"_HookProcGetMessage@12"),
      hInstance,(DWORD)ThreadID)))
```

Une action souris ou clavier, ou encore une fin de temporisation, provoque l'appel du code utilisateur :

```
extern "C" LRESULT __declspec(dllexport) CALLBACK HookProcGetMessage
                  (int code,WPARAM wParam,LPARAM lParam)
{
      unsigned int MessageWindowRecu=(*(MSG*)lParam).message;
      if (MessageWindowRecu==WM_LBUTTONDBLCLK) {
            code utilisateur pour le double-clic souris }
      if (MessageWindowRecu==WM_MOUSEMOVE) {
            code utilisateur pour un mouvement de la souris }
      if (MessageWindowRecu==WM_CHAR) {
            code utilisateur pour l'appui sur une touche du clavier }
      if (MessageWindowRecu==WM_TIMER) {
            code utilisateur pour une fin de temporisation }
}
```

L'intégration dans $I_{G/L}$ des principales API utiles (récupération d'informations sur le texte, la sélection..., affichage d'un message...) leur permettraient de présenter une syntaxe unifiée, (presque) indépendante du logiciel de base.

Les avantages de la séparation entre les codes « linguistique » et « général » sont les suivants (exemples donnés pour un traitement de texte) :

⇨ économie importante grâce à la réutilisation du code réalisé :
- la partie « linguistique » (clavier virtuel, sélection...) peut être utilisée dans des parties « générales » différentes (Word, OpenOffice.org...),
- la partie « générale » peut utiliser des parties « linguistiques » de différentes langues ou écritures,

⇨ spécialisation des contributeurs :
- la partie « linguistique » peut être traitée par des linguistes de manière plus autonome,
- la partie « générale » peut être développée par des spécialistes d'un environnement particulier (Word, OpenOffice.org...), voire de l'éditeur lui-même,

⇨ impact limité suite aux évolutions des logiciels de base :
- le complément général est le seul à nécessiter des évolutions,
- on peut supprimer la fonction linguistique si elle vient à se retrouver intégrée de manière satisfaisante au logiciel de base.

III.1.3.2 SÉPARATION DES PARTIES GÉNÉRIQUE ET SPÉCIFIQUE DES DÉVELOPPEMENTS « LINGUISTIQUES »

Nous pouvons entrer à l'intérieur de la partie « linguistique » et y distinguer encore deux types de développements selon qu'ils sont propres à une langue (ou à une écriture) ou plus généraux. Par exemple, la gestion d'un dictionnaire (création, modification et suppression d'une entrée) pourra être générique alors que les parties du discours et les articles eux-mêmes seront spécifiques. De même, l'algorithme général de sélection syllabique (voir le chapitre III.2.7) pourra être commun à toutes les écritures non segmentées (donc générique) alors que la fonction de reconnaissance des syllabes appelée par cet algorithme sera spécifique à une écriture. Dans le tableau du chapitre III.1.1.1, nous avons indiqué par un G les développements « linguistiques » génériques et par un S les développements « linguistiques » spécifiques (première colonne).

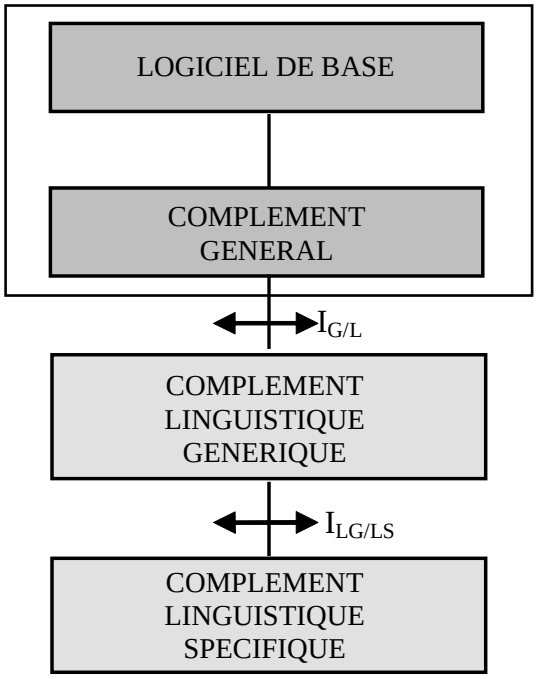


Figure 45 : Séparation des parties générique et spécifique des développements « linguistiques »

Nous sommes ainsi amené à définir une deuxième interface — $I_{LG/LS}$ — entre ces deux types de développements. En reprenant l'exemple d'un traitement de texte, cette interface entre les compléments « linguistiques » généraux et spécifiques peut être définie par :

- ⇨ un format pour le tableau de correspondance touche-code(s),
- ⇨ une fonction booléenne : Syllabe(chaîne),
- ⇨ une structure (XML...) contenant un dictionnaire...

Les avantages de cette séparation sont comparables à ceux cités précédemment pour l'interface $I_{G/L}$ (exemples donnés pour un traitement de texte) :

- ⇨ économie importante grâce à la réutilisation du code réalisé :
  - o la partie « spécifique » (modèle syllabique, dictionnaire...) peut être utilisée dans des parties « génériques » différentes,
  - o la partie « générique » (algorithme de sélection, gestion de dictionnaire...) peut utiliser des parties « spécifiques » de différentes langues ou écritures,
- ⇨ spécialisation des contributeurs :
  - o la partie « spécifique » peut être traitée par des linguistes de manière plus autonome,
  - o la partie « générique » peut être développée par des spécialistes d'un type de difficulté particulier (écritures non segmentées, lexicographie...).

III.2 EXEMPLE DE MISE EN ŒUVRE SUR LE GROUPE DES LANGUES À ÉCRITURE NON SEGMENTÉE

III.2.1 Introduction

Nous présentons dans ce chapitre les développements que nous avons réalisés pour apporter une réponse générique au problème de segmentation des systèmes d'écriture non segmentés d'Asie du Sud-Est. Nous nous sommes placé dans la perspective d'une informatisation suffisamment large — une vingtaine de systèmes d'écriture — pour étendre l'architecture modulaire proposée au chapitre précédent et y ajouter un module permettant aux linguistes de produire les compléments linguistiques spécifiques sans avoir besoin de programmer, cela grâce à des outils adaptés.

Ce principe **d'outils destinés à des activités linguistiques** — ou **outils linguiciels** — est général. Il peut apporter un gain de temps significatif lorsque plusieurs langues sont visées ainsi qu'une plus grande lisibilité et une facilité de mise au point dans tous les cas. Ces outils doivent être simples d'emploi et clairement documentés pour permettre à des personnes désirant contribuer à un projet d'informatisation de le faire sans qu'elles soient nécessairement informaticiennes.

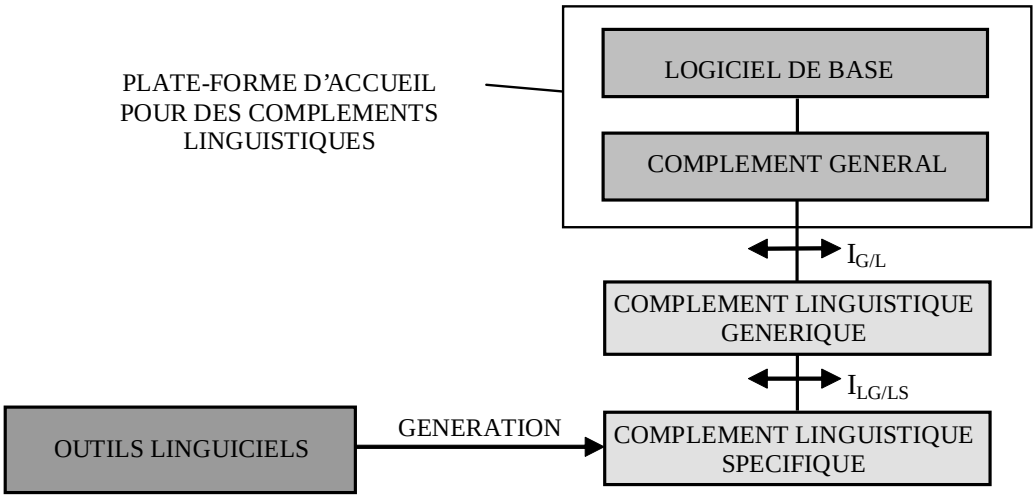


Figure 46 : Une architecture modulaire pour la segmentation de vingt systèmes d'écriture

L'outil réalisé pour la segmentation syllabique — Sylla — est présenté au chapitre III.2.6. Il permet à des personnes connaissant bien un système d'écriture de décrire simplement les règles de formation de ses syllabes, et ce dans toute langue utilisant un système d'écriture alphabétique. La description obtenue peut alors être transformée en un programme utilisable dans des algorithmes de segmentation, de tri ou de transcription phonétique.

III.2.2 Travaux existant sur les écritures non segmentées

III.2.2.1 SYSTÈMES D'ÉCRITURE NON SEGMENTÉS D'ASIE DU SUD-EST[1]

Les populations d'Asie du Sud-Est parlent un ensemble de langues provenant de familles aussi différentes que les familles austro-asiatique, austro-nésienne, hmong-mien, sino-tibétaine et taï-kadaï. Cette diversité de langues s'appuie sur différents systèmes d'écriture dérivés des écritures indiennes. Cet héritage est dû à la grande influence que la civilisation indienne a eue dans cette zone, au moins depuis le premier siècle de l'ère chrétienne.

Systèmes d'écriture en Asie du sud-est continentale

Dans cette région du monde, les écritures peuvent être regroupées en deux familles dérivant toutes deux de l'écriture pallava[2] et appelées du nom de leur prototype[3] (adapté de [Ferlus 1988]) :

- **khmer** – <u>sous-familles</u> : cham [4], khmer, laotien-thaï (lao, thaï), tai du nord-est (tai dam, tai khao, tai deng, tai yo, lai pao),
- **môn** – <u>sous-familles</u> : môn-birman (birman, môn), tai du nord-ouest[5] (dehong dai/tai nüa, khamti, shan, tai mau), tham (tham laotien et isan [6], lü, lanna, khün).

[1] Seules les écritures du Cambodge, du Laos, du Myanmar, de la Thaïlande, du Vietnam et du Yunnan sont prises en compte ici. Ainsi, des écritures considérées comme venant de l'Inde comme l'Ahom, l'Aiton et le Phake n'ont pas été incluses bien que proches de plusieurs écritures de la région. Le Khamti étant présent à la fois en Inde et au Myanmar, il a été inclus. De la même manière, plusieurs autres écritures n'ont pas été prises en compte car ne dérivant pas des écritures indiennes ou récemment inventées. C'est le cas, par exemple, des écritures Hmong (Shong Lue Yang's et Samuel Pollard's), de l'écriture Kommadam ainsi que des récentes écritures Chin, Kayah et Kayin. D'autres systèmes d'écriture peuvent aussi avoir été omis.
Sur ces systèmes d'écriture et sur d'autres, voir [Coyaud 1995] et [Coyaud 1997].

[2] Ce modèle pallava était en usage en Inde du sud entre le 3e et le 5e siècles [Ferlus 1988].

[3] Le tai khao est aussi appelé tai don. Le tai yo est aussi appelé quy chau. Le lai pao est aussi appelé tai muong. Le dehong dai est aussi appelé tai nüa. Le lanna est aussi appelé yuon.

[4] Le cham est parfois considéré comme une partie du monde indonésien. Les écritures cham sont les plus anciennes. Elles sont apparues au 3e siècle (en langue indienne, [Février 1959], p. 370). Le khmer est attesté en 611 (en khmer, [Huffman 1970], p. 4) et le môn au 6e ou 7e siècle (en môn, [Ferlus 1988], p. 8).

[5] Les écritures des Tai du nord-ouest sont très proches des écritures môn-birmanes.

[6] Les écritures tham lao et isan sont principalement utilisées dans les textes bouddhistes écrits en pali, langue traditionnelle du bouddhisme.

Les populations peuplant aujourd'hui l'Asie du Sud-Est continentale sont arrivées à diverses périodes. Chams, Môns et Khmers étaient là depuis longtemps lorsque les Birmans et les Thaïs arrivèrent entre les onzième et treizième siècles. Chaque langue a adapté les écritures indiennes présentes alors en fonction de ses besoins phonologiques : des phonèmes et des tons nouveaux devaient être transcrits. L'adaptation fut parfois incomplète, comme par exemple dans le cas de certaines écritures tai du nord-est qui ne permettent pas de rendre compte entièrement de la phonologie des langues qu'elles transcrivent.

Famille	Sous-Famille	Système d'écriture	Exemple
Khmer	Khmer	Khmer	មានរឿងមួយដំណាលថា មានភិក្ខុ ១ អង្គ
	Lao-thaï	Lao	ບ້ານຂ້ອຍມີໂຮງຮຽນຫຼັງນຶ່ງ
		Thaï	ควรจะกล่าวรวบรัดให้กระชับ ไม่ยืดยาดน่าเบื่อ
	Tai du nord-est	Tai Dam	[illegible]
	Cham	Cham	[illegible]
Môn	Môn	Môn	[illegible]
	Tham	Tham Lao	[illegible]
		Lü	[illegible]
		Lanna	[illegible]
	Tai du nord-ouest	Shan	[illegible]
	Birman	Birman	[illegible]

Figure 47 : Exemples d'écritures non segmentées d'Asie du sud-est continentale

Ces écritures ne forment pas un ensemble homogène, certaines étant presque éteintes (lai pao), écrites verticalement et en déclin (tai yo) ou variant d'un village à l'autre (tai khao et tai deng).

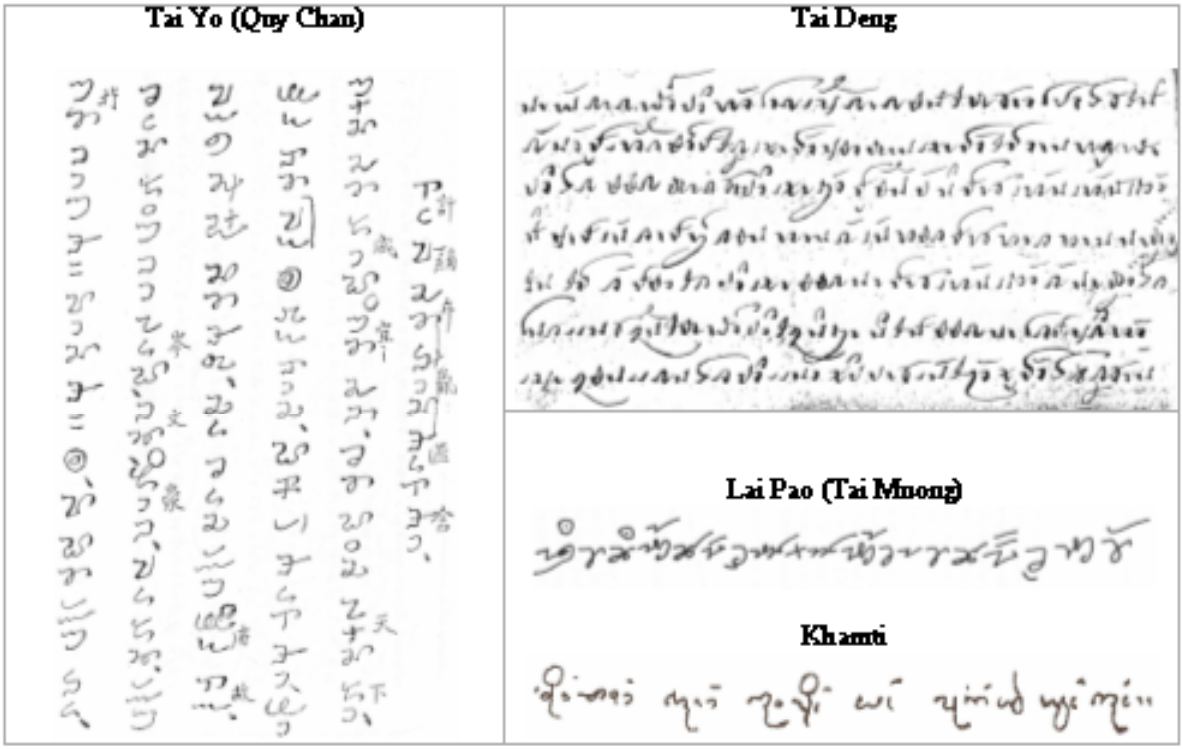


Figure 48 : Autres exemples d'écritures non segmentées (pas de police de caractères)

La plupart des langues associées à ces systèmes d'écriture sont des langues-π. Des polices de caractères sur huit bits ont cependant été développées pour plusieurs d'entre elles. Dans ce contexte, la langue thaïe est une exception brillante grâce aux recherches dynamiques qui l'ont dotée de riches ressources informatiques incluant la traduction automatique et une prise en compte par Microsoft Windows. Les travaux menés par les Thaïs sur la segmentation sont présentés au chapitre suivant.

Dans sa version 3.0, le standard Unicode prend en compte les écritures thaïe, laotienne, khmère et birmane. Grâce à Unicode, les systèmes d'écriture laotien, khmer et birman bénéficient donc d'un standard d'encodage (le thaï avait déjà sa norme TIS 620-2533). Plusieurs autres systèmes d'écriture (cham, tai nüa, lanna...) avaient été proposés pour les versions ultérieures d'Unicode. La nouvelle version 4.0 n'a pris en compte que le tai nüa (appelé *tai le* dans le standard : zone allant de 1950 à 197F).

III.2.2.2 TRAVAUX SUR LA SEGMENTATION

La segmentation des textes a été principalement étudiée, d'un point de vue général, aux deux niveaux du **mot** et de la **phrase** ([Palmer 2000]). La raison en est que ces deux niveaux doivent être correctement traités pour que des analyseurs de texte puissent fournir des résultats corrects.

À l'intérieur de cette problématique générale, les **systèmes d'écritures non segmentés**[1] ont une place particulière du fait que les frontières des mots ne sont pas marquées. La technique générale de segmentation des mots utilisée avec ces écritures emploie un algorithme qui recherche dans un dictionnaire les mots correspondant à ceux du texte et qui, en cas d'ambiguïté, sélectionne celui qui optimise un paramètre dépendant de la stratégie choisie. Dans les stratégies les plus courantes, l'optimisation consiste à :

- ⇨ maximiser la taille des mots, pris un par un de gauche à droite, avec retour arrière en cas d'échec (« plus longue chaîne d'abord » ou « *longest match* »),
- ⇨ minimiser le nombre de mots dans la phrase entière (« plus petit nombre de mots » ou « *maximal match* »).

Dans la zone géographique qui nous intéresse, la segmentation de texte a été principalement étudiée pour le thaï[2]. Les techniques de **segmentation de mots**, en particulier, ont été améliorées de manière continue depuis les premiers travaux. Les premières approches, qui datent des années 1980, ont obtenu des résultats corrects avec des algorithmes de « plus longue chaîne d'abord » ([Rarunrom 1991][3]) et de « plus petit nombre de mots » ([Sornlertlamvanich 1993]). Les résultats obtenus sont les suivants (nombre de mots correctement segmentés / nombre total de mots[4] ; [Theeramunkong et al. 2000]) :

- ⇨ *plus longue chaîne d'abord* : 97,03 % sans mot inconnu et 86,21 avec 20 % de mots inconnus,
- ⇨ *plus petit nombre de mots* : 97,24 % sans mot inconnu et 82,60 avec 20 % de mots inconnus,

Cependant, ces méthodes ont des performances médiocres en présence de cas d'ambiguïté et de mots inconnus. Le premier type de problèmes a été récemment réduit grâce à des algorithmes dans lesquels l'environnement des mots est analysé linguistiquement (parties du discours, mots contextuels, collocations) pour déterminer la segmentation la plus probable ([Meknavin et al. 1997])[5]. Cette méthode résout correctement les ambiguïtés dans au moins 86,6 % des cas lorsque la résolution nécessite de prendre en compte le contexte[6] et dans au moins 91,27 % dans le cas contraire.

Le problème des mots inconnus a été traité par des algorithmes ayant recours à un marquage sémantique et à des règles contextuelles pour extraire des informations syntaxiques et sémantiques des mots inconnus ([Kawtrakul et al. 1997])[7].

---

[1] Palmer appelle « écritures non segmentées » (*unsegmented writings*) les systèmes d'écriture dans lesquels les mots ne sont pas séparés par une espace, par opposition aux « écritures délimitées par des espaces » (*space-delimited writings*) que nous appelons systèmes d'écriture segmentés. Dans cette catégorie des systèmes d'écriture non segmentés, le chinois, le japonais, le coréen et le thaï ont été étudiés.

[2] Pour une liste étoffée de publications thaïes, voir http://crcl.th.net/crcl/bib/by-date.htm.

[3] Cette méthode de « plus longue chaîne d'abord » avec retour arrière avait déjà été appliquée sur les syllabes en 1986 par Poowarawan et Imarom ([Poowarawan et Imarom 1986]).

[4] Si on a une ambiguïté tous les dix mots, un résultat de 97 % correspondra à (1 – ((1 - 0,97) x 10) = 70 % d'ambiguïtés correctement résolues.

[5] Une technique dérivée du « modèle de trigramme » a aussi été proposée pour résoudre la question mais n'a obtenu que des résultats limités en présence d'ambiguïtés dépendant du contexte (voir par exemple [Kawtrakul et al. 1995a]). Un tableau de comparaison des performances est fourni dans [Sornlertlamvanich et al. 1997].

[6] Exemple de cas (en thaï) où le contexte est nécessaire : ที่รัก peut être un seul mot (« *chéri* ») ou deux mots (« *qui aime* ») : ที่ + รัก.

[7] Nous ne disposons pas de résultats sur les performances de cet algorithme.
Nota : Le problème des mots inconnus est pris en compte dans des algorithmes de segmentation plus généraux (voir par exemple [Kanlayanawat and Prasitjutrakul 1997]).

Toutes ces techniques recourent intensivement à des dictionnaires, qu'il faut donc créer[1]. Une approche sans dictionnaire a récemment été proposée pour la segmentation en mots. Cette technique s'appuie sur des arbres de décision intégrant des groupes de caractères appelés TCC (Thai Character Clusters) qui forment des unités inséparables. Cette méthode donne des résultats de l'ordre de 85 %, ce qui correspond aux résultats que les méthodes de *plus longue chaîne d'abord* et de *plus petit nombre de mots* donnent avec un taux de mots inconnus de 20 %. Elle peut être utilisée en complément d'autres méthodes pour améliorer les performances en présence de mots inconnus ([Theeramunkong et al. 2000]).

Les algorithmes de **segmentation en phrases thaïes** utilisent les espaces qui marquent traditionnellement les fins de phrases en thaï. Comme ces espaces peuvent avoir d'autres rôles que d'être des séparateurs de phrases, un marquage statistique en parties du discours est appliqué pour ne retenir que les séparateurs de phrase ([Mittrapiyanuruk et Sornlertlamvanich 2000]).

La **segmentation syllabique** a aussi été étudiée pour le thaï, en particulier à des fins de synthèse vocale. L'approche basée sur des règles utilisée dans ces études consiste à décrire l'ensemble des syllabes par une expression régulière[2] et à en dériver un automate ou un transducteur.

Au moins deux développements différents ont été décrits[3] :

- ⇨ *Thai Soundex* ([Karoonboonyanan et al. 1997]),
- ⇨ *Thai TTS* ([Mittrapiyanuruk et al. 2000]).

Thai Soundex diffère fondamentalement de Thai TTS en ce qu'il intègre des règles phonétiques dans le modèle de syllabe et dérive ainsi un transducteur produisant directement la transcription phonétique des syllabes passées en paramètre. Il s'appuie sur l'outil FIRE de Bruce Watson ([Watson 1994]) pour générer ce transducteur. Conséquence de la nature ambiguë de la lecture du thaï, le transducteur obtenu est non déterministe (dans le cas général, une séquence de caractères thaïs peut être lue de plusieurs façons).

Thai TTS utilise d'abord l'analyseur lexical LEX pour générer un automate déterministe qui reconnaît le langage des syllabes. À partir des syllabes reconnues, Thai TTS construit alors la transcription phonétique en utilisant une table de correspondance syllabe-prononciation.

Des études de transcription phonétique ont aussi été menées par les chercheurs thaïs. Elles se basent sur les mêmes solutions que *Thai Soundex* (transducteur non déterministe) et utilisent des techniques probabilistes pour résoudre les cas d'ambiguïté ([Charoenporn et al. 1999]).

[1] Bien que cela puisse être fait automatiquement par apprentissage de formes à partir d'un corpus, ces dictionnaires ont été créés manuellement ([Sornlertlamvanich et al. 2000]).

[2] Les premiers travaux ([Thairatananond 1981] et [Charnyapornpong 1983], cités dans [Aroonmanakun 2002]) utilisaient des modèles de syllabes. Ils ont permis d'obtenir des taux de bon découpage respectivement de 85 % et de 96 %. Puis, des méthodes à base de dictionnaire sont apparues. Par exemple, Poowarawan et Imarom ([Poowarawan et Imarom 1986]) ont utilisé une liste de 5400 syllabes (50 ko) issus d'un corpus de 140 000 mots, en association avec un algorithme de « plus longue chaîne d'abord » avec retour arrière. Le taux de bon découpage atteint était de l'ordre de 99 %. Des méthodes mixtes à base de règles et de dictionnaire ont été proposées plus récemment, ainsi que divers algorithmes (voir [Aroonmanakun 2002]).

[3] Nota : Un logiciel de synthèse vocale appelé CU-TTS a été développé à l'université de Chulalongkorn.

III.2.3 Modèle grammatical des syllabes

III.2.3.1 UN MODÈLE SYLLABIQUE NATUREL

Nous cherchons à réaliser un outil permettant à des linguistes de décrire un modèle de syllabes dans leur propre terminologie. Examinons quelques modèles de syllabes d'Asie du Sud-Est tels que décrits par des linguistes.

- ⇨ À la page 11 de son livre sur l'écriture **khmère** (*Cambodian system of writing and beginning reader* [Huffman 1970]), Franklin Huffman écrit « *The structure of monosyllables is shown by the formula $C_1(C_2)(C_3)V_1(V_2)(C_4)$, with the limitation that if V_2 doesn't occur, then C_4 must occur.* », les C_i et les V_i étant des consonnes et des voyelles décrites précédemment dans le livre.
- ⇨ Pour le **laotien**, Marc Reinhorn écrit à la page 11 de [Reinhorn 1975] que les syllabes peuvent être de la forme (en négligeant les tons) CV, CVC, CaCV ou CaCVC (CaC représentant un pseudo-groupe consonantique).
- ⇨ Pour le **thaï**, Gilles Delouche[1] liste dans [Delouche 1988], page LP.3, les phonèmes consonantiques pouvant exister à l'initiale et en finale, les combinaisons de ces phonèmes pouvant se trouver à l'initiale, les phonèmes vocaliques et les diphtongues puis fait le lien avec l'écriture thaïe dans la deuxième partie du manuel. Dans un autre livre sur l'écriture thaïe (*Lire et écrire le thaï*, [Brown 1991]), Marie-Hélène Brown écrit, page 54, que les syllabes sont de la forme CV ou CVC. Elle précise par ailleurs que certaines consonnes n'existent qu'à l'initiale et que le premier C de la syllabe pouvait être constitué de deux consonnes, en listant les couples de consonnes possibles.
- ⇨ Enfin, pour le **birman**, Denise Bernot, Marie-Hélène Cardinaud et Marie Yin Yin Myint indiquent, dans [Bernot et al. 1990] pages xxx et xxxi, que les syllabes peuvent être de la forme CV, CVN (N pour nasale) ou CVC, et détaillent les différents cas possibles. Haigh Roop écrit, quant-à lui, à la page xi de [Roop 1972] : « *A syllable in Burmese consists of an initial consonant or a cluster of two consonants, a vowel nucleus pronounced with one of four tones or atonically, and sometimes final /n/. That is C_1 (C_2) V/(T) (n).* » puis détaille les différents constituants.

De telles descriptions, à cheval entre l'écriture et la phonologie, se retrouvent dans de nombreux manuels d'apprentissage des systèmes d'écriture qui nous intéressent. À partir d'une telle approche, nous pouvons facilement écrire un ensemble de règles de la forme [2]:

Syllabe = Ci V *ou* Ci V Cf *ou* ...
Ci = c_{i1} *ou* c_{i2} *ou* ...
V = v_1 *ou* v_2 *ou* ...
Cf = c_{f1} *ou* c_{f2} *ou* ...

En optant pour une représentation de cette forme, on obtient immédiatement la grammaire de réécriture reconnaissant le langage des syllabes. Par exemple, une description simplifiée des syllabes khmères peut être donnée par :

Syllabe = Cons Voyelle + Cons Voyelle Cons ;

Cons = : ក + : ខ + : គ + : ឃ + : ង + : ច + : ឆ + : ជ + : ឈ + : ញ +
: ដ + : ឋ + : ឌ + : ឍ + : ណ + : ត + : ថ + : ទ + : ធ + : ន +
: ប + : ផ + : ព + : ភ + : ម + : យ + : រ + : ល + : វ + : ស + : ហ + : ឡ + : អ ;

Voyelle = : ា + : ិ + : ី + : ឹ + : ឺ + : ុ + : ូ + : ួ ;

[1] Gilles Delouche est professeur de siamois (thaï) à l'INALCO, institut dont il est aussi le président.
[2] Ci, Cf et V représentent respectivement la consonne initiale, la consonne finale et la voyelle.

Ce type de représentation des syllabes par une grammaire, à la fois naturel et formel, permet de tendre un pont entre la terminologie des linguistes et les langages formels que « comprennent » les machines. Nous l'avons donc choisi pour décrire les syllabes des systèmes d'écriture étudiés.

III.2.3.2 GRAMMAIRE DES SYLLABES LAOTIENNES PAS À PAS

Dans ce chapitre, nous détaillons le processus de réalisation d'une grammaire de syllabes, dans le cas du laotien. Le système d'écriture du laotien inclut les éléments suivants[1,2].

- 27 consonnes simples : ກ ຂ ຄ ງ ຈ ສ ຊ ຍ ດ ຕ ຖ ທ ນ ບ ປ ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ວ ຫ ອ ຮ.

- 6 groupes consonantiques associatifs[3]: ຫງ ຫຍ ຫນ/ໜ ຫມ/ໝ ຫລ/ໜ ຫວ.

En plus de ces « groupes associatifs » traditionnellement considérés comme des consonnes, il existe une autre catégorie de groupes consonantiques, celle des « groupes consonantiques englobants ». Ces derniers sont constitués d'une consonne simple suivie de la semi-voyelle ວ.

- 38 voyelles[4], faits de l'arrangement de 21 graphèmes de base: xະ ັx xา[x] ິx[x] ີx[x] ຶx[x] ືx[x] ຸx[x] ູx[x] ເxะ ເັx ເx[x] ແxะ ແັx ແx[x] ໂxะ ົx ໂx[x] ເxาะ xັx ໍx xx ເິx[x] ເີx[x] ເັຍ xັ້x ເxຍ xົ້x ເຶx[x] ເືx[x] ົxะ xັx ົx xx ໄx ໃx ເິา xำ.

- 4 accents[5,6] : x ◌่ x ◌้ x ◌๊ x ◌๋.

La forme générale des syllabes laotiennes est : **C_i [A] V [C_f]** [7] où :

- ⇨ C_i peut être, soit une consonne simple quelconque, soit un groupe consonantique,
- ⇨ C_f peut être l'une des consonnes ກ, ດ, ບ, ງ, ນ, ມ, ວ et ຍ.

La difficulté vient avec les voyelles. Comme nous l'avons fait remarquer au chapitre II.1.1.2, les voyelles laotiennes peuvent être situées avant, après, sur ou sous la consonne initiale. Par exemple, la voyelle ເ◌ິ est constituée de deux graphèmes : ເ situé avant la consonne et ◌ິ situé après (au dessus, visuellement). La partie droite de la règle *Syllabe = C_i [A] V [C_f]* devient ainsi plus complexe car on ne peut pas se contenter d'une règle *V = liste* ; elle doit être réécrite comme une disjonction de termes tels que : **ເ C_i [A] ◌ິ [C_f]**.

---

[1] Voir par exemple [Reinhorn 1970], [Reinhorn 1975] ou [Inthamone 1987]. La typologie des groupes consonantiques est détaillée dans [Inthamone 1987]. Nous résumons ici les éléments présentés au chapitre II.1.1.2 qui sont utiles pour décrire les syllabes.

[2] Le système d'écriture laotien contient aussi dix chiffres (໐ ໑ ໒ ໓ ໔ ໕ ໖ ໗ ໘ ໙) ainsi que quelques autres signes, en particulier de ponctuation (voir le chapitre II.1.1.2).

[3] Trois d'entre eux présentent deux formes, comme cela est indiqué. Le groupe théorique ຫຣ n'est pas pris en compte ici.

[4] Nouvelle écriture.

Nota : Le signe X indique la position de la consonne prononcée juste avant la voyelle. Lorsque deux X sont présents, la syllabe accepte une consonne finale. La forme [X] indique que X est facultatif.

[5] Le signe ◌ົ est parfois inclus dans la liste des accents. Il n'est pas pris en compte dans notre modèle.

[6] Nous adoptons, dans la grammaire qui suit, un modèle simplifié pour les accents consistant à les accepter lorsque la syllabe inclut une consonne finale. Un modèle plus précis pourrait être développé, contenant des règles plus restrictives, en particulier n'acceptant la présence d'accents que dans les syllabes ouvertes. Lorsqu'ils sont présents, les accents sont supposés suivre immédiatement la consonne initiale ou le groupe consonantique (forme canonique).

[7] C_i=Consonne initiale (simple ou groupe consonantique), A=Accent, V=Voyelle, C_f=Consonne finale. L'expression [X] signifie que X est facultatif.

À partir de la description ci-dessus, nous pouvons dériver l'ensemble de règles suivant (grammaire des syllabes laotiennes, notations basées sur l'anglais) en transformant C_i en CC, C_iA en CCA, A en Acc, C_f en FC et [C_f] en OFC [1,2,3]. La syntaxe de cette grammaire est définie par une grammaire — la grammaire des grammaires — donnée en annexe A.11.

Syllabes = CC :ະ + CCA :ັ FC + CCA :ໍ OFC +

CCA (:ິ + :ີ + :ຶ + :ື + :ຸ + :ູ) OFC +

:ເ CC :ະ + :ເ CCA :ັ FC + :ເ CCA OFC +

:ແ CC :ະ + :ແ CCA :ັ FC + :ແ CCA OFC +

:ໂ CC :ະ + CCA :ົ FC + :ໂ CCA OFC +

:ເ CC :ໍາ + CCA :ັ FC + CCA :ໍ + CCA :ອ FC +

:ເ CCA (:ິ + :ີ) OFC +

:ເ CC :ັຍ + CCA :ັ FC + :ເ CCA :ຍ + CCA :ງ FC +

:ເ CCA (:ິອ + :ີອ) OFC +

CC :ົວະ + CCA :ັ FC + CCA :ົວ + CCA :ອ FC +

(:ໄ + :ໃ) CCA + :ເ CCA :ົ + CCA :ໍ ;

CCA = CC + CC Acc ;

CC = Cluster + IC ;

Cluster = :ຫ (:ງ + :ຍ + :ນ + :ມ + :ລ + :ົ + :ວ) + :ຫ + :ໝ + IC :ອ ;

IC = :ກ + :ຂ + :ຄ + :ງ + :ຈ + :ສ + :ຊ + :ຍ + :ດ + :ຕ + :ຖ + :ທ + :ນ + :ບ + :ປ +

:ຜ + :ຝ + :ພ + :ຟ + :ມ + :ຢ + :ຣ + :ລ + :ວ + :ຫ + :ອ + :ຮ ;

OFC = FC + {} ;

FC = :ກ + :ດ + :ບ + :ງ + :ນ + :ມ + :ຍ + :ວ ;

Acc = :ˈ + :ᴮ + :ᵐ + :ˣ ;

[1] **Notations** (en anglais) : CC=*Consonant* ou *Cluster* ; CCA=*Consonant* ou *Cluster* suivi d'un accent optionnel ; IC=*Initial Consonant* ; FC=*Final Consonant* ; OFC=*Optional Final Consonant* ; Acc=*Accent*.

[2] L'existence d'homographes entraîne que des dérivations peuvent être redondantes, comme par exemple les syllabes « :ເ CCA :ຍ » qui sont aussi produites par « :ເ CCA OFC » du fait que ຍ est une consonne finale.

[3] Sous cette forme, la grammaire des syllabes laotiennes produit des syllabes qui ne sont que théoriques. Par exemple, la forme ນວວ sera acceptée comme syllabe valide bien qu'elle n'existe pas. Les règles doivent être davantage détaillées si l'on veut éliminer de telles syllabes théoriques. La création de la liste des syllabes considérées comme valides (productible automatiquement à partir de la grammaire) peut aider à détecter les syllabes indésirables en les visualisant.

III.2.3.3 NOMBRE DES SYLLABES GÉNÉRÉES PAR LES MODÈLES

Il est possible d'obtenir un majorant du nombre de syllabes générées par les modèles grammaticaux utilisés en observant la forme générale des syllabes. Par exemple, en prenant pour forme générale des syllabes laotiennes « **C [C] [A] V [C]** », le nombre de possibilités pour C, A et V étant 27, 4 et 38, on obtient le majorant 27x28x5x38x28, soit 4 021 920 syllabes.

En prenant la forme générale plus fine « **C [A] V [CF]** ou **GC [A] V [CF]** » (GC représentant les groupes consonantiques), on obtient un majorant plus proche de la réalité. Avec |C| = 27, |A| = 4, |GC| = 36, |V| = 38 et |CF| = 8, cela donne 95 760 syllabes, la valeur réelle étant de 56 670.

Pour les grammaires du birman, du khmer et du thaï présentées ci-dessous, nous pouvons donner les majorants grossiers suivants.

- ⇨ **birman** :
  - Forme générale : (C ou CS) [L] V [C ou ◌် ou ◌ဳ [◌ ်] [◌့ ou ◌း]]
    (CS = consonnes souscrites et L = ligatures)
  - Cardinaux des constituants : |C| = 33, |CS| = 20, |L| = 15, |V| = 35
  - Majorant : (33+20)x16x35x(34+2)x2x3 = 6 410 880 syllabes

- ⇨ **khmer** :
  - Forme générale[1] : (C [CS [CS]] [D1] ou CS [CS]) V [C [CS] [D2]] ou VI ou L
    (D1 = diacritiques ◌ៈ, ◌ៗ et ◌ៗ, D2 = diacritiques ◌ៗ et ◌ៗ, CS = consonnes souscrites,
    VI = voyelles indépendantes et L = ligatures)
  - Cardinaux des constituants : |C| = 33, |CS| = 32, |V| = 33, |VI| = 14, |L| = 10
  - Majorant : (33x33x33x4+32x33)x33x(34x33x3)+14+10 = 16 084 538 736 syllabes

- ⇨ **thaï** :
  - Forme générale : (C ou GC) [A] V [CF]
  - Cardinaux des constituants : |C| = 44, |GC| = 140, |A| = 4, |V| = 41, |CF| = 38
  - Majorant : (44+140)x5x41x39 = 1 471 080 syllabes

Il est remarquable que le nombre de syllabes puisse être considérablement réduit en regroupant les symboles jouant le même rôle. Par exemple, dans la forme générale des syllabes thaïes ci-dessus — (C ou GC) [A] V [CF] — les symboles C et GC jouent le même rôle. De plus, les consonnes et groupes consonantiques qu'ils représentent forment une chaîne connexe de caractères, comme en laotien. Cela est vrai aussi pour les consonnes finales (les CF de la formule).

Les grammaires produisent de nombreuses syllabes. Parmi elles, une grande partie est théorique dans le sens où elles ne représentent pas des sons de la langue. Cependant, la transcription de mots étrangers peut faire apparaître certaines de ces formes, comme par exemple le son *fr* en khmer qui est rendu par la forme ផ្រ, comme dans ផ្រាំ (franc).

[1] Forme rigoureuse (voir III.2.3.5.1). Un début en « C [CS [CS]] » est cependant rare et le majorant qui en découle est une mesure très surévaluée de la réalité. Ne pas en tenir compte donnerait (33x33x4+32x33)x33x (34x33x3)+13+10 = 601 154 159, soit environ 27 fois moins.

La grammaire des syllabes thaïes (voir le chapitre III.2.3.6) produit plus d'un million de syllabes. Une version simplifiée ne gardant comme consonne initiale ou finale que ก, et ne tenant pas compte des accents, produit la liste ci-dessous qui ne contient que 58 syllabes. Les autres syllabes s'en déduisent en remplaçant ก par les différentes valeurs possibles des formes consonantiques initiales et finales et en ajoutant un accent, lorsque la syllabe le permet.

กะ	กี	กู	กรร	เกอ	เกือก	ไก	ฤ
กัก	กีก	กูก	กรรก	เกอะ	แก	ใก	ฤๅ
กัว	กึ	กก	เก	เกิก	แกะ		ฦ
กัวะ	กึก	ก็อก	เกะ	เกีย	แก็ก		ฦๅ
กา	กือ	กอ	เก็ก	เกียะ	แกก		อยู่
กาก	กุ	กอก	เกก	เกียก	ก		อย่า
กิ	กือก	กวก	เกา	เกือ	โกะ		อย่าง
กิก	กุก	กำ	เกาะ	เกือะ	โกก		อย่าก

III.2.3.4 SYLLABES BIRMANES

III.2.3.4.1 Alphabet birman

L’alphabet birman contient :

- **33 consonnes de base** :

က, ခ(*), ဂ(*), ဃ, င(*), စ, ဆ, ဇ, ဈ, ည, ဋ, ဌ, ဍ, ဎ, ဏ, တ, ထ, ဒ(*), ဓ, န / ဿ, ပ(*), ဖ, ဗ, ဘ, မ, ယ, ရ, လ, ဝ(*), သ, ဟ, ဠ, အ.

(*) : Ces six consonnes s’associent avec les voyelles ါ, ါး, ေါ့, ေါ် et ော au lieu de ာ, ားး, ော့, ော် et ော.

- **20 consonnes souscrites** :

က်, ခ်, ဂ်, ဃ်, စ်, ဆ်, ဇ်, ဟ်, ဿ်, ဏ်, တ်, ထ်, ဒ်, ဓ်, န်, ပ်, ဗ်, ဘ်, မ်, လ်.

- **21 voyelles dont 14 revêtent 2 formes** :

ံ, ာ, ါ, ားး, ါး, ိ, ည့်, ီ, ည်, ီး, ည်း, ု, ူ, ူး, ေ့, ေ, ေး, ဲ့, ယ်, ဲ, ော့, ေါ့, ော်, ော်, ော, ော်, ို့, ို့, ို, ီ့, ိုး, ီး.

Nota : La série ည့်, ည်, ည်း est une variante de la série ိ, ီ, ီး, mais aussi de la série ဲ့, ယ်, ဲ.

- **4 ligatures simples, dont une peut avoir deux largeurs** (*) : ျ, ◌ြ , ◌ြ , ◌ှ, ◌ွ.

- **5 ligatures doubles, dont deux peuvent avoir deux largeurs** (*) : ◌ျ, ◌ျ , ◌ြ, ◌ြ, ◌ြ, ◌ွ, ◌ွ.

- **2 ligatures triples, dont une peut avoir deux largeurs** (*) : ◌ျ, ◌ျ , ◌ြ.

(*) : ◌ြ et ◌ြ s’utilisent respectivement avec les consonnes constitués d’un rond (par exemple ○) et deux ronds (par exemple ∞).

- **8 formes rares (initiales de mots pali et de quelques mots birmans)** :

ၐ, ၑ, ၒ, ၓ, ၔ, ၕ , ၖ , ၗ .

- **5 abréviations représentant des syllabes** :

၍, ၎, ၏, ၌, ၍.

- **2 signes supplémentaires représentant des consonnes finales** :

ံ, ဳ.

III.2.3.4.2 Grammaire élémentaire des syllabes birmanes

syllabes = :၎ (CC1 + CS) (:ျ + :ြ + :ာ) OFC +
:၎ (CC2 + CS) (:ိ + :ု + :ါ) OFC +
:၎ (CC + CS) (:္ + {} + :း) OFC +
(CC1 + CS) (:ာ + :ား) OFC +
(CC2 + CS) (:ါ + :ါး) OFC +
(CC + CS) ({} + V1) OFC +
FR +
AB ;

OFC = {} + (C + CS + :ံ + :်) ({} + :ှ) (:့ + {} + :း) ;

C = C1 + C2 ;

CC = CC1 + CC2 ;

CC1 = C1 YARAWAHA + C1 ;

CC2 = C2 YARAWAHA + C2 ;

C1 = :က + :ဃ + :စ + :ဆ + :ဇ + :ဈ + :ည +
:ဋ + :ဌ + :ဍ + :ဎ + :ဏ + :တ + :ထ + :ဓ + :န + :ဒ +
:ဖ + :ဗ + :ဘ + :မ + :ယ + :ရ + :လ + :သ + :ဟ + :ဠ + :အ ;

C2 = :ခ + :ဂ + :ငၚ + :ဒ + :ပ + :ဝ ;

CS = :က + :ခ + :ဂ + :ဃ + :စ + :ဆ + :ဇ + :ဟ + :ၛ + :ဏ + :တ +
:ထ + :ဒ + :ဓ + :န + :ပ + :ဗ + :ဘ + :မ + :လ ;

V1 = :ဝ + :ည် + :ီ + :ည် + :ီး + :ည်း + :ု + :ူ + :ၥ + :ဦ +
:ၥး + :ဦး + :ွ + :ယ် + :ှ + :ဲ + :ွ + :ူ + :ု + :ုး + :ူး ;

YARAWAHA = :ျ + :ြ + :ွ + :ၟ + :ၠ +
:ြွ + :ြူ + :ွၠ + :ွၡ + :ြဲ + :ြွ +
:ြွှ + :ြှ + :ှၡ ;

FR = :ဣ + :ဤ + :ဥ + :ဦ + :ဧ + :ဩ + :ဪ + :ဪာ ;

AB = :၏ + :၍ + :၌ + :၎င်း + :၎ ;

III.2.3.5 SYLLABES KHMÈRES

III.2.3.5.1 Alphabet khmer

- 33 consonnes simples réparties en deux séries :

	k	kh	ng	c	ch	ñ	d	t	th		n	b	p	ph	m
1	ក	ខ	ង	ច	ឆ	ញ	ដ	ត	ថ	ឋ	ណ	ប	ប៉	ផ	ម៉
2	គ	ឃ	ង	ជ	ឈ	ញ	ឌ	ទ	ធ	ឍ	ន	ប៊	ព	ភ	ម

	y	r	l	w	s	h	q
1	យ៉	រ៉	ឡ	វ៉	ស	ហ	អ
2	យ	រ	ល	វ	ស៊	ហ៊	អ៊

Les consonnes de la première série ង, ញ, ម, យ, រ, វ et la consonne ប de la deuxième peuvent être surmontées du caractère ◌៉ [1] (dent de souris, pouvant être remplacé par ◌ុ lorsqu'une suscrite fait que la place est insuffisante au dessus de la consonne) qui en fait des consonnes de la deuxième série (ou l'assourdit, dans le cas du ប).

Les consonnes de la deuxième série ប, ស, ហ, អ peuvent être surmontées du caractère ◌៊ (cheveu, pouvant être remplacé par ◌ុ lorsqu'une suscrite fait que la place est insuffisante au dessus de la consonne), qui en fait des consonnes de la première.

- 32 consonnes souscrites dont 2 ont la même forme et 1 a deux formes :

◌្ក, ◌្ខ, ◌្គ, ◌្ឃ, ◌្ង, ◌្ច, ◌្ឆ, ◌្ជ, ◌្ឈ, ◌្ញ, ◌្ឋ, ◌្ឌ, ◌្ឍ, ◌្ណ, ◌្ត, ◌្ថ, ◌្ទ, ◌្ធ, ◌្ន, ◌្ប, ◌្ផ, ◌្ព, ◌្ភ, ◌្ម, ◌្យ, ◌្រ, ◌្ល, ◌្វ, ◌្ស, ◌្ហ, ◌្អ.

- 28 voyelles simples dont 5 ont deux formes [2] :

X[X], X◌ា[X], X◌ិ[X], X◌ី[X], X◌ឹ[X], X◌ឺ[X], X◌ុ[X], X◌ូ[X] , X◌ួ[X], X◌ើ[X], X◌ឿ[X], X◌ៀ[X], X◌េ[X], X◌ែ[X], X◌ៃ[X], X◌ោ[X], X◌ៅ[X], X◌ំ, X◌ះ[X], X◌ៈ, X◌ា, X◌ៅ, X◌ុ, X◌ូ, X◌ាំ, X◌ះ, X◌ិះ, X◌ីះ, X◌ុះ, X◌ះ, X◌ាះ, X◌ីះ.

- 14 voyelles indépendantes : ឥ, ឦ, ឧ, ឩ, ឪ, ឫ, ឬ, ឭ, ឮ, ឯ, ឰ, ឱ, ឳ, ឵.

[1] Pour des raisons esthétiques, des ligatures ont été créées pour les consonnes រ and វ surmontées des dents de souris. Ainsi ร៉ et ว៉ remplaceront ร៉ et ว៉ dans la grammaire proposée (dents de souris décalées à gauche).

[2] Le signe X indique l'emplacement des consonnes. Quand deux X sont présents, la syllabe accepte une forme consonantique finale. La forme [X] indique que X est optionnel.

- 10 ligatures : ឈា, ឱ, ខ្ញុំ, ក្ដុ, ប៉ៅ, ប៉ៅ, ឈោ, ក្ដឹ, រ៉, រ៉ៃ.

Nous prendrons C[C]V[C[C]] comme forme générale des syllabes khmères[2]. Les consonnes finales peuvent être surmontées du caractère ៉ ou du caractère ៍. Dans les cas où la syllabe considérée est la deuxième d'une dissyllabe, il se peut que toutes les consonnes initiales soient des souscrites, ce qui peut entraîner la superposition de trois niveaux de consonnes, deux niveaux de souscrites s'ajoutant à un troisième pour la consonne finale[3]. Ces cas à trois niveaux sont rares, par exemple : ស្ត្រី[4](sat-trey) qui signifie *femme*, ou កន្ត្រៃ (kân-tray) qui signifie *ciseaux*. Les polices ne possédant généralement qu'un niveau de souscrites, le troisième niveau est obtenu en jouant sur la hauteur du caractère, par exemple : អង់គ្លេស (ang-klais) ou លក្ខណ៍ (Lak[-shma-na], les deux dernières syllabes ne sont pas prononcées). Nous voyons dans les deux premières dissyllabes qu'un prétraitement doit être réalisé pour reformer leurs deuxièmes syllabes. En effet, dans ស្ត្រី, ស្ត représente *sat* et ្ត្រី représente *trey*. De même, dans កន្ត្រៃ, កន représente *kân* et ្ត្រៃ représente *tray* et dans អង់គ្លេស, អង់ représente *ang* et ្គ្លេស représente *klais*. La grammaire proposée ci-dessous suppose que ce prétraitement est réalisé.

III.2.3.5.2 Grammaire élémentaire des syllabes khmères

syllabes = CC ({} + :ា + :ិ + :ី + :ឹ + :ឺ + :ុ + :ូ + :ួ + :ុ + :ូ + :ួ) OFC +
:េ CC (:ី + :ឿ + :ឿ + :ៀ + :ៀ + {}) OFC +
:ែ CC (:ះ + OFC) + :ែ CC +

---

[1] La lettre ឲ្យ n'est pas une voyelle indépendante, mais est considérée comme une variante graphique de ces autres possibilités : អោយ, ឱយ et ឱ្យ.

[2] En toute rigueur, le modèle général des syllabes est C[C[C]]V[C[C]]. Par exemple, ហ្វ្រង់ (*frang*) qui signifie franc (la monnaie) ou ហ្វ្លរីដ qui signifie Floride, commencent par trois consonnes. Cependant, ces cas sont rares et ne sont pas pris en compte dans la grammaire proposée.

[3] Les cas de doubles consonnes en fin de syllabes n'existant qu'en fins de mots (par exemple សាស្ត្រ, *sastr* qui signifie sciences), il n'existe pas de cas où une fin de syllabe à deux consonnes est suivie par deux souscrites.

[4] ស្ត្រី (sat-trey) est parfois considéré comme une seule syllabe (strey).

: េ CC : ា OFC + : េ CC : ៅ +

CC (: ុំ + : ំ + : ាំ + : ះ + : ិះ + : ឹះ + : ុះ) +

: េ CC (: ះ + : ាះ + : ើះ) + IndVowels + Ligatures ;

IndVowels = : ឥ + : ឦ + : ឧ + : ឩ + : ឪ + : ឫ + : ឬ + : ឭ + : ឮ +

: ឯ + : ឰ + : ឱ + : ឳ + : ឲ្យ;

Ligatures = : បា + : ឱ + : ខ្ញុំ + : ក្នុ + : ប៉ា + : ប៉ោ + : បា + : រ្ដ៍ + : រ៉ី + : រ៉ ;

CC = IC + : ្រ IC + IC BC + : ្រ IC BC + BC + BC BC + : ្រ BC ;

OFC = (IC + IC BC + {}) ({} + : ់ + : ៍) + IC IC : ៍ ;

IC = : ក + : ខ + : គ + : ឃ + : ង ({} + : ៉ + : ុ) +

: ច + : ឆ + : ជ + : ឈ + : ញ ({} + : ៉ + : ុ) +

: ដ + : ឋ + : ឌ + : ឍ + : ណ +

: ត + : ថ + : ទ + : ធ + : ន +

: ប ({} + : ៉ + : ុ) + : ផ + : ព + : ភ + : ម ({} + : ៉ + : ុ) +

: យ ({} + : ៊ + : ៉ + : ុ) + : រ ({} + : ុ) + : រ៉ + : ល + : វ ({} + : ុ) + : វ៉ +

: ស ({} + : ៊ + : ុ) + : ហ ({} + : ៊ + : ុ) + : ឡ + : អ ({} + : ៊ + : ុ) ;

BC = : ្ក + : ្ខ + : ្គ + : ្ឃ + : ្ង +

: ្ច + : ្ឆ + : ្ជ + : ្ឈ + : ្ញ +

: ្ដ + : ្ឋ + : ្ឌ + : ្ឍ + : ្ណ +

: ្ត + : ្ថ + : ្ទ + : ្ធ + : ្ន +

: ្ប + : ្ផ + : ្ព + : ្ភ + : ្ម +

: ្យ + : ្ល + : ្វ + : ្ស + : ្ហ + : ្អ ;

III.2.3.6 SYLLABES THAÏES

III.2.3.6.1 Alphabet thaï

- 44 consonnes simples :

ก, ข, ฃ, ค, ฅ, ฆ, ง, จ, ฉ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ผ, ฝ, พ, ฟ, ภ, ม, ย, ร, ล, ว, ศ, ษ, ส, ห, ฬ, อ, ฮ.

- 8 consonnes composées :

หง, หญ, หน, หม, หย, หร, หล, หว auxquelles il faut ajouter les groupes de consonnes constituées d'une consonne simple suivie de ร, ล ou ว[1].

- 37 consonnes finales :

ก, ข, ฃ, ค, ฅ, ฆ, ง, จ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, พ, ภ, ม, ย, ร, ล, ว, ศ, ษ, ส, ฬ auxquelles il faut ajouter le groupe de consonnes ตร.

- 25 voyelles simples [2] :

Xะ, Xั X, Xา[X], Xิ[X], Xี[X], Xึ[X], Xื อ, Xื X, Xุ[X], Xู[X], เXะ, เX็ X, เX[X], แXะ, แX็ X, แX[X], โXะ, X[X], โX[X], เXาะ, X็ อX, Xอ[X], เXอะ, เXอ, เXิ X.

- 16 diphtongues et autres :

เXียะ, เXีย[X], เXือะ, เXือ[X], Xัวะ, Xัว, XวX, ไX, ใX, เXา, Xำ, ฤ, ฤๅ, ฦ, ฦๅ, Xรร[X].

- 4 formes en อย :

อย่า, อยาก, อย่าง, อยู่.

- 4 accents [3] :

่ ้ ๊ ๋

X, X, X, X .

III.2.3.6.2 Grammaire élémentaire des syllabes thaïes

syllabes = CC :ะ + CCA : ั FC + CCA :า OFC +

CCA (: ิ + : ี) OFC +

CCA : ึ OFC + CCA : ื อ + CCA : ื FC +

[1] Une optimisation possible est de limiter aux consonnes constatées en thaï (voir [Delouche 1988], page LP.4) .

[2] Le signe X indique l'emplacement des consonnes. Quand deux X sont présents, la syllabe accepte une consonne finale. La forme [X] indique que X est optionnel.

[3] Le signe X ์ n'est pas pris en compte dans cette version du modèle.

CCA (:ุ + :ู) OFC +

:เ CC :ะ + :เ CCA :็ FC + :เ CCA OFC +

:แ CC :ะ + :แ CCA :็ FC + :แ CCA OFC +

:โ CC :ะ + CCA FC + :โ CCA OFC +

:เ CC :าะ + CCA :็อ FC + CCA :อ OFC +

:เ CC :อะ + :เ CCA :อ + :เ CCA :ิ FC +

:เ CC :ีียะ + :เ CCA :ีีย OFC +

:เ CCA :ืือะ + :เ CCA :ืือ OFC +

CC :ัวะ + CCA :ัว + CCA :ว FC +

(:ไ + :ใ) CCA + :เ CCA :า + CCA :ำ +

:ฤ + :ฤๅ + :ฦ + :ฦๅ +

CCA :รร OFC +

+ :อย่า + :อยาก + :อย่าง + :อยู่ ;

CCA = CC + CC Acc ;
CC = Cluster + IC ;

Cluster = :ห (:ง + :ญ + :น + :ม + :ย + :ร + :ล + :ว) + IC (:ร + :ล + :ว) ;

IC = :ก + :ข + :ฃ + :ค + :ฅ + :ฆ + :ง + :จ + :ฉ + :ช + :ซ + :ฌ + :ญ + :ฎ + :ฏ + :ฐ + :ฑ + :ฒ + :ณ + :ด + :ต + :ถ + :ท + :ธ + :น + :บ + :ป + :ผ + :ฝ + :พ + :ฟ + :ภ + :ม + :ย + :ร + :ล + :ว + :ศ + :ษ + :ส + :ห + :ฬ + :อ + :ฮ ;

OFC = FC + {} ;

FC = :ก + :ข + :ฃ + :ค + :ฅ + :ฆ + :ง + :จ + :ช + :ซ + :ฌ + :ญ + :ฎ + :ฏ + :ฐ + :ฑ + :ฒ + :ณ + :ด + :ต + :ถ + :ท + :ธ + :น + :บ + :ป + :พ + :ภ + :ม + :ย + :ร + :ล + :ว + :ศ + :ษ + :ส + :ฬ + :ตร ;

Acc = : + : + : + : ;

III.2.4 Compilation des grammaires

Dans nos logiciels pour la langue laotienne[1], la reconnaissance syllabique est réalisée par du code C++ obtenu à partir de la grammaire des syllabes grâce au compilateur de grammaires hors contexte *saint-jean*[2] (Del Vigna, 2000). La description des syllabes sous forme de grammaire hors-contexte permet d'utiliser directement ce compilateur qui génère un analyseur syntaxique reconnaissant les syllabes.

Le problème de la reconnaissance des syllabes n'est cependant pas de nature hors contexte[3] et un automate d'états finis paraît mieux adapté qu'un analyseur syntaxique pour déterminer si une syllabe fait partie du langage décrit par la grammaire. Les études que nous avons menées sur les ambiguïtés[4] nous ont conduit à rechercher un calcul plus efficace pour déterminer si une chaîne de caractères appartient à un langage fini (celui des syllabes) décrit par une grammaire donnée.

Dans un premier temps, les syllabes ont été générées une à une à partir de la grammaire puis rangées dans un arbre lexicographique (un cas particulier d'automate d'états finis). En plus de l'avantage de sa complexité linéaire, l'arbre lexicographique, du fait de sa structure, permet de supprimer les doublons apparaissant lors de la génération (voir la note 2 de la page 133) et ainsi d'éviter leur stockage.

La taille très importante des dictionnaires de syllabes pour certaines langues comme le khmer nous a cependant amené à remettre en question la technique d'arbre lexicographique. En effet, si le dictionnaire des syllabes du laotien compte 56 670 éléments, celui du thaï en compte plus d'un million et celui du khmer plus d'un milliard[5]. Ces volumes rendent impossible, en pratique, le stockage des arbres lexicographiques en mémoire centrale, voire sur disque dur pour le khmer. Nous pouvions d'ailleurs craindre que même un automate d'états finis minimal prenne trop de place et que la solution « analyseur syntaxique » soit la seule solution générique possible.

L'expérimentation réalisée par Yeu [Yeu 2003] a cependant montré à la fois que la construction de l'automate d'états finis minimal qui reconnaît le langage des syllabes pour ces différentes langues prend un temps raisonnable — moins d'une seconde pour l'automate des syllabes laotiennes et moins de quatre pour celles du khmer — et qu'il prend une place faible en mémoire. Les résultats sont les suivants (réalisé avec une machine Intel sous Windows 2000 cadencée à 2 GHz).

	Laotien	Khmer
Nombre de syllabes du langage	56 670	$> 10^9$
Nombre d'états de l'automate minimal	33	58
Nombre de transitions de l'automate minimal	480	2 726
Temps de calcul de l'automate minimal	< 1 s	< 4 s

Figure 49 : Caractéristiques d'automates reconnaissant les syllabes laotiennes et khmères

[1] Il s'agit des logiciels Tallao 3.2, LaoPad et LaoWord. La version 2 de Tallao reposait sur un interpréteur développé spécifiquement et trop lent, ce qui nous avait conduit à générer la liste triée des syllabes et à l'utiliser pour accélérer le traitement, en particulier lors des mouvements souris.

[2] S'écrit sans majuscule, voir http://cams-atid.ivry.cnrs.fr/saint-jean/grammaires.html.

[3] Les grammaires du chapitre III.2.3 ne contiennent pas de symbole non terminal récursif. Cela traduit le fait que le nombre des syllabes est fini et donc que le problème est de nature régulière.

[4] Voir le chapitre III.2.6 et l'article [Del Vigna et Berment 2004], à paraître, donné en annexe A.11.

[5] Le nombre important des syllabes vient en partie du fait que la grammaire recense non seulement les syllabes attestées mais aussi les syllabes « théoriques », c'est à dire celles qui, bien qu'en accord avec la langue, sont peu ou pas attestées mais dont l'occurrence dans un texte ne peut être exclue. Exemple de syllabe théorique en français : *xieng*. Bien que n'existant dans aucun mot français, elle est utilisée, par exemple, pour transcrire le nom de la ville laotienne de *Xieng-Khouang*.

La construction de cet automate minimal a suivi les étapes suivantes, proches de la démarche des travaux sur le thaï[1] (voir III.2.2.2) :

- ⇨ calcul d'une expression régulière à partir de la grammaire,
- ⇨ calcul d'un automate non déterministe à partir de l'expression régulière,
- ⇨ calcul de l'automate minimal à partir de l'automate non déterministe.

Les minimisations d'automates ont été réalisées grâce à l'algorithme de Brzozowski [Watson 1993] et les déterminisations grâce à l'algorithme décrit dans [Leslie 1995].

Le programme calculant les automates minimaux à partir des grammaires sera prochainement adapté pour produire ces automates sous forme de code C++. Le code généré sera intégré dans nos logiciels « laotiens » ainsi que dans Sylla (voir le chapitre III.2.6).

[1] Nota : Cette démarche n'est pas la plus directe. On peut, en effet :

- ⇨ fabriquer directement un automate d'états finis non déterministe à partir d'une grammaire hors contexte sans symbole non terminal récursif,
- ⇨ fabriquer directement l'automate d'états finis minimal à partir d'une expression régulière en construisant son arbre abstrait (algorithme de Glushkov).

III.2.5 Traitement des ambiguïtés

III.2.5.1 SEGMENTS AMBIGUS ET AMBIGUS IRRÉDUCTIBLES

Dans les situations pratiques, la juxtaposition des syllabes entraîne l'existence de cas où plusieurs découpages sont possibles[1]. Nous appellerons **segment ambigu** toute chaîne de caractères pouvant se découper de plusieurs façons en syllabes[2,3]. Par exemple, la phrase thaïe[4] ผมตากลม, contient trois syllabes qui peuvent être ผม ตา กลม ou ผม ตาก ลม. Ces deux découpages ont un sens : « j'ai les yeux globuleux » (ผม ตา กลม) ou « je prends l'air » (ผม ตาก ลม).

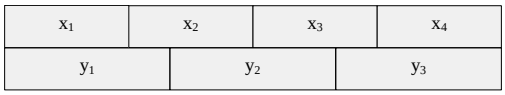


Figure 50 : Exemple de segment ambigu

Pour analyser le problème, adoptons le modèle de syllabes simplifié suivant[5] :

1 :	CV	3 :	CCV
2 :	CVC	4 :	CCVC

Si l'on fait une concaténation de deux syllabes, cela donne seize (2^4) combinaisons possibles parmi lesquelles huit sont ambiguës :

11 :	CV|CV		31 :	CCV|CV	
12 :	CV|CVC		32 :	CCV|CVC	
13 :	CV|CCV	= 21	33 :	CCV|CCV	= 41
14 :	CV|CCVC	= 22	34 :	CCV|CCVC	= 42
21 :	CVC|CV	= 13	41 :	CCVC|CV	= 33
22 :	CVC|CVC	= 14	42 :	CCVC|CVC	= 34
23 :	CVC|CCV		43 :	CCVC|CCV	
24 :	CVC|CCVC		44 :	CCVC|CCVC	

Les ambiguïtés proviennent donc, dans ce cas, de l'existence de syllabes sans consonne finale et de syllabes avec un groupe consonantique à l'initiale.

Bien qu'il soit très rare, l'exemple de segment ambigu suivant montre qu'il n'y a pas forcément le même nombre de syllabes dans les différentes décompositions possibles et que l'explication précédente n'est pas générale. En laotien, le mot ແຫລະ, qui est constitué du groupe consonantique ຫລ (hl) et de la voyelle ແ◌ະ (è bref), peut aussi se lire comme deux syllabes : ແຫ (hè) et ລະ (la, a bref).

[1] Ce n'est pas forcément le cas. Par exemple, les syllabes birmanes terminées par une consonne non souscrite sont toujours suivies du « signe qui tue » ou d'une consonne souscrite. À cause de leurs groupes consonantiques, les syllabes khmères, à l'instar des syllabes thaïes et laotiennes, connaissent ce problème de découpage. Cependant, en khmer comme en birman, une syllabe peut ne pas coïncider avec une forme écrite « normale ». Par exemple, ကုမ္ပဏီ (kum-**p**a-ni : compagnie), en birman, et អង្គរ (ang-**k**or : Angkor), en khmer, commencent par une souscrite.

[2] Les ambiguïtés dans les chaînes de caractères ont été étudiées dans le cadre de la théorie des codes (par exemple, Sardinas et Patterson 1953, cité dans [Berstel et Perrin 1985]) dans une optique inverse de la notre puisque ne s'intéressant pas aux non-codes que sont justement nos segments ambigus.

[3] Les segments ambigus doivent être traités au niveau linguistique (lexical, syntaxique, sémantique, ...) comme nous l'avons vu au chapitre III.2.2.2.

[4] Exemple donné par M. Gilles Delouche.

[5] C=Consonne, V=Voyelle, CC=Groupe consonantique (Consonant Clusters).

Ce cas provient de l'existence de groupes consonantiques et de voyelles réparties avant et après la consonne initiale, chaque partie existant aussi séparément[1].

L'ensemble des syllabes pouvant être trouvées dans un texte donné peut être représenté comme un ensemble de couples (i,j) d'indices des caractères de la chaîne associée, i étant le premier caractère de la syllabe et j le dernier. Certaines de ces syllabes — celles représentées en pointillés dans la figure ci-dessous — ne peuvent pas faire partie du texte car elles ne sont pas rattachées au début ou à la fin du texte par d'autres syllabes. Nous les appellerons **syllabes mortes**[2].

L'ensemble des couples (i,j) forme un graphe. Un découpage du texte en syllabes sera donc un chemin dans ce graphe qui part du début et se termine à la fin. Nous appellerons **point de convergence** un début ou une fin de syllabe tel que tout découpage possible du texte passe par lui. Le début et la fin du texte sont donc des points de convergence particuliers. Enfin, nous appellerons **œil** l'ensemble des chemins reliant deux points de convergence adjacents.

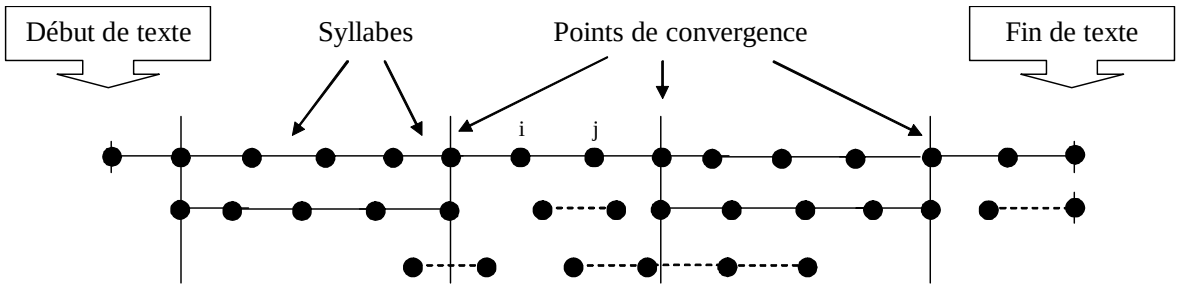


Figure 51 : Graphe de syllabisation – Points de convergence

Un texte sera donc constitué d'une suite d'yeux qui pourront être des segments non ambigus (1) ou ambigus (2). Cette décomposition est unique par construction.

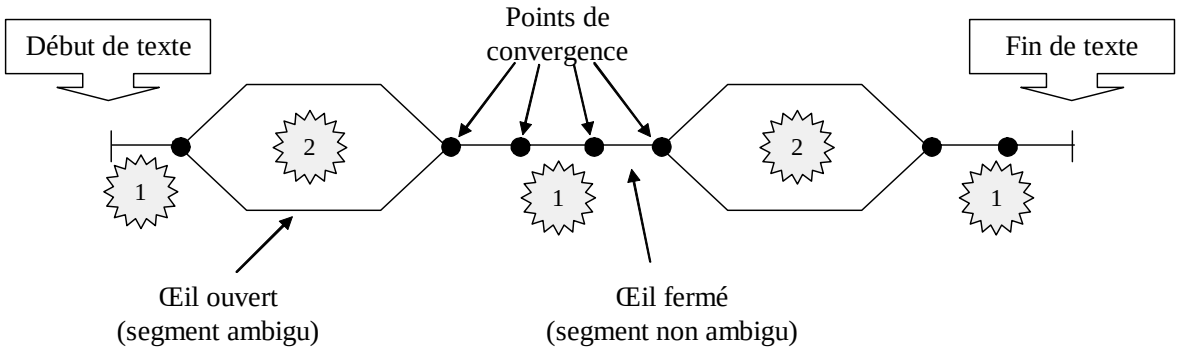
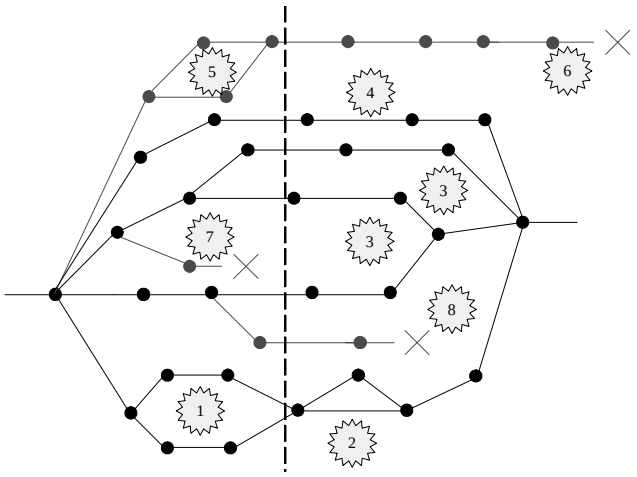


Figure 52 : Graphe de syllabisation – Segments ambigus et non-ambigus

[1] Ce cas est l'équivalent pour les syllabes à celui du concept des mots inconnus cachés (hidden unknown words) présenté dans [Kawtrakul et al. 1997].
[2] Cette notion de syllabe morte est différente de celle, phonologique, définie au chapitre II.1.1.2.

Les différents types de décomposition possible d'un œil ouvert en syllabes sont résumés dans la figure ci-dessous[1]. Nous y avons indiqué les **branches mortes**[2] que nous rencontrons en pratique lorsque, pour trouver les yeux, nous progressons de la gauche vers la droite en calculant, à chaque étape, toutes les syllabes pouvant partir du point final précédent sur la branche.



1- Segment ambigu avec le même nombre de syllabes dans les différentes branches
2- Segment ambigu avec un nombre différent de syllabes dans les différentes branches
3- Segment ambigu appartenant à AI[3]
4- Segment ambigu n'appartenant pas à AI
5- Segment faussement ambigu
6- Branche morte longue (générant un segment faussement ambigu)
7- Branche morte courte à une seule syllabe (ne générant pas de segment faussement ambigu)
8- Branche morte courte à plusieurs syllabes (ne générant pas de segment faussement ambigu)

Figure 53 : Graphe de syllabisation – Détails à l'intérieur d'un segment ambigu

La barre verticale discontinue coupe les différentes syllabes pouvant exister autour d'un caractère donné du texte[4]. Dans notre exemple, la barre verticale discontinue coupe huit syllabes dont six appartiennent à des branches vivantes — branches qui ne conduisent pas à des non-syllabes — et deux à des branches mortes[5]. Le nombre maximum de ces syllabes, que nous appellerons **degré d'ambiguïté syllabique du caractère**, est borné. Si l'on appelle *lmin* et *lmax* les nombres minimum et maximum de caractères que peut contenir une syllabe dans une langue et un système d'écriture donnés, cette **borne d'ambiguïté syllabique** vaut B_{AS} = **(*lmax**(*lmax*+1) - *lmin**(*lmin*-1)) / 2** [6].

On voit sur la figure que les branches mortes peuvent générer des fausses ambiguïtés (zone 5). Cela peut poser problème si l'on souhaite détecter localement les segments ambigus (sans partir du début du texte), ce qu'on est tenté de faire, par exemple, pour traiter le « problème du double-clic » (voir le chapitre suivant).

La notion d'œil est, en général, contextuelle, en ce sens qu'un œil est défini par rapport à une chaîne de caractères particulière. Nous nous sommes demandé s'il existait, pour un système d'écriture donné, un ensemble « **d'yeux intrinsèques »** ou **métasyllabes** qui restent des yeux quelles que soient les chaînes de caractères pouvant se trouver à droite et à gauche. L'intérêt de l'existence de telles métasyllabes réside dans le fait qu'il serait alors possible de déterminer localement les branches mortes et de les éliminer sans devoir aller jusqu'au début et à la fin du texte.

[1] Le texte est supposé bien orthographié.
[2] Il s'agit des branches constituées de syllabes mortes ; elles sont grisées et terminées par une croix. Par exemple, la chaîne laotienne ອດາຍ, peut commencer par deux syllabes, ອ (ou) et ອດ (oud), mais seule la première est possible car ດາຍ (*dai*̊) est une syllabe alors que າຍ (qu'on pourrait transcrire *ai*̊) n'en est pas une.
[3] AI = Ambigu Irréductible (voir plus bas).
[4] La figure représente un cas particulier dans lequel seules les syllabes trouvées en partant de la gauche apparaissent.
[5] Nous remarquons sur ce graphe que les branches mortes peuvent, théoriquement, n'être détectées qu'à une distance relativement importante de la barre verticale discontinue.
[6] En effet, il pourra y avoir *l* syllabes de longueur *l* contenant le caractère considéré (*l* positions possibles pour ce caractère dans la syllabes). Donc $B_{AS} = \Sigma_{lmin,lmax}\ l = \Sigma_{1,lmax}\ l - \Sigma_{1,lmin-1}\ l = lmax*(lmax+1)/2 - lmin*(lmin-1)/2$.

Cela nous a conduit à nous intéresser aux **segments ambigus irréductibles**, segments ambigus ne contenant pas de segment ambigu plus court. Par exemple, la chaîne laotienne[1] ກາມອວກວັນ peut se décomposer de deux façons possibles : ກາ-ມອວ-ກວັນ (ka-muao-kuan) et ກາມ-ອວກ-ວັນ (kam-vuak-van)[2] et aucune sous-chaîne formant une suite de syllabes n'est ambiguë : ມອວ-ກວັນ, ອວກ-ວັນ, ກາມ-ອວກ et ກາ-ມອວ. L'ensemble des segments ambigus irréductibles est noté AI. Nous avons montré que cet ensemble AI forme un langage rationnel et donné une méthode pour calculer son automate minimal (voir l'annexe A.11). Cet automate est suffisant pour reconnaître les segments ambigus irréductibles, ce qui nous évite ainsi le calcul, plus complexe, du graphe associé.

Nous continuons d'étudier le rôle de l'ensemble AI dans la décomposition des yeux en syllabes mais son « applicabilité » reste, au moment où nous rédigeons ce mémoire, à étudier et à expérimenter.

III.2.5.2 PROBLÈME DU « DOUBLE-CLIC »

Lorsque l'on double-clique sur un caractère ou, plus généralement, lorsque l'on utilise la souris pour sélectionner du texte écrit dans un système d'écriture non segmenté, il est souhaitable que la partie de texte sélectionnée ait un sens, c'est à dire qu'elle corresponde à une entité linguistique naturelle, par exemple un mot ou une syllabe. Nous nous intéressons ici au cas où la partie sélectionnée est une syllabe. Un algorithme simple peut alors déterminer l'ensemble des syllabes possibles autour du caractère situé sous le pointeur souris, puis, en cas d'ambiguïté, un algorithme devra éliminer les syllabes parasites (syllabes mortes) et élire une syllabe parmi les possibles.

Une première possibilité pour éliminer les syllabes parasites de la liste des syllabes consiste à calculer le graphe de syllabisation depuis le début du texte ou à partir d'un élément indiquant un début de syllabe de manière non ambiguë puis à en supprimer les branches mortes pour ne garder que les syllabes possibles. Pour les trois raisons suivantes, nous préférons une méthode heuristique :

- ⇨ comme cela est suggéré dans la figure précédente, les branches mortes sont, au moins théoriquement, à une distance non bornée du caractère sous le pointeur souris (« caractère de base »),
- ⇨ des fautes de frappes peuvent fausser le graphe, par exemple supprimer le point de convergence,
- ⇨ le calcul du graphe est coûteux en temps.

[1] La chaîne est théorique. De plus, les syllabes de la forme ◌ວວ, bien que théoriquement possibles et conformes au modèle général, n'existent pas. Cependant, le cas peut se rencontrer lorsque le modèle syllabique, comme celui du III.2.3.2, ne contient pas toutes les limitations de la langue, et dans le cas d'une écriture quelconque.

[2] Syllabes possibles en tête : ກາ et ກາມ, en queue : ວັນ et ກວັນ.

L'heuristique que nous proposons consiste à lister toutes les syllabes candidates autour du caractère de base[1] et à supprimer les branches mortes trouvées dans un voisinage borné autour de ce caractère. Pour déterminer toutes les syllabes candidates, une technique simple consiste à remplir une table de vérité représentant les caractères autour du caractère de base. Lorsque plus d'une syllabe est trouvée, il y a ambiguïté, vraie ou fausse. Dans l'exemple suivant, correspondant à des syllabes pouvant être constituées de deux à sept caractères (cas du laotien avec les encodages les plus courants), nous avons supposé que deux syllabes étaient possibles :

- ⇨ une syllabe commençant deux caractères avant le caractère de base et finissant un caractère après,
- ⇨ une syllabe commençant deux caractères avant le caractère de base et finissant deux caractères après.

	-6	-5	-4	-3	-2	-1	0
0							
1					x		
2					x		
3							
4							
5							
6							

Figure 54 : Table de vérité des syllabes candidates

Après avoir listé toutes les syllabes candidates, les branches mortes sont détectées en examinant, pour chacune d'elles, si leur voisinage immédiat, à droite et à gauche, est formé de syllabes, et en recommençant ce processus un nombre borné de fois. Dans l'exemple ອຸດາຍ, donné précédemment, l'initialisation de la table de vérité indiquera les deux syllabes, ອຸ (ou) et ອຸດ (out), et l'algorithme de détection des branches mortes éliminera la seconde candidate car son voisinage droit — າຍ… — ne peut pas commencer par une syllabe.

Une fois les syllabes parasites éliminées, et en l'absence d'autres traitements de désambiguïsation, nous proposons de sélectionner la syllabe la plus longue. Ce choix étant en partie arbitraire, une commande simple (raccourci clavier, touche de fonction...) doit permettre de faire défiler les différentes possibilités à l'issue du double-clic.

Les principes proposés ici pour le double-clic s'étendent à la sélection de texte à la souris (appui sur le bouton [gauche sous Windows] de la souris → déplacement de la souris → relachement du bouton souris). Ils sont mis en œuvre dans LaoWord. Le défilement des différentes syllabes possibles tient alors compte de l'extrémité active de la sélection :

- ⇨ défilement des syllabes ambiguës à gauche lorsque la sélection va de la droite vers la gauche,
- ⇨ défilement des syllabes ambiguës à droite lorsque la sélection va de la gauche vers la droite.

[1] Avec les notations du chapitre précédent, le nombre de ces syllabes candidates est borné par (lmax*(lmax+1) - lmin*(lmin-1)) / 2.

III.2.6 Sylla, un outil de mise au point des grammaires syllabiques

Pour réaliser la segmentation syllabique, nous avons développé Sylla, un prototype d'outil linguiciel au sens du III.2.1. Il est téléchargeable à l'adresse cams-atid.ivry.cnrs.fr/Syllabication. Sylla est destiné à la mise au point de modèles syllabiques. Il s'appuie sur le formalisme grammatical décrit précédemment.

Dans la fenêtre de droite (« Syllables Grammar »), la grammaire des syllabes est entrée, soit directement, soit par copie à partir d'un fichier texte préparé par ailleurs[1]. Après compilation de la grammaire, il est possible de tester si une syllabe est valide (en bas à gauche : « *Check syllable* ») ou de segmenter un texte (au dessus à gauche : « *Segment text* »). L'algorithme implémenté pour la segmentation du texte calcule le degré d'ambiguïté syllabique de chaque caractère, après élimination des branches mortes (voir le chapitre III.2.5.1). Un segment ambigu correspondra à une suite de caractères de degré supérieur à un. Dans la fenêtre de segmentation du texte, les segments ambigus sont colorés en bleu sans être segmentés.

La réussite d'une compilation est indiquée par le passage en vert de la grammaire. Dans le cas contraire, la grammaire passe en rouge à partir du point où l'erreur de syntaxe a été détectée, ce qui permet de trouver l'erreur plus facilement et d'annoncer la fin de la compilation, même lorsqu'elle a échoué.

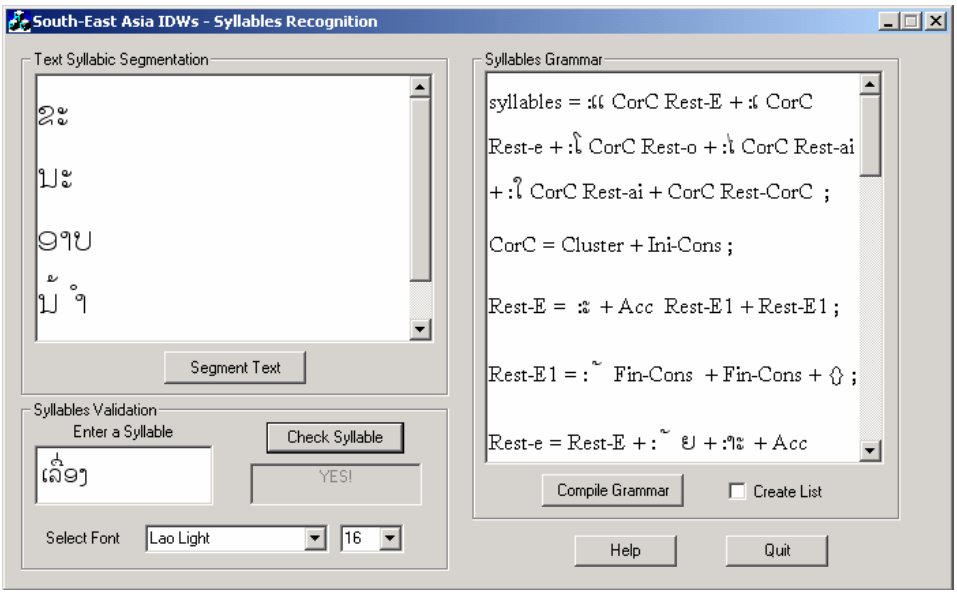


Figure 55 : Interface utilisateur de Sylla

Dans la version actuelle de l'outil, la compilation de la grammaire génère un arbre lexicographique[2] (*trie*) qui est alors utilisé pour reconnaître les syllabes (un fichier texte contenant la liste des syllabes peut être généré lors de la compilation). Il a été testé avec les grammaires du birman, du khmer (en version simplifiée), du laotien et du thaï présentées précédemment. Une aide en ligne, en version préliminaire, donne les explications de base pour l'utilisation du logiciel.

[1] La forme des grammaires de syllabes est définie par une grammaire des grammaires (voir annexe A.11).

[2] Sylla est écrit en C++. La liste des syllabes et l'arbre lexicographique pourraient être créés de façon plus « naturelle » à partir d'un programme en Prolog. La partie « génération » des syllabes a d'ailleurs été vérifiée à l'aide d'une liste de syllabes générée avec ce langage. Le code en est donné en annexe A.14.

Les créateurs de grammaires ont actuellement les contraintes suivantes :

- ⇨ en l'absence de standard d'encodage, la grammaire créée sera spécifique à un seul encodage, celui de la police qu'ils utilisent pour décrire la grammaire,
- ⇨ dans les cas où plusieurs chaînes de caractères différentes produisent le même rendu visuel (saisie non univoque), la grammaire créée ne sera applicable qu'à une forme canonique du texte[1].

Ces deux difficultés ont été mentionnées aux chapitres II.1.2.1 et II.1.2.2. Elles ont été résolues pour le laotien grâce à une fonction de mise en forme canonique (voir II.2.4.1). Cette fonction n'a pas encore été généralisée et il est donc supposé, dans la version actuelle de Sylla, que les textes sont standardisés d'une manière externe avant d'être testés. Une extension de Sylla pourra prendre en compte ces deux difficultés.

Plusieurs autres améliorations de Sylla sont prévues :

- ⇨ vérification sur la grammaire avant la compilation :
  - o non récursivité des règles de la grammaire,
  - o présence de tous les non terminaux en partie gauche des règles,
  - o présence des non terminaux en partie gauche d'une seule règle,
- ⇨ assouplissement de la grammaire des grammaires :
  - o autorisation des commentaires,
  - o autorisation de la numérotation des règles,
- ⇨ amélioration de la documentation.

Un autre problème résiduel est la limitation de l'outil à des grammaires « pas trop productives ». Cela provient du principe même de construction de l'automate reconnaissant les syllabes qui génère un arbre lexicographique dont la taille peut dépasser celle des disques durs (voir le chapitre III.2.4). L'intégration des développements réalisés par Yeu ([Yeu 2003]) — création de l'automate minimal correspondant à la grammaire — résoudra prochainement ce problème.

La génération automatique du code C++ réalisant les diverses fonctions de sélection — reconnaissance des syllabes et segmentation locale (algorithme « du double-clic ») — reste à définir et à réaliser.

[1] Plusieurs suites de caractères peuvent donner le même résultat visuel, par exemple, en birman : ◌ျ + ◌် = ◌် + ◌ျ = ◌်ျ, en laotien : ໆ = ◌ ◌ + ◌ ◌. La répétition des caractères sans avance (accents, voyelles suscrites ou souscrites) donne aussi le même résultat visuel, puisqu'ils se superposent. Ces caractères pouvant être répétés un nombre quelconque de fois, comme dans l'exemple suivant (birman) : ◌် = ◌် + ◌် = ◌် + ... + ◌်, le nombre des syllabes est potentiellement infini. Dans le but d'éviter cela et de garder des règles non récursives, la suppression des caractères répétés a été intégré au prétraitement de mise en forme canonique (voir II.2.4.1.1).

III.2.7 Un traitement de texte pour un groupe de langues : GMSWord

Dans ce chapitre, nous montrons qu'il est possible de généraliser rapidement le logiciel LaoWord aux systèmes d'écritures non segmentés d'Asie du Sud-Est dérivés de l'écriture pallava et décrits au chapitre III.2.2.1, en appliquant les principes du chapitre III.1.3. Nous nous limiterons ici à montrer la mise en œuvre de la méthode sur l'exemple de la composante « sélection » de LaoWord (sélection à la souris et au clavier, voir le chapitre II.2.2), qui en est probablement la composante la plus complexe. Nous avons vu par ailleurs au chapitre III.1.1.3 que l'adaptation de LaoWord à la saisie du bengali (BanglaWord) s'est faite très facilement.

Comme les systèmes d'écriture considérés sont utilisés au Cambodge, au Laos, au Myanmar, en Thaïlande, au Vietnam et au Yunnan, nous avons appelé GMSWord cette généralisation de LaoWord, GMS étant l'acronyme de *Greater Mekong Subregion*, l'ensemble des pays participant au réseau universitaire et de recherche GMSARN (http://www.ait.ac.th/gmsarn/). La limitation à ces systèmes d'écriture est artificielle et les mêmes techniques peuvent aussi s'appliquer, *a priori*, aux autres systèmes d'écriture non segmentés comme ceux mentionnés en note à la section III.2.2.1, que nous n'avons pas encore étudiés.

Notons que, outre son utilité pour réaliser les sélections, la segmentation en syllabes peut être appliquée au tri lexicographique, au calcul de transcriptions phonétiques et à la segmentation de textes en mots, en particulier en présence de mots inconnus, comme cela a été montré pour le laotien aux sections II.2.3, II.2.4.3 et II.3.1.2.

L'architecture proposée[1] s'appuie à la fois sur la forme générique qui a été dérivée de LaoWord pour réaliser BanglaWord et sur les grammaires mises au point grâce à Sylla ainsi sur que le code généré (dès que cette génération sera réalisée). L'interface d'accueil et les algorithmes de sélection — automate des événements, sélection sur double-clic, sur mouvement de souris, au clavier à l'aide des flèches — sont repris de LaoWord. Ces sélections reprennent les principes utilisés pour les mots par Microsoft Word :

- ⇨ sélections à la souris :
  - **double-clic** : selectionne une syllabe,
  - **bouton gauche (Windows) enfoncé + mouvement souris** : selectionne le texte syllabe par syllabe,
- ⇨ sélections au clavier :
  - **Ctrl +** → : déplace le curseur au début de la syllabe suivante,
  - **Ctrl +** ← : déplace le curseur au début de la syllabe précédente,
  - **Ctrl + Maj +** → : ajoute la syllabe suivante à la sélection courante,
  - **Ctrl + Maj +** ← : ajoute la syllabe précédente à la sélection courante,
  - **Ctrl + Suppr** : supprime la syllabe suivant la sélection courante,
  - **Ctrl + Backspace** : supprime la syllabe précédant la sélection courante.

[1] Non implémentée. Nous prévoyons de l'implémenter après la soutenance de cette thèse.

L'interface $I_{LG/LS}$ sera implémentée par une fonction *bool SyllabeGMS(langue,char*,debut,fin)* prenant en entrée une chaîne de caractères et renvoyant *VRAI* ou *FAUX* selon que la chaîne de caractères est une syllabe valide ou non pour la langue et le système d'écriture considérés.

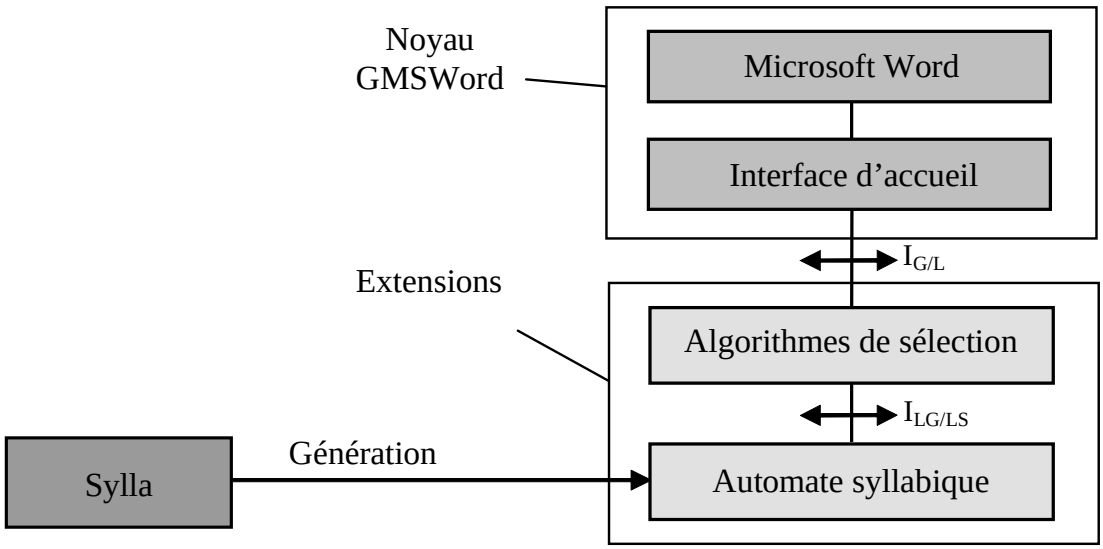


Figure 56 : Architecture du chapitre III.2.1 dans le cas de GMSWord

Les algorithmes de sélection de LaoWord devront être adaptés pour reconnaître le système d'écriture utilisé et déterminer le paramètre « langue » à passer à la fonction. Pour cela, nous supposerons que les textes sont en Unicode, ce qui ne pose pas de problème pour le birman, le dehong dai (tai nüa), le khmer, le laotien et le thaï. Pour les autres, qui ne sont pas actuellement prévus dans Unicode, nous nous proposons de leur affecter les plages suivantes dans la zone d'usage privé Unicode (voir I.1.4.1) :

- ⇨ cham: E000 à E0FF,
- ⇨ tai dam : E100 à E1FF,
- ⇨ tai khao : E200 à E2FF,
- ⇨ tai deng : E300 à E3FF,
- ⇨ tai yo : E400 à E4FF,
- ⇨ lai pao : E500 à E5FF,
- ⇨ môn : E600 à E6FF,
- ⇨ khamti : E700 à E7FF,
- ⇨ shan : E800 à E8FF,
- ⇨ tai mau : E900 à E9FF,
- ⇨ tham laotien : F000 à F0FF,
- ⇨ tham isan : F100 à F1FF,
- ⇨ lü : F200 à F2FF,
- ⇨ lanna : F300 à F3FF,
- ⇨ khün : F400 à F4FF.

Connaissant ces plages de code, les algorithmes de sélection appelleront la fonction *SyllabeGMS* avec le bon paramètre. Ils devront, par ailleurs, être modifiés pour traiter les spécificités des différents systèmes d'écriture, c'est-à-dire les deux difficultés rappelées au chapitre III.2.6 : nombre maximum de caractères des syllabes et mise en forme canonique.

Notons qu'un important travail préalable de définition et de réalisation des polices est à réaliser pour ces systèmes d'écriture. Le travail de définition (tables de correspondance codes-caractères, liste des glyphes, règles de réarrangement codes-glyphe et de saisie...) pourra s'appuyer sur les solutions mises en œuvre pour les systèmes d'écriture déjà traités (thaï, khmer...).

En résumé, les algorithmes de sélection auront à traiter les événements clavier ou souris et à réaliser l'affichage de la sélection du texte. Les événements (messages Windows) seront gérés par un automate provoquant l'appel de ces traitements dans les cas suivants (voir le principe de mise en oeuvre de l'interception des messages Windows au chapitre III.1.3.1 et celui de l'automate des événements souris[1] au chapitre II.2.2) :

- ⇨ WM_LBUTTONDBLCLK (double clic),
- ⇨ WM_MOUSEMOVE (mouvement de souris), le bouton gauche de la souris étant enfoncé,
- ⇨ WM_KEYDOWN (appui touche), avec les combinaisons :
  - o Ctrl + → et Ctrl + ←,
  - o Ctrl + Maj + → et Ctrl + Maj + ←,
  - o Ctrl + Suppr et Ctrl + Backspace.

L'algorithme réalisant ces traitements est le suivant[2].

```
Affichage(Chaine,i,evt)
// Chaine est une chaîne de caractères Unicode contenant le texte
// i est l'indice du caractère concerné par l'événement
// evt est l'événement ayant causé l'appel
{
   FormeCanonique(Chaine)              // Mise en forme canonique
   lg ← langue(Chaine[i])              // Détermination de la langue et du système d'écriture
   Tableau ← faux                      // Initialisation du tableau des syllabes candidates (III.2.5.2)
   pour j = Lmin[lg] à Lmax[lg]        // Lmin et Lmax : longueurs min et max des syllabes
      pour k = 0 à j-1
         si (SyllabeGMS(lg,Chaine,i-j+k+1,i+k)) Tableau[i-j+k+1,i+k] ← vrai
   EliminerBranchesMortes(Chaine,i,Tableau)  // Voir le chapitre III.2.5.2
   Afficher(i,Tableau,evt)          // L'affichage dépend de l'événement ayant causé l'appel[3]
}
```

L'automate et la fonction *Affichage* sont des compléments linguistiques génériques (algorithmes de sélection) et la fonction *SyllabeGMS* est un complément linguistique spécifique pour une langue donnée. L'interface $I_{LG/LS}$ est matérialisée par l'appel de cette fonction qui, avec *FormeCanonique*, *Lmin* et *Lmax*, sont les seuls éléments utilisés par la fonction *Affichage* qui dépendent de la langue et du système d'écriture[4]. La reconnaissance des syllabes intégrée dans la fonction *SyllabeGMS* sera réalisée par le code généré par Sylla pour chaque système d'écriture à partir des grammaires.

---

[1] Un automate des événements est nécessaire, en plus du seul appel de la fonction *Affichage* sur les événements WM_LBUTTONDBLCLK et WM_MOUSEMOVE, afin de substituer entièrement un automate GMSWord à l'automate de Word. Par exemple, l'événement WM_LBUTTONDOWN activera le « glisser-souris » et provoquera l'initialisation des variables associées (début et fin de la sélection), l'événement WM_LBUTTONUP provoquant les effets inverses à la fin de l'appui sur le bouton gauche de la souris.
L'automate général des événements se déduit simplement de l'automate des événements souris en y ajoutant la gestion des événements clavier (Ctrl + →, ...).

[2] Cet algorithme est très simplifié. Par exemple, il ne tient pas compte des effets de bord rencontrés, par exemple en début ou en fin de texte, du fait que le pointeur souris peut se trouver sur des éléments non textuels (images...) ou hors du document principal, du comportement particulier de certains objets (tableaux, commentaires...), de la présence de nombres, de ponctuations et de caractères d'autres systèmes d'écriture, ou encore de la modification de la valeur de *i* lors de la mise en forme canonique.

[3] Par exemple, l'effet du double clic est symétrique alors que celui du « glisser-souris » ne l'est pas puisque l'extrémité de la sélection est soit à droite soit à gauche.

[4] Ces éléments sont aussi utilisés dans la fonction *EliminerBranchesMortes* qui consiste à éliminer les syllabes déclarées candidates dans le tableau lorsque leurs voisinages gauche ou droit sont impossibles, en particulier lorsqu'ils forment des non-syllabes.

III.3 ESQUISSE DE LIVRE BLANC POUR L'INFORMATISATION D'UN GROUPE DE LANGUES

III.3.1 Présentation

Dans ce chapitre, nous tentons d'appliquer tout ce qui précède à une « étude de cas », et plus précisément à de ce que pourrait être un grand projet d'informatisation des langues mené par les Nations Unies. Ainsi, nous supposerons, comme au III.2.1, que nous voulons informatiser un nombre important de langues-π ou μ — quelques dizaines, voire davantage — ce qui pose le problème des critères à retenir pour choisir les langues à informatiser en priorité.

Parmi ces critères, le nombre de locuteurs a jusqu'à maintenant été déterminant dans les processus d'informatisation. Ainsi, des langues complexes à informatiser mais parlées par de nombreux locuteurs[1] sont d'ores et déjà bien informatisées alors que d'autres langues, techniquement beaucoup plus simples, sont en retard. Roland Breton considère ([Breton 2003], page 36) que les langues parlées par moins d'un million de locuteurs ont une existence précaire (entre 100 000 et 1 000 000 locuteurs), immédiatement menacée (entre 10 000 et 100 000 locuteurs), voire déjà moribonde (moins de 10 000 locuteurs). Elles sont pourtant plus de cinq mille, soit la quasi-totalité des langues du monde.

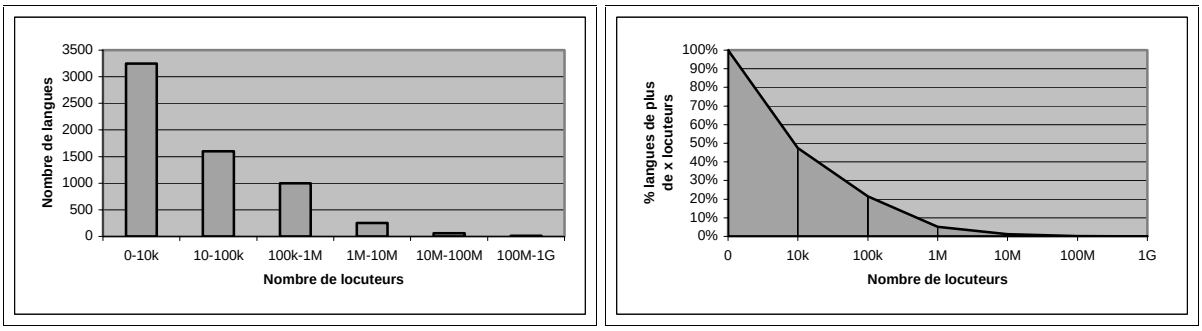


Figure 57 : Nombre de langues et nombres de locuteurs (d'après [Breton 2003])

L'informatisation de ces langues est dans une situation comparable. Si nous reprenons l'exemple d'Office XP (voir le chapitre I.1.4.2), seules deux des 48 langues prises en charge comptent moins d'un million de locuteurs : le **basque** et le **gallois** (580 000 chacune), et cinq moins de cinq millions : l'**estonien** (1 100 000), le **letton** (1 500 000), le **slovène** (2 000 000), le **galicien** (4 000 000) et le **lituanien** (4 000 000).

Notons néanmoins que seules 41 des 130 langues parlées par plus de cinq millions de locuteurs[2] font partie des quarante-huit prises en charge par cette suite bureautique, ce qui traduit l'existence d'autres critères pour le choix des langues à informatiser.

[1] Citons le chinois dont le système d'écriture idéographique pose des problèmes de saisie et de représentation ou encore le russe dont la morphologie est complexe.

[2] Ce nombre est tiré de l'annexe A.6 et suppose que toutes les langues de plus de cinq millions de locuteurs y figurent (ordre de grandeur).

Le choix de ces langues peut ainsi être guidé par des critères divers tels que :

- ⇨ le nombre de locuteurs,
- ⇨ le caractère officiel ou national de la langue,
- ⇨ le caractère central de la langue (voir I.2.2.1.2),
- ⇨ l'intérêt des populations pour des moyens informatiques dans leur langue,
- ⇨ la motivation des bailleurs pour l'informatisation d'une langue,
- ⇨ le niveau d'informatisation de la langue (indice-σ),
- ⇨ l'existence d'une grammaire et d'un dictionnaire,
- ⇨ l'existence d'une langue-τ proche,
- ⇨ la présence d'un bilinguisme permettant de faciliter la communication.

L'annexe A.6 est destinée à faciliter le choix des langues à informatiser. Elle donne, pour environ huit cents langues :

- ⇨ le nombre de locuteurs, ce qui fournit une indication sur leur possible situation informatique,
- ⇨ la famille de langues, ce qui permet d'identifier les langues proches,
- ⇨ le code Ethnologue, ce qui donne accès à des informations sur la langue[1] (autres noms...).

En complétant cette liste avec les critères retenus pour le choix des langues, nous obtenons un tableau d'aide au choix des langues à informatiser.

	Langue	Locuteurs	Famille	Ethno.	Off./Nat.	Indice-σ	Dict.	Intérêt	Bailleurs
1	abkhaze	105 000	nord-caucasienne	ABK					
2	aceh	3 000 000	austronésienne	ATJ					
3	achi, cubulco	45 000	maya	ACC					
4	achi, rabinal	37 300	maya	ACR					
5	acoli	773 800	nilo-saharienne	ACO					
6	adangme	825 900	nigéro-congolaise	DGM					
7	adygh	300 000	nord-caucasienne	ADY					
8	afar	1 579 000	afro-asiatique	AFR					
9	afrikaans	6 381 000	indo-européenne	AFK					
10	agariya	55 757	austro-asiatique	AGI					
11	aguacateco	18 000	maya	AGU					
12	akan	7 000 000	nigéro-congolaise	TWS					
13	albanais (gheg)	2 000 000	indo-européenne	ALS					
14	albanais (tosk)	3 000 000	indo-européenne	ALN					
15	aléoute	305	esquimo-aléoute	ALW					
16	allemand	100 000 000	indo-européenne	GER					
17	amharique	17 413 000	afro-asiatique	AMH					

Figure 58 : Exemple de tableau d'aide au choix des langues à informatiser

Ce choix étant fait, nous supposerons que nous opérons, pour chacune des langues, en deux phases :

- ⇨ première phase, visant la couverture des services qui nous paraissent les plus importants (voir la liste entière des services envisagés au chapitre I.1.1.2) :
  - traitement du texte, à l'exception des correcteurs grammaticaux et stylistiques,
  - ressources dictionnairiques (de type Papillon), avec un objectif de cinq mille articles pour chaque langue,
  - systèmes simples (de type Montaigne / LaoLex) d'aide à la traduction humaine,
- ⇨ deuxième phase, visant la couverture des autres services jugés critiques par les populations.

[1] Pour accéder à ces informations, aller à http://www.ethnologue.com/show_language.asp?code=XXX, où XXX est le code Ethnologue fourni dans l'annexe. D'autres informations sur d'éventuels points de contact et listes de diffusion existants sont proposées pour quelques langues sur la page http://www.evertype.com/langlist.html.

Pour réaliser efficacement un tel projet, mais aussi pour en maîtriser la complexité, nous proposons de le morceler en tâches élémentaires, conduisant à :

- ⇨ une organisation technique, suivant l'architecture du chapitre III.2.1,
- ⇨ une organisation chronologique, suivant une logique d'enchaînement des tâches.

III.3.2 Organisation technique en projets

Nous avons présenté au chapitre III.2.1 une architecture comprenant cinq parties (plates-formes d'accueil standardisées incluant un logiciel de base et un complément général, compléments linguistiques génériques, compléments linguistiques spécifiques, outils linguiciels) séparées par les interfaces $I_{G/L}$, $I_{LG/LS}$ et *génération*.

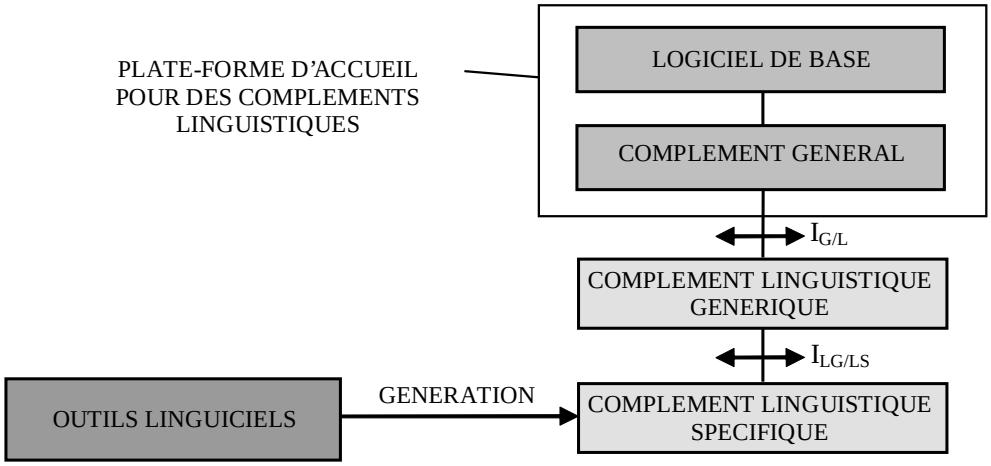


Figure 59 : Rappel de l'architecture du chapitre III.2.1

Nous proposons d'attribuer les **compléments généraux** aux **éditeurs des logiciels de base** associés. Ils sont, en effet, les mieux placés pour compléter correctement leurs produits[1] et en assurer efficacement les évolutions. Nous pensons que les éditeurs de logiciel accepteront de réaliser ces développements, motivés par la plus grande diffusion de leur offre dans un contexte très concurrentiel[2]. Notons que cela n'est possible que tant qu'un nombre important d'utilisateurs potentiels restera à atteindre. Les cinq mille langues les moins parlées concernant moins de deux cents millions de personnes ([Breton 2003] page 36), il faudra s'appuyer sur celles parlées par plusieurs millions de locuteurs chacune pour que les développements des compléments généraux couvrent aussi ces cinq mille « petites langues », à moins d'instituer une obligation de service universel au niveau mondial pour les éditeurs de logiciels comme cela peut exister pour l'électricité, le téléphone, les services postaux, ferroviaires, bancaires...

Les **compléments linguistiques génériques** sont des modules destinés à répondre à des besoins généraux pour certains types de langues ou de systèmes d'écriture (un module pour chaque besoin identifié) comme, par exemple, la partie générique de la sélection syllabique présentée au chapitre II.2.2 puis rappelée au III.1.3.2 et au III.2.7. Ils doivent être définis par une analyse du besoin (voir le chapitre suivant) et nous les supposerons pris en charge par un **réseau de laboratoires universitaires** réunissant des compétences informatiques et linguistiques.

[1] Par exemple, les limitations de l'API proposée par Microsoft pour les versions de Word antérieures à Word 2000 nous ont obligé à des contournements complexes pour réaliser les sélections syllabiques de LaoWord alors qu'une API adaptée aurait considérablement facilité le travail. De plus, il a fallu entièrement réécrire ces fonctions lorsque l'API de Word 2000 est arrivée.

[2] Les contacts nécessaires pourraient avoir lieu via la « commission spéciale des Nations Unies pour les technologies de l'information et des télécommunications » à laquelle participent plusieurs grands éditeurs de logiciels (voir le chapitre I.1.2.4.2).

Les **outils linguiciels** doivent permettre à des non-informaticiens de réaliser des ressources propres à une langue sans devoir suivre de formation lourde. Ils doivent être simples d'emploi et accompagnés d'une documentation claire. Nous suggérons que leur développement soit pris en charge par le même **réseau de laboratoires universitaires** que celui devant assurer le développement des compléments linguistiques génériques et qu'il soit précédé, comme ces derniers, par une analyse du besoin. Notons que ce réseau pourrait inclure des **projets multilingues** coopératifs (voir I.2.2.3) tels que Papillon pour les ressources lexicales, Montaigne / LaoLex pour les aides à la traduction humaine et Sylla pour les modèles syllabiques, projets déjà actifs dans ces domaines.

Avec les **compléments linguistiques spécifiques** se posent, dans toute leur dimension, le problème de la diversité des langues et celui de leur caractère politique. Nous ferons ici abstraction du problème politique, rappelant seulement que la *déclaration des droits des personnes appartenant à des minorités nationales ou ethniques, religieuses et linguistiques* et la *déclaration universelle de l'UNESCO sur la diversité culturelle* (voir le chapitre I.1.2.4.1) vont dans le sens d'une collaboration plus facile des États à des projets d'informatisation de langues minoritaires parlées sur leur territoire. En ce qui concerne le problème de la diversité des langues, nous proposons que les **populations linguistiques** — avec leur diaspora (voir I.2.2.5) — prennent elles-mêmes en charge les compléments propres à leurs langues, ce avec l'aide des outils réalisés par les laboratoires universitaires et en travaillant de manière coopérative comme proposé au chapitre I.2.2.3. Cette prise en charge par les populations linguistiques offre les avantages suivants :

- ⇨ ces populations connaissent bien leur langue et leurs besoins informatiques,
- ⇨ elles peuvent prendre, si besoin, des décisions sur leur langue (standardisation de l'orthographe...),
- ⇨ leur motivation pour une informatisation de leur langue est mieux identifiée, permettant d'éviter une informatisation « forcée » et inutile,
- ⇨ l'effort est ainsi mieux réparti, évitant qu'un organisme central en supporte la charge.

La réalisation des **ressources linguistiques** pose aussi le problème de la diversité des langues et, lorsque de telles ressources existent sous forme papier, la tâche est énormément facilitée. Cependant, certaines langues insuffisamment décrites — David Crystal estime qu'il en existe trois mille (voir [Calvet 2002], page 107) — nécessiteront l'élaboration d'une grammaire et d'un dictionnaire pour qu'une version exploitable informatiquement puisse en être dérivée. Nous avons vu au chapitre I.2.2.1.2 que la réalisation d'une grammaire et d'un dictionnaire coûte environ 150 000 € (trois hommes-ans). Nous supposerons ici encore une prise en charge par les **populations linguistiques**, avec l'aide d'outils tels que ceux proposés aux chapitres I.2.2.4 et II.3.1.1.

La mise en œuvre du projet pourrait donc faire intervenir cinq types d'acteurs :

- ⇨ les Nations Unies pour les aspects généraux, par exemple :
  - o bonne prise en compte de tous les systèmes d'écriture par les logiciels de base,
  - o projets de développement des ressources linguistiques (supervision),
- ⇨ des laboratoires universitaires pour les aspects « traitement des langues » :
  - o projets multilingues (Papillon, Montaigne / LaoLex, Sylla, UNL...),
  - o projets de développement des modules linguistiques génériques,
  - o projets de développement des environnements de programmation générant les modules spécifiques,
- ⇨ des instituts spécialisés en langues pour les aspects linguistiques :
  - o génériques, en collaboration avec les laboratoires d'informatique,
  - o spécifiques, en support des populations linguistiques,
- ⇨ les populations linguistiques elles-mêmes pour les aspects linguistiques spécifiques :
  - o projets de développement des ressources linguistiques (supervision par l'ONU),
  - o projets de développement des modules linguistiques spécifiques,
- ⇨ les éditeurs de logiciel pour les compléments généraux.

III.3.3 Organisation chronologique et planification

Étant donné le nombre important d'acteurs envisagé au chapitre précédent, nous proposons de mettre en place un certain nombre de moyens de communication pour leur permettre d'informer, de consulter et de dialoguer. Bien que ces acteurs puissent résider dans des pays différents, ils travailleraient effectivement ensemble, essentiellement grâce à ces moyens de communication, et beaucoup moins grâce aux classiques réunions ou séjours collaboratifs. Voici quelques-uns des moyens de communication qui pourront être utilisées tout au long du projet, en supplément du courrier électronique et d'Internet[1] qui sont des infrastructures de base maintenant largement répandues grâce, en particulier, à l'action des organismes onusiens.

- ⇨ Site Internet dédié, que nous appellerons UNITAL, sur lequel nous trouvons :
  - des explications générales sur le projet ainsi qu'un forum de discussion général,
  - des appels à participation à l'intention des éditeurs de logiciels, des instituts de langues, des laboratoires universitaires, des projets multilingues, des populations linguistiques et de divers organismes (consortium Unicode, SIL International...),
  - un espace d'expression pour chaque langue incluant un forum, la possibilité d'indiquer les éventuelles ressources existantes ou manquantes, de remplir le tableau d'évaluation du niveau d'informatisation[2], de proposer ses compétences...,
  - les logiciels et ressources réalisés dans le cadre du projet (compléments, outils pour linguistes, bases lexicales...), ainsi que les logiciels gratuits ou donnés par leurs auteurs,
  - des services en ligne tels que Papillon pour les bases lexicales et Montaigne pour les aides à la traduction humaine (outils pour linguistes ou autres),
  - un support technique et linguistique.
- ⇨ Diffusion sur les listes en traitement automatique des langues naturelles :
  Ce mode de diffusion atteint les éditeurs de logiciels, les laboratoires universitaires, les instituts de langues, les projets multilingues.
- ⇨ Diffusion par l'intermédiaire d'organismes tels que le PNUD :
  Ce mode de diffusion atteint les organismes *ad hoc* des États et certaines populations linguistiques.
- ⇨ Diffusion directe :
  Ce mode de diffusion peut être utile pour toucher certains acteurs (principaux éditeurs de logiciel, consortium Unicode, populations linguistiques, SIL International...).

Afin de réutiliser le plus possible les réflexions et développements déjà réalisés, et avant de lancer le projet lui-même, nous commençons par **recenser les besoins informatiques ainsi que les éventuelles réponses existantes**. Les motivations, l'état d'informatisation et les besoins restant à satisfaire sont compilés dans cette phase amont par un groupe de travail constitué des différents acteurs. Ce travail conduit, en particulier, à un découpage du projet en **étapes** traitant chacune un ensemble précis de langues à informatiser. Nous supposerons que la première étape contient une centaine de langues « déjà bien décrites ».
Les divers compléments nécessaires ainsi que les interfaces $I_{G/L}$, $I_{LG/LS}$ et *génération* sont alors spécifiés et l'émission d'un ensemble de recommandations, accompagnées de ces spécifications, marque le coup d'envoi du projet d'informatisation proprement dit.

[1] Le « tchat » (ICQ, Yahoo !Messenger, MSN...) est aussi un moyen de communication envisageable. Cependant, il nécessite une connexion permanente et ne peut être généralisé dans tous les cas.
[2] Pour chaque langue, un vote en ligne pourrait donner une première idée de l'indice-σ de satisfaction des populations linguistiques par rapport à l'informatisation de leur langue. Ce procédé, s'il introduit un biais du fait que la population s'exprimant sera, a priori, une minorité ayant accès à Internet, a néanmoins des chances de toucher les personnes les plus à même de juger du niveau d'informatisation de leur langue.

Langues liées

Notons que l'informatisation d'une langue peut être rendue plus efficace si elle bénéficie des acquis obtenus sur d'autres (voir I.1.2.3 et I.2.2.1). Dans la présente démarche, qui vise à informatiser de nombreuses langues, cette synergie doit s'inscrire dans la planification elle-même, les projets étant liés les uns aux autres chronologiquement lorsque les langues sont liées. En particulier, à l'intérieur d'une famille donnée, les langues les mieux disposées pourront être traitées en priorité, les autres pouvant alors tirer profit des acquis réalisés pour leur propre informatisation.

Voici une proposition pour une première phase d'un tel projet (les dates sont exprimées en années). Les moyens de communication, en particulier le site UNITAL, sont supposés exister dès le début du projet. Rappelons que l'objectif est de fournir, pour un grand nombre de langues, des traitements de texte simples (hors correcteurs grammaticaux et stylistiques) mais bien adaptés aux langues, des dictionnaires de cinq mille articles et des systèmes simples d'aide à la traduction. Pour fixer les idées, nous supposons que sont développés **dix compléments généraux** (éditeurs de logiciel [EL]), **cinq compléments linguistiques génériques** et **cinq outils pour linguistes** (réseau universitaire [RU]), **cent compléments linguistiques spécifiques** et **cent dictionnaires** (populations linguistiques [PL]), les autres tâches étant réalisées par les Nations Unies [NU] ou par des groupes *ad hoc* [GROUPE].

ETAPE 1 (cent langues, six ans) **Tâches de la phase 1**	**T0-1**		**T0+1**	**T0+2**	**T0+3**	**T0+4**	**T0+5**	**T0+6**
Travaux amont								
Préparation du projet	NU							
Développement du site web UNITAL	NU							
Travaux généraux, recensement, spécifications								
Travaux avec Unicode et les éditeurs de logiciel			GROUPE 1					
Recensement de la situation linguistique et choix des langues			GROUPE 2					
Recensement du besoin en compléments et outils			GROUPE 2					
Spécification des compléments et outils			GROUPE 2					
Travaux pour les langues insuffisamment décrites					PL			
Traitement du texte								
Développement des compléments généraux					EL			
Développement des compléments linguistiques génériques					RU			
Développement d'outils pour linguistes					RU			
Développement des compléments linguistiques spécifiques						PL		
Aides à la traduction								
Développement d'outils pour linguistes					RU			
Construction des ressources linguistiques						PL		
Ressources linguistiques								
Développement d'outils pour linguistes					RU			
Construction des ressources linguistiques							PL	

Figure 60 : Planning pour la première phase de la première étape du projet

Date	Acteurs	Contenu
T_0	Consortium Unicode, principaux éditeurs de logiciel, populations linguistiques concernées, consultants (groupe 1)	**Appel à participation** Groupe de travail sur la complétion du standard Unicode, des polices de caractères, et des classes d'édition de texte. **Objectif** : Régler définitivement les problèmes de saisie, d'affichage et d'impression et obtenir des logiciels de base intégrant tous les systèmes d'écriture existants. **Durée** : 2 ans.
T_0	Universités et instituts de langues, principaux éditeurs de logiciel, SIL International (groupe 2)	**Appel à participation** Groupe de travail sur le recensement de la situation (intérêt des populations, état d'informatisation, langues proches bien informatisées...) et des besoins en compléments et outils (travail par groupes de langues, en particulier pour les compléments linguistiques génériques). **Objectif** : Rédaction de spécifications techniques utilisables par des informaticiens pour les compléments et les outils. Publication du planning général incluant les étapes avec la liste des langues retenues pour chacune d'elles. **Durée** : 2 ans.
T_0+2	Populations linguistiques	**Appel à participation** Groupes de travail pour les langues insuffisamment décrites. **Objectif** : Création de groupes de linguistes et réalisation de dictionnaires et de grammaires pour les étapes suivantes. **Durée** : 4 ans.
T_0+2	Éditeurs de logiciel de base	**Diffusion de spécifications** Diffusion des spécifications des compléments généraux. **Objectif** : Intégration de l'interface $I_{G/L}$ aux logiciels de base. **Durée** : 1 an.
T_0+2	Universités et instituts de langues (inclut des projets multilingues)	**Diffusion de spécifications et appel d'offres** Diffusion des spécifications des compléments linguistiques génériques et des outils pour linguistes. Appel d'offres pour leur réalisation. Invitation des universités et instituts de langues à répondre en consortiums en fonction de leurs compétences. **Objectif** : Développement des compléments linguistiques génériques et des outils pour linguistes. **Durée** : 1 an.
T_0+3	Populations linguistiques	**Lancement des projets par langue** Diffusion des outils pour linguistes et lancement des projets de réalisation des compléments linguistiques spécifiques et des ressources linguistiques. Accompagnement des projets de compléments linguistiques spécifiques (participation des différents acteurs intéressés : projets multilingues...). **Objectif** : Développement des compléments linguistiques spécifiques. **Durée** : 3 ans.
T_0+6	Tous les acteurs	**Annonce de la fin de la phase 1** Compte-rendu des travaux.

Figure 61 : Tableau des tâches pour la première phase de la première étape du projet

CONCLUSION DE LA TROISIÈME PARTIE : DES MÉTHODES POUR UN GROUPE DE LANGUES

Les idées et méthodes développées dans la première partie rendent plus efficace l'informatisation d'une langue. Lorsque l'objectif est d'informatiser, non pas une, mais plusieurs langues en même temps, ces préceptes restent valables mais d'autres peuvent venir compléter leur arsenal de techniques d'optimisation des moyens à mettre en œuvre.

Nous partons de l'hypothèse que les logiciels grand public forment une base sur laquelle nous greffons des compléments linguistiques. Dans chaque catégorie — traitements de texte, systèmes de traduction automatique, systèmes de reconnaissance vocale... — plusieurs logiciels grand public peuvent être utilisés en même temps, qu'ils soient spécifiques à un système d'exploitation ou non. Notre expérience sur la langue laotienne a montré que les compléments linguistiques peuvent être largement réutilisés d'un logiciel de base à l'autre (à l'intérieur d'une catégorie donnée) : 98 % de gain lors du passage de nos compléments linguistiques de WordPad à Word.

De manière comparable, les adaptations réalisées dans les logiciels de base pour accueillir des compléments linguistiques se sont révélées très flexibles lorsque nous avons voulu les interfacer avec des compléments de la même catégorie mais pour d'autres langues : là encore, près de 98 % de gain lors de l'adaptation au bengali de LaoUniKey, logiciel développé initialement pour permettre la saisie du laotien en Unicode.

Des extrapolations nous ont permis d'évaluer l'économie d'échelle réalisée lorsque l'on informatise simultanément un grand nombre de langues en s'appuyant sur plusieurs logiciels de base à la fois. Par exemple, l'intégration de dix langues dans quatre logiciels de base différents conduit à une économie d'échelle de 80 % par rapport à ce que coûteraient quarante développements indépendants, dans le cas où sont définies des interfaces entre les modules liés à la langue et les autres. Cela nous a conduit à proposer une architecture modulaire comportant :

- ⇨ des modules de base offrant une interface d'accueil standard pour des compléments linguistiques,
- ⇨ des modules linguistiques conformes à ces interfaces.

Pour pouvoir apporter une solution à la diversité linguistique, nous proposons le développement d'outils simples destinés à des linguistes et générant le code des compléments linguistiques. Nous avons expérimenté cette voie. Après avoir défini l'interface d'un complément linguistique générique assurant la sélection de texte pour des écritures non segmentées, nous avons réalisé Sylla, un outil permettant de mettre au point des modèles syllabiques pour n'importe quel système d'écriture non segmenté, et nous l'avons utilisé pour mettre au point les modèles syllabiques du birman, du khmer, du laotien et du siamois (thaï) avec l'aide de Pierre Sein Aye, Michel Antelme et Gilles Delouche. La génération du code C++ correspondant au modèle pourra prochainement être produit automatiquement par cet outil.

Nous proposons un schéma de développement dans lequel les différents acteurs ont des rôles complémentaires : les éditeurs des logiciels se chargent de développer et de maintenir les modules de base, et les personnes le souhaitant développent les compléments pour leur langue. L'architecture modulaire et la création d'outils de génération de compléments linguistiques permettent ainsi à la fois d'abaisser très sensiblement le coût de l'informatisation d'un groupe de langues mais surtout de donner aux populations linguistiques la possibilité, si elles le souhaitent, de réaliser de façon autonome des logiciels et ressources adaptés à leur langue. Pour animer un tel projet, un soutien est nécessaire. Dans l'exemple de plan de développement sur six ans que nous avons présenté, nous avons fait jouer ce rôle aux Nations Unies. Ces derniers y supervisent l'informatisation d'une centaine de langues, mettant en relation les éditeurs de logiciel qui amènent les logiciels de base, un réseau d'universités et d'organismes qui réalise les logiciels, et les populations linguistiques qui développent les compléments linguistiques pour leur langue. L'ensemble des informations et des ressources nécessaires est supposé être mis gratuitement à la disposition des volontaires sur un site Internet.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Informatiser une ou plusieurs langues peu dotées, délaissées par les éditeurs de logiciel à cause de leur non-rentabilité, se fait souvent avec les seuls moyens de quelques programmeurs isolés parlant ces langues. Il est donc important, pour ces personnes, d'utiliser au mieux les leviers permettant d'obtenir rapidement des logiciels de qualité.

Le présent travail de thèse a montré comment quelques méthodes simples permettent à des développeurs d'améliorer significativement l'indice–σ[1] d'une langue faiblement dotée. Parmi elles, nous avons mis en œuvre sur la langue laotienne : le recours à **des environnements** et à **des projets génériques**, le principe de **contribution linguistique généralisé** et **l'utilisation d'Internet et du courrier électronique**. Ainsi, l'indice-σ du laotien, qui était d'environ 4/20 au début de nos travaux, est actuellement estimé à 8,68/20, notre contribution étant d'environ 4 points.

Les services développés sont des traitements de texte, un dictionnaire électronique et une aide à la traduction, le temps passé ayant été de 2,5 hommes-ans pour les traitements de textes et 0,5 hommes-ans pour l'aide à la traduction et le dictionnaire. Si nous comparons, par exemple, les deux hommes-ans passés à réaliser le logiciel LaoWord[2] et ce que coûte le développement ex-nihilo d'un traitement de textes de qualité équivalente, on peut estimer que le rapport est très inférieur à 1 %. Cela a été rendu possible grâce à l'utilisation d'un environnement générique comme logiciel de base : le traitement de texte Word de Microsoft.

Pour l'informatisation d'un groupe de langues, les développeurs peuvent tirer parti, en plus des méthodes précédentes, de la spécificité multi-langues du problème. Il peut s'agir des relations existant entre les langues, comme la parenté linguistique, la proximité géographique ou l'histoire commune. Dans ce cas, les problèmes des langues sont, sinon identiques, du moins liés entre eux : systèmes d'écriture similaires, existence de dictionnaires et de personnes bilingues, morphologie, syntaxe ou terminologie comparables. L'exploitation de ces similarités permet de réduire l'effort nécessaire.

Une autre économie peut provenir des factorisations rendues possibles du fait que plusieurs langues sont traitées dans le même projet : plates-formes communes à toutes les langues considérées, traitements linguistiques réutilisés dans plusieurs applications. Nous montrons que l'économie d'échelle obtenue peut alors s'avérer très significative. Par exemple, nous évaluons à 80 % l'économie réalisée en développant, dans le même projet, quatre applications adaptées à dix langues différentes, par rapport à ce que coûteraient quarante développements indépendants.

Ces gains supposent l'utilisation d'une architecture modulaire dont les interfaces internes sont définies. Deux interfaces sont importantes :

- $I_{G/L}$, entre les logiciels de base et les compléments linguistiques,
- $I_{LG/LS}$, à l'intérieur de ces derniers, entre :
  - les parties génériques pouvant s'appliquer à l'ensemble ou à un sous-ensemble des langues (par exemple : gestion d'un dictionnaire, algorithme de sélection syllabique),
  - les parties propres à une langue (par exemple : fonction de reconnaissance des syllabes pour une langue donnée).

[1] Mesure de la satisfaction des utilisateurs vis à vis des logiciels adaptés à leur langue.
[2] Logiciel ajoutant à Microsoft Word des fonctions spécifiquement laotiennes.

Ainsi, les éditeurs de logiciel peuvent intégrer l'interface $I_{G/L}$ dans la conception de leurs produits, des équipes spécialisées (laboratoires universitaires…) peuvent développer des « middleware » présentant les interfaces $I_{G/L}$ et $I_{LG/LS}$, enfin les populations linguistiques peuvent développer ce qui est propre à leurs langues. Pour faciliter cela, nous proposons que les équipes réalisant le « middleware » réalisent aussi des outils linguiciels faciles d'emploi et permettant à des non-informaticiens de produire des compléments linguistiques spécifiques conformes à l'interface $I_{LG/LS}$. Par cette façon de procéder, la plus grosse partie de l'effort d'informatisation des langues-π[1] est reportée sur les populations linguistiques qui peuvent réaliser les compléments linguistiques spécifiques de leurs langues sans nécessiter de formation lourde, ce qui peut amener une réduction des coûts considérable.

APPORTS DE LA THÈSE

Le premier apport de cette thèse réside dans la synthèse et la structuration de nombreuses informations, donnant des points de repère dans le domaine de l'informatisation des langues-π : langues du monde, systèmes d'écriture, droit des langues, actions et programmes pour la sauvegarde des langues, projets d'informatisation, technologie, éléments de coût.

Elle donne ensuite une liste de méthodes permettant d'employer l'effort plus efficacement et montre cette efficacité sur quelques langues test : principalement le laotien, mais aussi le birman, le khmer, le siamois (thaï) et le bengali.

Elle précise ce qu'est une langue faiblement dotée et, plus généralement, donne un moyen de mesurer le degré d'informatisation d'une langue.

Elle a été l'occasion d'améliorer sensiblement le degré d'informatisation du laotien et de fournir des outils linguiciels (Sylla, mais aussi LaoUniKey et LaoUniWeb) permettant, potentiellement, d'améliorer celui de plusieurs autres langues.

Elle propose une architecture et une méthodologie permettant d'informatiser efficacement et durablement de nombreuses langues, répartissant la charge entre les éditeurs de logiciel, les populations linguistiques et des organismes mettant en place les moyens nécessaires pour que ces derniers puissent adapter les logiciels des premiers à leurs langues.

POTENTIEL DE RECHERCHE ET DÉVELOPPEMENT

Les perspectives immédiates de recherche et développement consistent à identifier et à classer les besoins spécifiques des différentes langues du monde non couvertes par les logiciels existants. Cela permettrait de spécifier les interfaces $I_{G/L}$ et $I_{LG/LS}$ ainsi que les différents compléments linguistiques et outils linguiciels. Dès lors, une phase de réalisation pourrait être lancée, par exemple sous la forme d'un projet de type GNU en réseau et accompagné par l'UNESCO.

[1] Langues faiblement dotées.

ANNEXES

A.1 INDICATIONS SUR L'ÉTAT DE L'ART TECHNOLOGIQUE

A.1.1 PRÉSENTATION

Cette annexe propose quelques indications classées — essentiellement des liens vers des sites Internet et quelques éléments de bibliographie — donnant une première orientation dans les techniques du traitement des langues. Elle complète les informations générales fournies au chapitre I.1.3.3.
Les *informations bibliographiques* sont en **gras**.

A.1.2 GÉNÉRALITÉS

A.1.2.1 Introduction

Ingénierie des langues (ouvrage collectif dirigé par Jean-Marie Pierrel), Hermes, 2000
L'intelligence artificielle et le langage, vol. 1 (Gérard Sabah), Hermes, 1988
L'intelligence artificielle et le langage, vol. 2 (Gérard Sabah), Hermes, 1990
Traitement automatique des langues naturelles (collectif, dir. Pierrette Bouillon), Duculot, 1998
L'analyse syntaxique des langues naturelles (Éric Wehrli), Masson, 1997
Sémantique pour l'analyse (F. Rastier, M. Cavazza, A. Abeillé), Masson, 1994
Handbook of natural language processing (Robert Dale, Hermann Moisl, Harold Somers), Marcel Dekker, New York, 2000
Speech and language processing (Daniel Jurafsky, James Martin), Prentice Hall, 2000
Initiation au TAL et outils (Reptil) : http://www.up.univ-mrs.fr/veronis/Atala/reptil/
Survey of the State of the Art in Human Language Technology : http://www.cslu.ogi.edu/HLTsurvey/
Bibliographie (LIMSI) : http://www.limsi.fr/Individu/gs/BibliographieEndNote/f-bib-sab.html

A.1.2.2 Linguistique

Observatoire de la Linguistique Sens-Texte (OLST) : http://www.olst.umontreal.ca/
Phonétique (université de Lausanne): http://www2.unil.ch/ling/phon/index.html
Phonétique (université d'Alberta) : http://www.fsj.ualberta.ca/beaudoin/ling/phone.htm
Phonétique (université de Kingston - Queen's) : http://qsilver.queensu.ca/french/Cours/215/chap2.html
Alphabet Phonétique International : http://www2.arts.gla.ac.uk/IPA/ipa.html
Marges linguistiques (revue) : http://www.marges-linguistiques.com/
Langues de France : http://www.culture.fr/Groups/langues/home

A.1.2.3 Ressources de développement

PCCTS (création de compilateur[1]) : http://dynamo.ecn.purdue.edu/~hankd/PCCTS/
Ressources pour PCCTS (création de compilateur) : http://www.ocnus.com/pccts.html
Ressources pour PCCTS (création de compilateur) : http://www.polhode.com/pccts.html
ANTLR (création de compilateur) : http://www.antlr.org/
Ressources pour développeurs (IBM) : http://www-136.ibm.com/developerworks/
Jade (outil DSSSL de James Clark) : http://www.jclark.com/jade/
Outils SIL : http://www.sil.org/computing/catalog/
Librairie XML (univ. d'Edinburgh) : http://www.ltg.ed.ac.uk/corpora/xmldoc/release/xmlweb.html
Site Oasis (SGML, XML...) : http://www.oasis-open.org/cover/
Grammaires LFG (université d'Essex) : http://www.essex.ac.uk/linguistics/LFG/
Liste d'analyseurs syntaxiques : http://www.lpl.univ-aix.fr/~blache/analyse.html
MSDN (support développement Microsoft) : http://msdn.microsoft.com/library/
Ressources téléchargeables : http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat
Ressources (université de New-Mexico, CLR) : http://crl.nmsu.edu/crltoolkit/crl2.htm
Alep : http://www.franken.de/users/nicklas/das/projects/alep/www_demo/LingwareEditor.html

[1] Voir aussi Lex/Yacc, Flex/Bison et Javacc.

A.1.3 TRAITEMENT DU TEXTE (PREMIER NIVEAU)

A.1.3.1 Introduction

Unicode 3.0 (The Unicode Consortium), Addison-Wesley, 2000
Site Unicode : http://www.unicode.org/
Systèmes d'écriture Unicode : http://www.unicode.org/charts/

A.1.3.2 Polices de caractères

A.1.3.2.1 POLICES DU MONDE ENTIER

Yamada Language Center : http://babel.uoregon.edu/yamada/fonts.html
SIL International : http://www.sil.org/computing/fonts/
Linguist's Software : http://www.linguistsoftware.com/
Omniglot : http://omniglot.com/links/fonts.htm
Polices africaines : http://www.progiciels-bpi.ca/tcao/apercu.html#h-an4
Polices de caractères vietnamiens : http://www.ambafrance-vn.org/download/vnfont/infofont.htm
Unifont (GNU) : http://czyborra.com/unifont/

A.1.3.2.2 TECHNOLOGIES TRUETYPE ET OPENTYPE

TrueType & OpenType (MS): http://www.microsoft.com/typography/users.htm
Article sur OpenType (MS): http://www.microsoft.com/typography/developers/opentype/default.htm
Manuel TrueType (Apple) : http://developer.apple.com/fonts/TTRefMan/index.html

A.1.3.2.3 AUTRES TECHNOLOGIES ADAPTÉES AUX POLICES D'ÉCRITURES COMPLEXES

Graphite (SIL) : http://www.sil.org/computing/graphite/
FreeType : http://www.freetype.org/
AAT (Apple) : http://developer.apple.com/fonts/
ICU (IBM) : http://oss.software.ibm.com/icu/
ClearType (Microsoft) : http://www.microsoft.com/typography/cleartype/
WEFT (Microsoft) : http://www.microsoft.com/typography/web/embedding/weft3/default.htm

A.1.3.2.4 OUTILS DE DÉVELOPPEMENT

Fontlab : http://www.fontlab.com/html/fontlab.html
Fontlab et autres (Pyrus) : http://www.pyrus.com/
Fontographer (Macromedia) : http://www.macromedia.com/software/fontographer/
VOLT (Microsoft) : http://www.microsoft.com/typography/developers/volt/default.htm
PfaEdit (SourceForge) : http://pfaedit.sourceforge.net/
Panther : http://www.codepoetry.net/archives/2003/10/24/panthers_major_text_services_upgrade.php
Longhorn : http://longhorn.msdn.microsoft.com/lhsdk/layout/overviews/typography_ovw.aspx

A.1.3.2.5 INFORMATIONS DIVERSES

Liste de fonderies (MyFont) : http://www.myfonts.com/FontFoundry?level=-1
Typographie : http://typo.textbox.org/article/172
Typographie (Microsoft) : http://www.microsoft.com/typography/default.mspx
Création de polices en ligne : http://www.fontifier.com/
Site personnel : http://perso.wanadoo.fr/lefonds/old/html/typofr.html
Site personnel : http://www.eki.ee/letter/
Systèmes d'écriture (Omniglot) : http://omniglot.com/writing/alphabetic.htm

A.1.3.3 Claviers

Clavier virtuel Microsoft : http://microsoft.com/downloads/details.aspx?FamilyId=FB7B3DCD-D4C1-4943-9C74-D8DF57EF19D7&displaylang=en
Clavier virtuel pour langues africaines : http://www.progiciels-bpi.ca/tcao/clavier.html
IME : http://www.microsoft.com/windows/ie/downloads/recommended/ime/default.asp
Programmes de traitement du thaï : http://www.links.nectec.or.th/www-new/download.php

A.1.3.4 Tri lexicographique

ISO/CEI CD 14651 : http://anubis.dkuug.dk/jtc1/sc22/open/n2933.pdf (Classement international de chaînes de caractères - Méthode de comparaison de chaînes de caractères et description du modèle commun d'ordre de classement)
Tri du français : http://www.tresor.gouv.qc.ca/doc/techtri.htm
Tri du khmer : http://www.bauhahnm.clara.net/Khmer/Welcome.html#KHMERSORTING
Tri du thaï : http://www.linux.thai.net/thep/tsort.html
Tri de langues africaines (Bourbeau Pinard) : http://www.progiciels-bpi.ca/tcao/apercu.html#h-2.2.4

A.1.3.5 Ressources diverses

OpenOffice.org : http://www.openoffice.org/
Office : http://msdn.microsoft.com/office/
Ressources pour plusieurs langues : http://omniglot.com/links/software.htm
Liste de ressources multilingues : http://www.gy.com/source/lang.html
Traitement de texte multilingue (UniWrite) : http://www.softissimo.com/products/uniwrite.htm
Traitement de texte multilingue (GlobalOffice et GlobalWriter) : http://www.unitype.com/
Traitement de texte multilingue (Xenotype) : http://www.xenotypetech.com/
Traitement de texte multilingue (Universal Word et OnePen) : http://www.aramedia.com/
Site gy.com : http://www.gy.com/home.html
Site WorldLanguage.com : http://www.worldlanguage.com/

A.1.4 TRAITEMENT DU TEXTE (DEUXIÈME NIVEAU)

A.1.4.1 Introduction

Computer Programs for Spelling Correction, Lecture Notes in Computer Science, Vol 96 (James Lyle Peterson), Springer-Verlag , New York, 1980.
Bibliographie (correction d'orthographe) : http://www.math.utah.edu/pub/tex/bib/spell.html

A.1.4.2 Correction de texte

Ispell (correcteur d'orthographe) : http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html
Ispell (en français) : http://www.mime.univ-paris8.fr/~loscar/Papyrus/doc/node143.html
MySpell (OpenOffice.org, nombreuses langues) : http://lingucomponent.openoffice.org
Article (correcteurs d'orthographe) : http://web.dcs.bbk.ac.uk/~roger/spellchecking.html
Article (correcteur grammatical du français) : http://www.lim.nl/monitor/machina-sapiens-1.html
Synapse Cordial 10 (Correcteur du français) : http://www.synapse-fr.com/

A.1.4.3 Analyseurs morphologiques et syntaxiques

PILAF: http://clips.imag.fr/trilan/Pilaf/
Chunks : http://users.info.unicaen.fr/~jvergne/analyseur_GREYC/analyseur_du_GREYC.html
Segmentation du thaï (page perso) : http://www.geocities.com/ResearchTriangle/Thinktank/5593/
Analyseurs SIL : ftp://ftp.sil.org/software/win/pc-parse-win.zip (inclut le code source)
Inclut :

- PC-KIMMO : http://www.sil.org/computing/catalog/show_software.asp?id=33
- Ample : http://www.sil.org/computing/catalog/show_software.asp?id=1

- PC-Parse : http://www.sil.org/computing/catalog/show_software.asp?id=34
- Ktagger : http://www.sil.org/computing/catalog/show_software.asp?id=22
- Ktext : http://www.sil.org/computing/catalog/show_software.asp?id=23

A.1.4.4 Outils divers

Liste d'outils ATALA : http://www.biomath.jussieu.fr/ATALA/outil/
Liste d'outils ELSNET : http://www.elsnet.org/toolslist.html
Outils ISSCO : http://issco-www.unige.ch/tools/

A.1.5 SYNTHÈSE VOCALE

A.1.5.1 Introduction

Traitement de la parole (René Boite, Hervé Bourlard, Thierry Dutoit, Joël Hancq, Henri Leich), Presses polytechniques et universitaires romandes, 2000, ISBN 2-88074-388-5

23es journées d'étude sur la parole : http://www.icp.inpg.fr/jep2000/

A.1.5.2 Systèmes universitaires

Festvox (code et documentation) : http://festvox.org/festvox/festvox_toc.html
Démos Festvox : http://festvox.org/voicedemos.html
Festival (Edinburgh) : http://www.cstr.ed.ac.uk/projects/festival/
MBROLA (Mons) : http://tcts.fpms.ac.be/synthesis/mbrola/
Euler (Mons) : http://tcts.fpms.ac.be/synthesis/euler/
Strut (Mons) : http://www.tcts.fpms.ac.be/asr/project/strut/
PSOLA : http://www.ircam.fr/equipes/analyse-synthese/peeters/PSOLA/index.html

A.1.5.3 Systèmes commerciaux

Elan Speech : http://www.elan.fr/
Synthèse iSpeak et DECtalk : http://www.fonix.com/
ScanSoft (ex-Dragon) : http://www.scansoft.com
MySoft : http://www.mysoft.fr/speakbac.htm

A.1.5.4 Divers

VoiceXML : http://www.w3.org/TR/voicexml/
Page de Marie-Josée Leboeuf : http://www.esi.umontreal.ca/~leboeufm/blt6134/theorie.html
Page de Michel Divay : http://www.iut-lannion.fr/MD/MDRECH/
Interface Internet : http://www.trace.wisc.edu/world/web/index.html
ISCA (synthèse) : http://www.slt.atr.co.jp/cocosda/synthesis/synsig.html
ISCA (prosodie) : http://aune.lpl.univ-aix.fr:16080/projects/sprosig/

A.1.6 RECONNAISSANCE DE LA PAROLE

A.1.6.1 Systèmes universitaires

Outils de l'université d'Illinois : http://www.ifp.uiuc.edu/speech/software/
Sphinx (CMU) : http://fife.speech.cs.cmu.edu/sphinx/

A.1.6.2 Systèmes commerciaux

SpeechMagic (Philips) : http://www.speech.philips.com/
Naturally Speaking (ScanSoft, ex-Dragon) : http://www.scansoft.com

ViaVoice (IBM) : http://www.ibm.com/software/speech/
http://www-306.ibm.com/software/voice/viavoice/

A.1.6.3 Divers

Sphinx-4, FreeTTS, Java Speech API et SpeechActs (Sun) : http://research.sun.com/speech/
Traitement de la parole (SRI) : http://www.speech.sri.com/
Traitement de la parole (MIT) : http://www.sls.csail.mit.edu/sls/sls-orange-noflash.shtml
VoiceXML : http://www.w3.org/TR/voicexml/
VoiceXML : http://www.wirelessdevnet.com/training/voicexml/voicexmloverview.html
Browser VoiceXML : http://www.speech.cs.cmu.edu/openvxi/index.html
Microphones : http://www.microphones.com/

A.1.7 TRADUCTION AUTOMATIQUE ET AIDE À LA TRADUCTION

A.1.7.1 Introduction

Traduction assistée par ordinateur (Christian Boitet), dans « Ingénierie des langues » (2000).
Ariane (Christian Boitet), 1990 :
http://www-clips.imag.fr/geta/christian.boitet/pages_personnelles/zArticles_sur_la_TAO_pdf/TAO-90.II.PNTAO.pdf
TA fondée sur le dialogue (Christian Boitet), 1993 :
http://clips.imag.fr/geta/christian.boitet/pages_personnelles/zArticles_sur_la_TAO_pdf/LivreTA-93.pdf
Aide à la traduction (Christian Boitet), 1996 :
http://www-clips.imag.fr/geta/christian.boitet/pages_personnelles/zArticles_sur_la_TAO_pdf/Moncton-96.CB.v2.pdf
Traduction automatique (documentation) : http://www.transref.org/
Guide en traduction automatique : http://www.essex.ac.uk/linguistics/clmt/MTbook/

A.1.7.2 Systèmes universitaires

Documentation Ariane G5 (GETA) : http://clips.imag.fr/projets/arianey/AG5/index.html
Documentation Ariane Y (GETA) : http://clips.imag.fr/projets/arianey/
Présentation CStar (traduction de la parole) :
http://www-clips.imag.fr/projets/cstar/clips/IntroClips.html
Universal Networking Language (UNL) : http://www.unl.ias.unu.edu/
Projet BabelCode : http://www.babelcode.org
Pangloss (CMU, NMSU et USC) : http://www.lti.cs.cmu.edu/Research/Pangloss/Home.html
Pangloss (CMU, NMSU et USC) : http://citeseer.nj.nec.com/241393.html
Mikrokosmos (NMS) : http://citeseer.nj.nec.com/beale95semantic.html
MaTra (NCST, Inde) : http://www.ncst.ernet.in./kbcs/nlp.shtml

A.1.7.3 Systèmes commerciaux

Réverso (Softissimo) : http://www.reverso.com/
Systran : http://www.systransoft.com/

A.1.7.4 Liste de systèmes de traduction

Systèmes de traduction automatique : http://www.lpl.univ-aix.fr/~blache/prost.html
Liste de systèmes de traduction : http://www.megagiciel.com/logiciels/275.html
Liste de systèmes de traduction : http://www.foreignword.com/
Liste de systèmes de traduction : http://www.i18nguy.com/TranslationTools.html
Liste de systèmes de traduction : http://omniglot.com/links/translation.htm

A.1.7.5 Traduction en ligne

Traduction en ligne (Voilà) : http://trans.voila.fr/
Traduction en ligne (Babel Fish) : http://babel.altavista.com/
Traduction en ligne (Systran) : http://www.systransoft.com/
Traduction en ligne (Foreignword) : http://www.foreignword.com/
Traduction en ligne (Gist-in-time) : http://www.teletranslator.com/
Traduction en ligne (Google) : http://www.google.ca/language_tools?hl=fr
Traduction en ligne (Promt) : http://www.translate.ru/
Traduction en ligne (T-mail) : http://www.t-mail.com/
Traduction en ligne (Prasit) : http://www.suparsit.com/index2new.html
Traduction en ligne (Poltran) : http://www.poltran.com/

A.1.7.6 Aide à la traduction

Aides à la traduction (journée d'étude du 8 octobre 1998) :
http://www.adbs.fr/site/evenements/journees/journee.php?limit=0&annee=1998&id=30&version=1
Aide à la traduction (Aramedia) : http://www.aramedia.com/

A.1.8 RESSOURCES LEXICALES ET CORPUS

A.1.8.1 Introduction

Création et gestion de ressources linguistiques (plusieurs chapitres par Jean Véronis, Patrick Paroubek, Martin Rajman, Patrice Bonhomme, Laurent Romary), dans « Ingénierie des langues », ouvrage collectif dirigé par Jean-Marie Pierrel, 2000.
Méthodes d'acquisition lexicale en TAO (Christian Boitet), 2001 :
http://www.li.univ-tours.fr/taln-recital-2001/Actes/tome2_PDF/partie3_confAss/art1_p249_265.pdf
Bibliographie : http://www.termisti.refer.org/theoweb9.htm
Outil Shoebox : http://www.ethnologue.com/tools_docs/shoebox.asp

A.1.8.2 Lexicographie

Association Asiatique pour la Lexicographie (Asialex) : http://www.asialex.org/
Association Européenne pour la Lexicographie (Euralex) : http://www.ims.uni-stuttgart.de/euralex/
Association Africaine pour la Lexicographie (Afrilex) :
http://www.up.ac.za/academic/libarts/afrilang/homelex.html
Association Australasienne pour la Lexicographie (Australex) : http://australex.anu.edu.au/

A.1.8.3 Dictionnaires

Grand Dictionnaire Terminologique : http://www.granddictionnaire.com
Trésor de la Langue Française informatisé : http://www.inalf.fr/_ie/index_tlfi.htm
Base de données lexicales multilingue Papillon : http://www.papillon-dictionary.org
Base de données lexicales anglaises en ligne (Wordnet) : http://www.cogsci.princeton.edu/~wn/
Eurodicautom : http://europa.eu.int/eurodicautom/Controller
Dicologique et autres dictionnaires (Memodata) : http://www.memodata.com/index.shtml
Dictionnaire thaï-japonais et japonais-thaï en ligne (Saikam) : http://saikam.nii.ac.jp/
Dictionnaire français-malais : http://www-clips.imag.fr/geta/services/fem/
Dictionnaire français-malais : http://www.lirmm.fr/~lafourca/ML-projects/ALEX/wAlex-FeM-v3.html
Dictionnaire français-japonais en ligne : http://dico.fj.free.fr/index.php
Dictionnaires géorgien, russe, anglais, allemand : http://greg.iatp.org.ge/

A.1.8.4 Listes de dictionnaires

Listes de dictionnaires en ligne (Yakeo) : http://www.yakeo.com/fr/dictionnaires/
Liste de dictionnaires en ligne (université de Nancy) : http://www.sciences.bu.u-nancy.fr/net.htm#DICTIONNAIRES
Liste de dictionnaires en ligne (YourDictionary) : http://www.yourdictionary.com/
Liste de dictionnaires : http://www.megagiciel.com/logiciels/275.html
Liste de dictionnaires : http://www.foreignword.com/
Liste Omniglot : http://omniglot.com/links/dictionaries.htm
Travlang : http://www.travlang.com/languages/

A.1.8.5 Corpus

Text Encoding Initiative (TEI) : http://www.tei-c.org/
Corpus Encoding Standard (CES, EAGLES) : http://www.cs.vassar.edu/CES/
MULTEXT : http://www.lpl.univ-aix.fr/projects/multext/
Corpus thaï (NECTEC, Thaïlande) : http://www.links.nectec.or.th/orchid/
Silfide : http://www.loria.fr/projets/Silfide/Index.html
Silfide : http://www.loria.fr/projets/XSilfide/FR/index.html

A.1.9 DIVERS

A.1.9.1 Reconnaissance optique de caractères et standardisation

ROC multilingue (Aramedia) : http://www.aramedia.com/
Standardisation orthographique (Marilyn Mason) : http://hometown.aol.com/créoleCH/

A.1.9.2 Autres

Internationalisation d'Internet (Babel) : http://alis.isoc.org/
Analyse et Traitement informatique de la Langue Française (ATILF) : http://www.atilf.fr
Fouille de données textuelles (Lexicometrica) : http://www.cavi.univ-paris3.fr/lexicometrica/
Résumé automatique (Gragomir Radev) : http://www.summarization.com/
Thai Open Source Developer Network : http://developer.thai.net/
Projet coopératif thaï (ZzzTh@i) : http://www.fedu.uec.ac.jp/ZzzThai/
Projet coopératif japonais (Yakushite.net) : http://www.yakushite.net/
ELRA (ressources diverses) : http://www.elra.info/
Xerox (technologie des langues) : http://www.xrce.xerox.com/
Talana : http://talana.linguist.jussieu.fr/
Lingaware: http://www.linga.fr/LingFr/Homefr.htm
Linguacubun: http://www.chibcha.demon.co.uk/
Linguistic Data Consortium: http://www.ldc.upenn.edu/
Localisation Research Centre : http://www.localisation.ie/

A.2 TABLEAU ISO 639 DES CODES DE LANGUES

Le tableau suivant reproduit la liste des codes utilisés par les normes internationales ISO 639 : « Codes pour la représentation des noms de langue »[1] publiées :

⇨ en 2002 pour la version sur deux lettres « Partie 1 : Code alpha-2 » (ISO 639-1, 185 codes),

⇨ en 1998 pour la version sur trois lettres « Partie 2 : Code alpha-3 » (ISO 639-2, 465 codes).

Un tableau de correspondance entre les codes ISO, Microsoft et Macintosh est proposé sur le site d'Unicode : http://www.unicode.org/unicode/onlinedat/languages.html.

Langue	639-1	639-2	Numéro
abkhaze	ab	abk	1
aceh		ace	2
acoli		ach	3
adangme		ada	4
adygh		ady	5
afar	aa	aar	6
afrihili		afh	7
afrikaans	af	afr	8
afro-asiatiques, autres langues		afa	9
akan	ak	aka	10
akkadien		akk	11
albanais	sq	alb/sqi	12
aléoute		ale	13
algonquines, langues		alg	14
allemand	de	ger/deu	15
allemand, moyen haut (ca. 1050-1500)		gmh	16
allemand, vieux haut (ca. 750-1050)		goh	17
altaïques, autres langues		tut	18
amharique	am	amh	19
anglais	en	eng	20
anglais moyen (1100-1500)		enm	21
anglo-saxon (ca.450-1100)		ang	22
apache		apa	23
arabe	ar	ara	24
aragonais	an	arg	25
araméen		arc	26
arapaho		arp	27
araucan		arn	28
arawak		arw	29
arménien	hy	arm/hye	30
artificielles, autres langues		art	31
assamais	as	asm	32
asturien; bable		ast	33
athapascanes, langues		ath	34
australiennes, langues		aus	35
avar	av	ava	36
avestique	ae	ave	37

[1] Voir http://www.iso.org/iso/fr/prods-services/popstds/languagecodes.html, http://linux.infoterm.org/iso-e/i-iso.htm et http://www.loc.gov/standards/iso639-2/frenchlangn.html. Voir aussi les mises à jour disponibles sur le site http://lcweb.loc.gov/standards/iso639-2/codechanges.html.

Langue	639-1	639-2	Numéro
awadhi		awa	38
aymara	ay	aym	39
azéri	az	aze	40
bachkir	ba	bak	41
balinais		ban	42
baloutchi		bal	43
baltiques, autres langues		bat	44
bambara	bm	bam	45
bamilékés, langues		bai	46
banda		bad	47
bantoues, autres langues		bnt	48
bas allemand; bas saxon; allemand, bas; saxon, bas		nds	49
basa		bas	50
basque	eu	baq/eus	51
bas-sorabe		dsb	52
batak (Indonésie)		btk	53
bedja		bej	54
bemba		bem	55
bengali	bn	ben	56
berbères, autres langues		ber	57
bhojpuri		bho	58
bichlamar	bi	bis	59
biélorusse	be	bel	60
bihari	bh	bih	61
bikol		bik	62
bini		bin	63
birman	my	bur/mya	64
blackfoot		bla	65
bosniaque	bs	bos	66
bouriate		bua	67
braj		bra	68
breton	br	bre	69
bugi		bug	70
bulgare	bg	bul	71
caddo		cad	72
caribe		car	73
catalan; valencien	ca	cat	74
caucasiennes, autres langues		cau	75
cebuano		ceb	76
celtiques, autres langues		cel	77
chames, langues		cmc	78
chamorro	ch	cha	79
chan		shn	80
cherokee		chr	81
cheyenne		chy	82
chibcha		chb	83
chichewa; chewa; nyanja	ny	nya	84
chinois	zh	chi/zho	85
chinook, jargon		chn	86
chipewyan		chp	87
choctaw		cho	88

Langue	639-1	639-2	Numéro
chuuk		chk	89
copte		cop	90
coréen	ko	kor	91
cornique	kw	cor	92
corse	co	cos	93
couchitiques, autres langues		cus	94
cree	cr	cre	95
créoles et pidgins anglais, autres		cpe	96
créoles et pidgins divers		crp	97
créoles et pidgins français, autres		cpf	98
créoles et pidgins portugais, autres		cpp	99
croate	hr	scr/hrv	100
dakota		dak	101
danois	da	dan	102
dargwa		dar	103
dayak		day	104
delaware		del	105
dinka		din	106
dioula		dyu	107
diverses, langues		mis	108
djaghataï		chg	109
dogri		doi	110
dogrib		dgr	111
douala		dua	112
dravidiennes, autres langues		dra	113
dzongkha	dz	dzo	114
écossais		sco	115
efik		efi	116
égyptien		egy	117
ekajuk		eka	118
élamite		elx	119
erza		myv	120
esclave (athapascan)		den	121
espagnol; castillan	es	spa	122
espéranto	eo	epo	123
estonien	et	est	124
éwé	ee	ewe	125
éwondo		ewo	126
fang		fan	127
fanti		fat	128
féroïen	fo	fao	129
fidjien	fj	fij	130
finnois	fi	fin	131
finno-ougriennes, autres langues		fiu	132
fon		fon	133
français	fr	fre/fra	134
français ancien (842-ca.1400)		fro	135
français moyen (1400-1800)		frm	136
frioulan		fur	137
frison	fy	fry	138
gaélique; gaélique écossais	gd	gla	139

Langue	639-1	639-2	Numéro
galicien	gl	glg	140
galla	om	orm	141
gallois	cy	wel/cym	142
ganda	lg	lug	143
gayo		gay	144
gbaya		gba	145
géorgien	ka	geo/kat	146
germaniques, autres langues		gem	147
gond		gon	148
gorontalo		gor	149
gothique		got	150
goudjrati	gu	guj	151
grebo		grb	152
grec ancien (jusqu'à 1453)		grc	153
grec moderne (après 1453)	el	gre/ell	154
groenlandais	kl	kal	155
guarani	gn	grn	156
guèze		gez	157
gwich´in		gwi	158
haida		hai	159
haïtien; créole haïtien	ht	hat	160
haoussa	ha	hau	161
haut-sorabe		hsb	162
hawaïen		haw	163
hébreu	he	heb	164
herero	hz	her	165
hiligaynon		hil	166
himachali		him	167
hindi	hi	hin	168
hiri motu	ho	hmo	169
hittite		hit	170
hmong		hmn	171
hongrois	hu	hun	172
hupa		hup	173
iakoute		sah	174
iban		iba	175
ido	io	ido	176
igbo	ig	ibo	177
ijo		ijo	178
ilocano		ilo	179
indéterminée		und	180
indiennes d'Amérique centrale,		cai	181
indiennes d'Amérique du Nord,		nai	182
indiennes d'Amérique du Sud,		sai	183
indo-aryennes, autres langues		inc	184
indo-européennes, autres langues		ine	185
indonésien	id	ind	186
ingouche		inh	187
interlingua (langue	ia	ina	188
interlingue	ie	ile	189
inuktitut	iu	iku	190

Langue	639-1	639-2	Numéro
inupiaq	ik	ipk	191
iraniennes, autres langues		ira	192
irlandais	ga	gle	193
irlandais ancien (jusqu'à 900)		sga	194
irlandais moyen (900-1200)		mga	195
iroquoises, langues (famille)		iro	196
islandais	is	ice/isl	197
italien	it	ita	198
japonais	ja	jpn	199
javanais	jv	jav	200
judéo-arabe		jrb	201
judéo-espagnol		lad	202
judéo-persan		jpr	203
kabardien		kbd	204
kabyle		kab	205
kachin		kac	206
kachoube		csb	207
kalmouk		xal	208
kamba		kam	209
kannada	kn	kan	210
kanouri	kr	kau	211
karakalpak		kaa	212
karatcha balkar		krc	213
karen		kar	214
kashmiri	ks	kas	215
kawi		kaw	216
kazakh	kk	kaz	217
khasi		kha	218
khmer	km	khm	219
khoisan, autres langues		khi	220
khotanais		kho	221
kikuyu	ki	kik	222
kimbundu		kmb	223
kirghize	ky	kir	224
kiribati		gil	225
kom	kv	kom	226
kongo	kg	kon	227
konkani		kok	228
kosrae		kos	229
koumyk		kum	230
kpellé		kpe	231
krou		kro	232
kuanyama; kwanyama	kj	kua	233
kurde	ku	kur	234
kurukh		kru	235
kutenai		kut	236
lahnda		lah	237
lamba		lam	238
langues des signes		sgn	239
laotien	lo	lao	240
latin	la	lat	241

Langue	639-1	639-2	Numéro
letton	lv	lav	242
lezghien		lez	243
limbourgeois	li	lim	244
lingala	ln	lin	245
lituanien	lt	lit	246
lojban		jbo	247
lozi		loz	248
luba-katanga	lu	lub	249
luba-lulua		lua	250
luiseno		lui	251
lunda		lun	252
luo (Kenya et Tanzanie)		luo	253
lushai		lus	254
luxembourgeois	lb	ltz	255
macédonien	mk	mac/mkd	256
madourais		mad	257
magahi		mag	258
maithili		mai	259
makassar		mak	260
malais	ms	may/msa	261
malayalam	ml	mal	262
malayo-polynésiennes,		map	263
maldivien	dv	div	264
malgache	mg	mlg	265
maltais	mt	mlt	266
mandar		mdr	267
mandchou		mnc	268
mandingue		man	269
manipuri		mni	270
manobo, langues		mno	271
manx; mannois	gv	glv	272
maori	mi	mao/mri	273
marathe	mr	mar	274
mari		chm	275
marshall	mh	mah	276
marvari		mwr	277
massaï		mas	278
maya, langues		myn	279
mendé		men	280
micmac		mic	281
minangkabau		min	282
mohawk		moh	283
moksa		mdf	284
moldave	mo	mol	285
mongo		lol	286
mongol	mn	mon	287
môn-khmer, autres langues		mkh	288
moré		mos	289
mounda, langues		mun	290
multilingue		mul	291
muskogee		mus	292

Langue	639-1	639-2	Numéro
nahuatl		nah	293
napolitain		nap	294
nauruan	na	nau	295
navaho	nv	nav	296
ndébélé du Nord	nd	nde	297
ndébélé du Sud	nr	nbl	298
ndonga	ng	ndo	299
néerlandais moyen (ca. 1050-1350)		dum	300
néerlandais; flamand	nl	dut/nld	301
népalais	ne	nep	302
newari		new	303
nias		nia	304
nigéro-congolaises, autres langues		nic	305
nilo-sahariennes, autres langues		ssa	306
niué		niu	307
nogaï; nogay		nog	308
norrois, vieux		non	309
norvégien	no	nor	310
norvégien bokmål; bokmål, norvégien	nb	nob	311
norvégien nynorsk; nynorsk, norvégien	nn	nno	312
nubiennes, langues		nub	313
nyamwezi		nym	314
nyankolé		nyn	315
nyoro		nyo	316
nzema		nzi	317
occitan (après 1500); provençal	oc	oci	318
ojibwa	oj	oji	319
oriya	or	ori	320
osage		osa	321
ossète	os	oss	322
otomangue, langues		oto	323
oudmourte		udm	324
ougaritique		uga	325
ouïgour	ug	uig	326
ourdou	ur	urd	327
ouszbek	uz	uzb	328
pachto	ps	pus	329
pahlavi		pal	330
palau		pau	331
pali	pi	pli	332
pampangan		pam	333
pangasinan		pag	334
papiamento		pap	335
papoues, autres langues		paa	336
penjabi	pa	pan	337
persan	fa	per/fas	338
perse, vieux (ca. 600-400 av. J.-C.)		peo	339
peul	ff	ful	340
phénicien		phn	341
philippines, autres langues		phi	342
pohnpei		pon	343

Langue	639-1	639-2	Numéro
polonais	pl	pol	344
portugais	pt	por	345
prâkrit		pra	346
provençal ancien (jusqu'à 1500)		pro	347
quetchua	qu	que	348
rajasthani		raj	349
rapanui		rap	350
rarotonga		rar	351
réservée à l'usage local		qaa-qtz	352
rhéto-roman	rm	roh	353
romanes, autres langues		roa	354
roumain	ro	rum/ron	355
rundi	rn	run	356
russe	ru	rus	357
rwanda	rw	kin	358
salish, langues		sal	359
samaritain		sam	360
sami de Lule		smj	361
sami d'Inari		smn	362
sami du Nord	se	sme	363
sami du Sud		sma	364
sami skolt		sms	365
sami, autres langues		smi	366
samoan	sm	smo	367
sandawe		sad	368
sango	sg	sag	369
sanskrit	sa	san	370
santal		sat	371
sarde	sc	srd	372
sasak		sas	373
selkoupe		sel	374
sémitiques, autres langues		sem	375
serbe	sr	scc/srp	376
sérère		srr	377
shona	sn	sna	378
sidamo		sid	379
sindhi	sd	snd	380
singhalais	si	sin	381
sino-tibétaines, autres langues		sit	382
sioux, langues		sio	383
slaves, autres langues		sla	384
slavon d'église; vieux slave; slavon liturgique; vieux bulgare	cu	chu	385
slovaque	sk	slo/slk	386
slovène	sl	slv	387
sogdien		sog	388
somali	so	som	389
songhai		son	390
soninké		snk	391
sorabes, langues		wen	392
sotho du Nord		nso	393
sotho du Sud	st	sot	394

Langue	639-1	639-2	Numéro
soundanais	su	sun	395
soussou		sus	396
suédois	sv	swe	397
sukuma		suk	398
sumérien		sux	399
swahili	sw	swa	400
swati	ss	ssw	401
syriaque		syr	402
tadjik	tg	tgk	403
tagalog	tl	tgl	404
tahitien	ty	tah	405
tamatchek		tmh	406
tamoul	ta	tam	407
tatar	tt	tat	408
tatar de Crimée		crh	409
tchèque	cs	cze/ces	410
tchétchène	ce	che	411
tchouvache	cv	chv	412
télougou	te	tel	413
temne		tem	414
tereno		ter	415
tetum		tet	416
thaï	th	tha	417
thaï, autres langues		tai	418
tibétain	bo	tib/bod	419
tigré		tig	420
tigrigna	ti	tir	421
tiv		tiv	422
tlingit		tli	423
tok pisin		tpi	424
tokelau		tkl	425
tonga (Nyasa)		tog	426
tongan (Îles Tonga)	to	ton	427
touva		tyv	428
tsigane		rom	429
tsimshian		tsi	430
tsonga	ts	tso	431
tswana	tn	tsn	432
tumbuka		tum	433
tupi, langues		tup	434
turc	tr	tur	435
turc ottoman (1500-1928)		ota	436
turkmène	tk	tuk	437
tuvalu		tvl	438
twi	tw	twi	439
ukrainien	uk	ukr	440
umbundu		umb	441
vaï		vai	442
venda	ve	ven	443
vietnamien	vi	vie	444
volapük	vo	vol	445

Langue	639-1	639-2	Numéro
vote		vot	446
wakashennes, langues		wak	447
walamo		wal	448
wallon	wa	wln	449
waray		war	450
washo		was	451
wolof	wo	wol	452
xhosa	xh	xho	453
yao		yao	454
yapois		yap	455
yi de Sichuan	ii	iii	456
yiddish	yi	yid	457
yoruba	yo	yor	458
yupik, langues		ypk	459
zandé		znd	460
zapotèque		zap	461
zenaga		zen	462
zhuang; chuang	za	zha	463
zoulou	zu	zul	464
zuni		zun	465

A.3 TABLEAU ISO 15924 DES CODES DE SYSTÈMES D'ÉCRITURE

Le tableau suivant reproduit la liste des 99 codes utilisés par le projet de norme internationale ISO 15924 : « Technologies de l'information - Code pour la représentation des noms d'écritures »[1].

Groupe	Écriture	Symbole	Code
Écritures hiéroglyphiques et cunéiformes (000-099)			
	cunéiforme suméro-akkadien	Xsux	20
	démotique égyptien	Egyd	70
	hiératique égyptien	Egyh	60
	hiéroglyphes égyptiens	Egyp	50
	hiéroglyphes mayas	Maya	90
Écritures alphabétiques rédigées de droite à gauche (100-199)			
	ancien hongrois	Hung	176
	arabe	Arab	160
	araméen	Aram	130
	avestique	Aves	151
	cunéiforme ougaritique	Xuga	106
	cunéiforme persépolitain	Xpeo	105
	hébreu	Hebr	125
	mandéen	Mnda	140
	méroïtique	Mero	100
	mongol	Mong	145
	orkhon	Orkh	175
	pehlavî	Palv	150
	phénicien	Phnx	115
	syriaque	Syrc	135
	syriaque (variante estranghélo)	Syre	138
	syriaque (variante occidentale)	Syrj	137
	syriaque (variante orientale)	Syrn	136
	thâna	Thaa	170
	tifinagh (berbère)	Tfng	120
Écritures alphabétiques rédigées de gauche à droite (200-299)			
	ancien italique (étrusque, osque, etc.)	Ital	210
	ancien permien	Perm	227
	arménien	Armn	230
	bopomofo	Bopo	285
	cirth	Cirt	291
	cyrillique	Cyrl	220
	cyrillique (variante slavon)	Cyrs	221
	déséret (mormon)	Dsrt	250
	géorgien (assomtavrouli)	Geoa	240
	géorgien (mkhédrouli)	Geor	242
	géorgien (nouskhouri)	Geon	241
	glagolitique	Glag	225
	gotique	Goth	206
	grec	Grek	200
	hangul	Hang	286
	latin	Latn	215

[1] Voir http://www.evertype.com/standards/iso15924/document/index.html. Achat de la norme sur le site ISO : http://www.iso.org/iso/fr/CatalogueDetailPage.CatalogueDetail?CSNUMBER=29546&ICS1=1&ICS2=140&ICS3=10.

Groupe	Écriture	Symbole	Code
	latin (variante brisée)	Latf	217
	latin (variante gaélique)	Latg	216
	ogam	Ogam	212
	osmanais	Osma	260
	parole visible	Visp	280
	phonétique de Pollard	Plrd	282
	runique	Runr	211
	shavien (Shaw)	Shaw	281
	tengwar	Teng	290
Écritures dérivées du brâhmî (300-399)			
	batak	Batk	365
	bengali	Beng	325
	birman	Mymr	350
	bouguis	Bugi	367
	bouhide	Buhd	372
	brâhmî	Brah	300
	cham (čam, tcham)	Cham	358
	dévanâgarî (nâgarî)	Deva	315
	goudjarâtî (gujrâtî)	Gujr	320
	gourmoukhî	Guru	310
	hanounóo	Hano	371
	javanais	Java	360
	kannara (canara)	Knda	345
	kayah li	Kali	357
	kharosthî	Khar	305
	khmer	Khmr	354
	laotien	Laoo	356
	lepcha (róng)	Lepc	335
	malayâlam	Mlym	347
	oriyâ	Orya	327
	singhalais	Sinh	348
	tagal	Tglg	370
	tagbanoua	Tagb	373
	tamoul	Taml	346
	télougou	Telu	340
	thaï	Thai	352
	tibétain	Tibt	330
Écritures syllabiques (400-499)			
	(alias pour hiragana + katakana)	Hrkt	412
	chypriominoen	Cpmn	402
	éthiopique (éthiopien, ge'ez)	Ethi	430
	hiragana	Hira	410
	katakana	Kana	411
	linéaire A	Lina	400
	linéaire B	Linb	401
	pahawh hmong	Hmng	450
	syllabaire autochthone canadien unifié	Cans	440
	syllabaire chypriote	Cprt	403
	tchérokî	Cher	445
	vaï	Vaii	470
	yi	Yiii	460

Groupe	Écriture	Symbole	Code
Écritures idéographiques (500-599)			
	braille	Brai	570
	idéogrammes han	Hani	500
	symboles Bliss	Blis	550
Écritures non déchiffrées (600-699)			
	indus	Inds	610
	phaistos	Phst	600
	rongorongo	Roro	620
Non attribué (700-799)			
Non attribué (800-899)			
Codets à usage privé, alias, codets spéciaux (900-999)			
	(alias pour han + hiragana + katakana)	Jpan	930
	codet pour écriture indéterminée	Zyyy	998
	codet pour écriture non codée	Zzzz	999
	codet pour les langues non ? crites	Zxxx	997
	réservé à l'usage privé	Qaaa-Qtzz	900-919

A.4 TABLEAU LANGUES - SYSTÈMES D'ÉCRITURE

Le tableau ci-dessous est dérivé du site Unicode[1]. Il présente une liste de quatre-vingt dix-neuf langues et les systèmes d'écriture utilisés pour les écrire.

Les annotations suivantes sont utilisées.
[1] = Pas encore dans Unicode.
[2] = Possède une ou plusieurs écritures éteintes ou mineures pas encore dans Unicode.
[3] = Ecriture utilisée dans le passé, une autre écriture étant utilisée actuellement.

Langue	Officielle	Ecriture	Numéro
abaza	N	cyrillique	1
abkhaz	N	cyrillique	2
adygei	N	cyrillique	3
afrikaans	O	latin	4
aïnou	N	katakanas, latin	5
aisor	N	cyrillique	6
albanais	O	latin [2]	7
altai	N	cyrillique	8
allemand	O	latin	9
amharique	O	éthiopien	10
amo	N	latin	11
anglais	O	latin	12
arabe	O	arabe	13
arménien	O	arménien, syriaque [3]	14
assamais	N	bengali	15
assyrien moderne	N	syriaque	16
avar	N	cyrillique	17
awadhi	N	devanagari	18
aymara	N	latin	19
azéri	N	cyrillique, latin	20
azerbaïdjanais	O	arabe, cyrillique, latin	21
badaga	N	tamoul	22
bagheli	N	devanagari	23
balear	N	latin	24
balkar	N	cyrillique	25
balti	N	devanagari, balti [2]	26
bashkir	N	cyrillique	27
basque	N	latin	28
batak	N	batak [1], latin	29
batak toba	N	batak [1], latin	30
bateri	N	devanagari	31
bengali	O	bengali	32
bhili	N	devanagari	33
bhojpuri	N	devanagari	34
bichlamar	O	latin	35
biélorusse	O	cyrillique	36
bihari	N	devanagari	37

[1] http://www.unicode.org/unicode/onlinedat/languages-scripts.html.

Langue	Officielle	Ecriture	Numéro
birman	O	birman	38
bokmål	O	latin	39
bulgare	O	cyrillique	40
cantonais	O	han	41
catalan	O	latin	42
chinois	O	han	43
coréen	O	hangul + han	44
danois	O	latin	45
dzongkha	O	tibétain	46
espagnol	O	latin	47
estonien	O	latin	48
farsi /dari	O	arabe	49
filipino/tagalog	O	latin, tagalog	50
finnois	O	latin	51
français	O	latin	52
géorgien	O	géorgien	53
grec	O	grec	54
hébreu	O	hébreu	55
hindi	O	devanagari	56
hongrois	O	latin	57
irlandais	O	latin	58
islandais	O	latin	59
italien	O	latin	60
japonais	O	han + hiraganas + katakanas	61
kazakh	O	cyrillique	62
khmer	O	khmer	63
kirghiz	O	arabe [3], latin, cyrillique	64
laotien	O	laotien	65
letton	O	latin	66
lituanien	O	latin	67
macédonien	O	cyrillique	68
malais	O	arabe [3], latin	69
maldivien	O	thaana	70
malgache	O	latin	71
maltais	O	latin	72
mongol	O	mongol, cyrillique	73
néerlandais	O	latin	74
népali	O	devanagari	75
ourdou	O	ourdou	76
ouzbek	O	cyrillique, latin	77
pashtou	O	arabe	78
polonais	O	latin	79
portugais	O	latin	80
roumain/moldave	O	latin, cyrillique [3] (roumain), cyrillique (moldave)	81
russe	O	cyrillique	82
serbo-croate	O	cyrillique (serbe), latin (croate)	83
singhalais	O	singhalais	84
slovaque	O	latin	85

Langue	Officielle	Ecriture	Numéro
slovène	O	latin	86
somali	O	latin	87
suédois	O	latin	88
swahili	O	latin	89
tadjik	O	arabe [3], latin, cyrillique (=> latin)	90
tamoul	O	tamoul	91
tchèque	O	latin	92
thaï	O	thaï	93
tingrinya	O	éthiopien	94
turc	O	arabe [3], latin	95
turkmène	O	arabe [3], latin, cyrillique (=> latin)	96
tuvaluan	O	cyrillique	97
ukrainien	O	cyrillique	98
vietnamien	O	latin, chu nom	99

A.5 PRINCIPALES FAMILLES DE LANGUES

La liste ci-dessous présente l'arborescence des familles de langues fournie par Ethnologue. Elle se limite à trois niveaux qui peuvent être des sous-familles (en minuscules) ou des langues, dans le cas de ramifications peu nombreuses (en majuscules). Le seul premier niveau comporte cent huit familles qui sont marquées en gras. L'arborescence complète est accessible dans [Grimes & Grimes 2000] ou sur le site d'Ethnologue, à l'adresse http://www.ethnologue.org/family_index.asp. Cette liste des familles de langues est cohérente avec l'annexe A.6 qui donne, pour plus de huit cent langues, la famille à laquelle elle est rattachée, le nombre de locuteurs et le code ethnologue. C'est ce même code que l'on peut voir ci-dessous après le nom des langues. Les données de cette annexe sont fournies telles que présentées sur le site Ethnogue, en particulier sans la traduction en français des noms de familles et de langues.

- **Afro-Asiatic** (372)
  - Berber (26)
    - Eastern (3)
    - Guanche (1)
    - Northern (17)
    - Tamasheq (4)
    - Zenaga (1)
  - Chadic (195)
    - Biu-Mandara (79)
    - East (34)
    - Masa (9)
    - West (73)
  - Cushitic (47)
    - Central (5)
    - East (34)
    - North (1)
    - South (7)
  - Egyptian (1)
  - Omotic (28)
    - North (24)
    - South (4)
  - Semitic (74)
    - Central (57)
    - South (17)
  - Unclassified (1)
    - BIRALE [BXE] (Ethiopia)
- **Alacalufan** (2)
  - KAKAUHUA [KBF] (Chile)
  - QAWASQAR [ALC] (Chile)
- **Algic** (40)
  - Algonquian (38)
    - Central (23)
    - Eastern (10)
    - Plains (4)
    - Unclassified (1)
  - Wiyot (1)
    - WIYOT [WIY] (USA)
  - Yurok (1)
    - YUROK [YUR] (USA)
- **Altaic** (65)
  - Mongolian (13)
    - Eastern (12)
    - Western (1)
  - Tungus (12)
    - Northern (4)
    - Southern (8)
  - Turkic (40)
    - Bolgar (1)
    - Eastern (7)
    - Northern (8)
    - Southern (12)
    - URUM [UUM] (Georgia)
    - Western (11)
- **Amto-Musan** (2)
  - AMTO [AMT] (Papua New Guinea)
  - MUSAN [MMP] (Papua New Guinea)
- **Andamanese** (13)
  - Great Andamanese (10)
    - Central (6)
    - Northern (4)
  - South Andamanese (3)
    - JARAWA [ANQ] (India)
    - ÖNGE [OON] (India)
    - SENTINEL [STD] (India)
- **Arauan** (8)
  - ARUA [ARA] (Brazil)
  - BANAWÁ [BNH] (Brazil)
  - CULINA [CUL] (Brazil)
  - DENÍ [DAN] (Brazil)
  - JAMAMADÍ [JAA] (Brazil)
  - JARUÁRA [JAP] (Brazil)
  - PAUMARÍ [PAD] (Brazil)
  - SURUAHÁ [SWX] (Brazil)
- **Araucanian** (2)
  - HUILLICHE [HUH] (Chile)
  - MAPUDUNGUN [ARU] (Chile)
- **Arawakan** (60)
  - Maipuran (54)
    - Central Maipuran (6)
    - Eastern Maipuran (1)
    - Northern Maipuran (24)
    - Southern Maipuran (21)
    - Western Maipuran (2)
  - Unclassified (6)
    - CHANÉ [CAJ] (Argentina)
    - CUMERAL [CUM] (Colombia)
    - OMEJES [OME] (Colombia)
    - PONARES [POD] (Colombia)
    - TOMEDES [TOE] (Colombia)
    - XIRIÂNA [XIR] (Brazil)
- **Artificial language** (3)
  - ESPERANTO [ESP] (France)
  - EUROPANTO [EUR] (Belgium)
  - INTERLINGUA [INR] (France)
- **Arutani-Sape** (2)
  - ARUTANI [ATX] (Brazil)
  - SAPÉ [SPC] (Venezuela)
- **Australian** (258)
  - Bunaban (2)
    - BUNABA [BCK] (Australia)
    - GOONIYANDI [GNI] (Australia)
  - Burarran (4)
    - BURARRA [BVR] (Australia)
    - DJEEBBANA [DJJ] (Australia)
    - GURAGONE [GGE] (Australia)
    - NAKARA [NCK] (Australia)

- Daly (19)
  - Bringen-Wagaydy (13)
  - Malagmalag (4)
  - Moil (1)
  - Murrinh-Patha (1)
- Djamindjungan (2)
  - DJAMINDJUNG [DJD] (Australia)
  - NUNGALI [NUG] (Australia)
- Djeragan (3)
  - Kitjic (1)
  - Miriwungic (2)
- Enindhilyagwa (1)
  - ANINDILYAKWA [AOI] (Australia)
- Gagudjuan (1)
  - GAGADU [GBU] (Australia)
- Garawan (1)
  - GARAWA [GBC] (Australia)
- Gungaraganyan (1)
  - KUNGARAKANY [GGK] (Australia)
- Gunwingguan (13)
  - Djauanic (1)
  - Gunwinggic (3)
  - Mangarayic (1)
  - Ngalakanic (1)
  - Ngandic (1)
  - Nunggubuan (1)
  - Rembargic (1)
  - Warayan (1)
  - Yangmanic (3)
- Laragiyan (2)
  - LARAGIA [LRG] (Australia)
  - WULNA [WUX] (Australia)
- Mangerrian (3)
  - Mangerric (1)
  - Urninganggic (2)
- Maran (3)
  - Alawic (1)
  - Mara (2)
- Nyulnyulan (8)
  - BAADI [BCJ] (Australia)
  - DJAWI [DJW] (Australia)
  - DYABERDYABER [DYB] (Australia)
  - DYUGUN [DYD] (Australia)
  - NIMANBUR [NMP] (Australia)
  - NYIGINA [NYH] (Australia)
  - NYULNYUL [NYV] (Australia)
  - YAWURU [YWR] (Australia)
- Pama-Nyungan (177)
  - Arandic (6)
  - Baagandji (2)
  - Bandjalangic (1)
  - Barrow Point (1)
  - Dyirbalic (2)
  - Flinders Island (1)
  - Galgadungic (2)
  - Gumbaynggiric (1)
  - Kala Lagaw Ya (1)
  - Karnic (11)
  - Maric (12)
  - Muruwaric (1)
  - Ngarinyeric-Yithayithic (1)
  - Nyawaygic (1)
  - Paman (44)
  - South-West (51)
  - Tangic (4)
  - Wagaya-Warluwaric (3)
  - Waka-Kabic (4)
  - Warumungic (1)
  - Wiradhuric (3)
  - Yalandyic (4)
  - Yanyuwan (1)
  - Yidinic (2)
  - Yuin-Kuric (7)
  - Yuulngu (10)
- Tiwian (1)
  - TIWI [TIW] (Australia)
- Unclassified (3)
  - LIMILNGAN [LMC] (Australia)
  - NGURMBUR [NRX] (Australia)
  - UMBUGARLA [UMR] (Australia)
- West Barkly (3)
  - Jingalic (1)
  - Wambayan (2)
- Wororan (7)
  - Ungarinjinic (2)
  - Wororic (1)
  - Wunambalic (4)
- Yiwaidjan (4)
  - Amaragic (1)
  - Margic (1)
  - Yiwaidjic (2)

Austro-Asiatic (168)
- Mon-Khmer (147)
  - Aslian (19)
  - Eastern Mon-Khmer (67)
  - Monic (2)
  - Nicobar (6)
  - Northern Mon-Khmer (38)
  - Palyu (1)
  - Unclassified (4)
  - Viet-Muong (10)
- Munda (21)
  - North Munda (12)
  - South Munda (9)

Austronesian (1262)
- Formosan (23)
  - Atayalic (2)
  - Paiwanic (17)
  - Tsouic (4)
- Malayo-Polynesian (1239)
  - Central-Eastern (706)
  - Unclassified (2)
  - Western Malayo-Polynesian (531)

Aymaran (3)
- AYMARA, CENTRAL [AYM] (Bolivia)
- AYMARA, SOUTHERN [AYC] (Peru)

Barbacoan (7)
- Andaqui (1)
  - ANDAQUI [ANA] (Colombia)
- Cayapa-Colorado (2)
  - CHACHI [CBI] (Ecuador)
  - COLORADO [COF] (Ecuador)
- Coconucan (2)
  - GUAMBIANO [GUM] (Colombia)
  - TOTORO [TTK] (Colombia)
- Pasto (2)
  - BARBACOAS [BPB] (Colombia)
  - AWA-CUAIQUER [KWI] (Colombia)

Basque (3)
- BASQUE [BSQ] (Spain)
- BASQUE, NAVARRO-LABOURDIN [BQE] (France)
- BASQUE, SOULETIN [BSZ] (France)

Bayono-Awbono (2)

- AWBONO [AWH] (Indonesia (Irian Jaya))
- BAYONO [BYL] (Indonesia (Irian Jaya))

Caddoan (5)
- Northern (4)
  - Pawnee-Kitsai (3)
  - Wichita (1)
- Southern (1)
  - CADDO [CAD] (USA)

Cahuapanan (2)
- CHAYAHUITA [CBT] (Peru)
- JEBERO [JEB] (Peru)

Cant (1)
- English-Tahitian (1)
  - PITCAIRN-NORFOLK [PIH] (Norfolk Island)

Carib (29)
- Northern (21)
  - Coastal (3)
  - East-West Guiana (12)
  - Galibi (1)
  - Northern Brazil (2)
  - Western Guiana (3)
- Southern (8)
  - Southeastern Colombia (1)
  - Southern Guiana (3)
  - Xingu Basin (4)

Chapacura-Wanham (5)
- Guapore (2)
  - ITENE [ITE] (Bolivia)
  - KABIXÍ [KBD] (Brazil)
- Madeira (3)
  - ORO WIN [ORW] (Brazil)
  - PAKAÁSNOVOS [PAV] (Brazil)
  - TORÁ [TRZ] (Brazil)

Chibchan (22)
- Aruak (3)
  - ICA [ARH] (Colombia)
  - COGUI [KOG] (Colombia)
  - MALAYO [MBP] (Colombia)
- Chibchan Proper (5)
  - CHIBCHA [CBF] (Colombia)
  - Tunebo (4)
- Cofan (1)
  - COFÁN [CON] (Ecuador)
- Guaymi (2)
  - NGÄBERE [GYM] (Panama)
  - BUGLERE [SAB] (Panama)
- Kuna (2)
  - KUNA, SAN BLAS [CUK] (Panama)
  - KUNA, BORDER [KUA] (Colombia)
- Motilon (1)
  - MOTILÓN [MOT] (Colombia)
- Paya (1)
  - PECH [PAY] (Honduras)
- Rama (2)
  - MALÉKU JAÍKA [GUT] (Costa Rica)
  - RAMA [RMA] (Nicaragua)
- Talamanca (4)
  - BORUCA [BRN] (Costa Rica)
  - BRIBRI [BZD] (Costa Rica)
  - CABÉCAR [CJP] (Costa Rica)
  - TERIBE [TFR] (Panama)
- Unclassified (1)
  - CHIMILA [CBG] (Colombia)

Chimakuan (1)
- QUILEUTE [QUI] (USA)

Choco (10)
- Embera (6)
  - Northern (2)
  - Southern (4)
- ANSERMA [ANS] (Colombia)
- ARMA [AOH] (Colombia)
- RUNA [RUN] (Colombia)
- WOUN MEU [NOA] (Panama)

Chon (2)
- ONA [ONA] (Argentina)
- TEHUELCHE [TEH] (Argentina)

Chukotko-Kamchatkan (5)
- Northern (4)
  - Chukot (1)
  - Koryak-Alyutor (3)
- Southern (1)
  - ITELMEN [ITL] (Russia (Asia))

Chumash (7)
- BARBAREÑO [BOI] (USA)
- CHUMASH [CHS] (USA)
- CRUZEÑO [CRZ] (USA)
- INESEÑO [INE] (USA)
- OBISPEÑO [OBI] (USA)
- PURISIMEÑO [PUY] (USA)
- VENTUREÑO [VEO] (USA)

Coahuiltecan (1)
- TONKAWA [TON] (USA)

Creole (81)
- Afrikaans based (2)
  - TSOTSITAAL [FLY] (South Africa)
  - OORLAMS [OOR] (South Africa)
- Arabic based (3)
  - ARABIC, BABALIA CREOLE [BBZ] (Chad)
  - NUBI [KCN] (Uganda)
  - ARABIC, SUDANESE CREOLE [PGA] (Sudan)
- Assamese based (1)
  - NAGA PIDGIN [NAG] (India)
- Dutch based (4)
  - BERBICE CREOLE DUTCH [BRC] (Guyana)
  - DUTCH CREOLE [DCR] (U.S. Virgin Islands)
  - PETJO [PEY] (Indonesia (Java and Bali))
  - SKEPI CREOLE DUTCH [SKW] (Guyana)
- English based (30)
  - Atlantic (22)
  - Pacific (7)
  - SARAMACCAN [SRM] (Suriname)
- French based (11)
  - AMAPÁ CREOLE [AMD] (Brazil)
  - TAYO [CKS] (New Caledonia)
  - SESELWA CREOLE FRENCH [CRS] (Seychelles)
  - LESSER ANTILLEAN CREOLE FRENCH [DOM] (St. Lucia)
  - FRENCH GUIANESE CREOLE FRENCH [FRE] (French Guiana)
  - HAITIAN CREOLE FRENCH [HAT] (Haiti)
  - KARIPÚNA CREOLE FRENCH [KMV] (Brazil)
  - LOUISIANA CREOLE FRENCH [LOU] (USA)
  - MORISYEN [MFE] (Mauritius)

- RÉUNION CREOLE FRENCH [RCF] (Reunion)
- SAN MIGUEL CREOLE FRENCH [SME] (Panama)

German based (1)
- UNSERDEUTSCH [ULN] (Papua New Guinea)

Iberian based (1)
- PAPIAMENTU [PAE] (Netherlands Antilles)

Indonesian based (1)
- INDONESIAN, PERANAKAN [PEA] (Indonesia (Java and Bali))

Kongo based (2)
- KITUBA [KTU] (Democratic Republic of Congo)
- MUNUKUTUBA [MKW] (Congo)

Malay based (6)
- MALAY, AMBONESE [ABS] (Indonesia (Maluku))
- MALAY, BABA [BAL] (Singapore)
- BETAWI [BEW] (Indonesia (Java and Bali))
- MALACCAN CREOLE MALAY [CCM] (Malaysia (Peninsular))
- MALAY, KUPANG [MKN] (Indonesia (Nusa Tenggara))
- SRI LANKAN CREOLE MALAY [SCI] (Sri Lanka)

Ngbandi based (2)
- SANGO [SAJ] (Central African Republic)
- SANGO, RIVERAIN [SNJ] (Central African Republic)

Portuguese based (13)
- ANGOLAR [AOA] (São Tomé e Principe)
- CAFUNDO CREOLE [CCD] (Brazil)
- SÃOTOMENSE [CRI] (São Tomé e Principe)
- FA D'AMBU [FAB] (Equatorial Guinea)
- INDO-PORTUGUESE [IDB] (Sri Lanka)
- KABUVERDIANU [KEA] (Cape Verde Islands)
- MALACCAN CREOLE PORTUGUESE [MCM] (Malaysia (Peninsular))
- MACANESE [MZS] (China)
- CRIOULO, UPPER GUINEA [POV] (Guinea-Bissau)
- PRINCIPENSE [PRE] (São Tomé e Principe)
- TERNATEÑO [TMG] (Indonesia (Maluku))
- PIDGIN, TIMOR [TVY] (Timor Lorosae)
- KORLAI CREOLE PORTUGUESE [VKP] (India)

Spanish based (2)
- CHAVACANO [CBK] (Philippines)
- PALENQUERO [PLN] (Colombia)

Swahili based (1)
- CUTCHI-SWAHILI [CCL] (Kenya)

Tetun based (1)
- TETUM PRASA [TDT] (Timor Lorosae)

Deaf sign language (114)
- ADAMOROBE SIGN LANGUAGE [ADS] (Ghana)
- ALGERIAN SIGN LANGUAGE [ASP] (Algeria)
- AMERICAN SIGN LANGUAGE [ASE] (USA)
- ARGENTINE SIGN LANGUAGE [AED] (Argentina)
- ARMENIAN SIGN LANGUAGE [AEN] (Armenia)
- AUSTRALIAN ABORIGINES SIGN LANGUAGE [ASW] (Australia)
- AUSTRALIAN SIGN LANGUAGE [ASF] (Australia)
- AUSTRIAN SIGN LANGUAGE [ASQ] (Austria)
- BALI SIGN LANGUAGE [BQY] (Indonesia (Java and Bali))
- BAMAKO SIGN LANGUAGE [BOG] (Mali)
- BAN KHOR SIGN LANGUAGE [BLA] (Thailand)
- BELGIAN SIGN LANGUAGE [BVS] (Belgium)
- BOLIVIAN SIGN LANGUAGE [BVL] (Bolivia)
- BRAZILIAN SIGN LANGUAGE [BZS] (Brazil)
- BRITISH SIGN LANGUAGE [BHO] (United Kingdom)
- BULGARIAN SIGN LANGUAGE [BQN] (Bulgaria)
- CATALONIAN SIGN LANGUAGE [CSC] (Spain)
- CHADIAN SIGN LANGUAGE [CDS] (Chad)
- CHIANGMAI SIGN LANGUAGE [CSD] (Thailand)
- CHILEAN SIGN LANGUAGE [CSG] (Chile)
- CHINESE SIGN LANGUAGE [CSL] (China)
- COLOMBIAN SIGN LANGUAGE [CSN] (Colombia)
- COSTA RICAN SIGN LANGUAGE [CSR] (Costa Rica)
- CZECH SIGN LANGUAGE [CSE] (Czech Republic)
- DANISH SIGN LANGUAGE [DSL] (Denmark)
- DOMINICAN SIGN LANGUAGE [DOQ] (Dominican Republic)
- DUTCH SIGN LANGUAGE [DSE] (Netherlands)
- ECUADORIAN SIGN LANGUAGE [ECS] (Ecuador)
- ESTONIAN SIGN LANGUAGE [ESO] (Estonia)
- ETHIOPIAN SIGN LANGUAGE [ETH] (Ethiopia)
- FINNISH SIGN LANGUAGE [FSE] (Finland)
- FRENCH SIGN LANGUAGE [FSL] (France)
- GERMAN SIGN LANGUAGE [GSG] (Germany)
- GHANAIAN SIGN LANGUAGE [GSE] (Ghana)
- GREEK SIGN LANGUAGE [GSS] (Greece)
- GUATEMALAN SIGN LANGUAGE [GSM] (Guatemala)
- GUINEAN SIGN LANGUAGE [GUS] (Guinea)
- HAIPHONG SIGN LANGUAGE [HAF] (Viet Nam)
- HANOI SIGN LANGUAGE [HAB] (Viet Nam)
- HAUSA SIGN LANGUAGE [HSL] (Nigeria)
- HAWAI'I PIDGIN SIGN LANGUAGE [HPS] (USA)

HO CHI MINH CITY SIGN LANGUAGE [HOS] (Viet Nam)
HUNGARIAN SIGN LANGUAGE [HSH] (Hungary)
ICELANDIC SIGN LANGUAGE [ICL] (Iceland)
INDIAN SIGN LANGUAGE [INS] (India)
INDONESIAN SIGN LANGUAGE [INL] (Indonesia (Java and Bali))
IRISH SIGN LANGUAGE [ISG] (Ireland)
ISRAELI SIGN LANGUAGE [ISL] (Israel)
ITALIAN SIGN LANGUAGE [ISE] (Italy)
JAMAICAN COUNTRY SIGN LANGUAGE [JCS] (Jamaica)
JAPANESE SIGN LANGUAGE [JSL] (Japan)
JORDANIAN SIGN LANGUAGE [JOS] (Jordan)
KENYAN SIGN LANGUAGE [XKI] (Kenya)
KOREAN SIGN LANGUAGE [KVK] (Korea, South)
KUALA LUMPUR SIGN LANGUAGE [KGI] (Malaysia (Peninsular))
LAOS SIGN LANGUAGE [LSO] (Laos)
LATVIAN SIGN LANGUAGE [LSL] (Latvia)
LIBYAN SIGN LANGUAGE [LBS] (Libya)
LITHUANIAN SIGN LANGUAGE [LLS] (Lithuania)
LYONS SIGN LANGUAGE [LSG] (France)
MALAYSIAN SIGN LANGUAGE [XML] (Malaysia (Peninsular))
MALTESE SIGN LANGUAGE [MDL] (Malta)
MARITIME SIGN LANGUAGE [NSR] (Canada)
MARTHA'S VINEYARD SIGN LANGUAGE [MRE] (USA)
MEXICAN SIGN LANGUAGE [MFS] (Mexico)
MONGOLIAN SIGN LANGUAGE [QMM] (Mongolia)
MOROCCAN SIGN LANGUAGE [XMS] (Morocco)
MOZAMBICAN SIGN LANGUAGE [MZY] (Mozambique)
NAMIBIAN SIGN LANGUAGE [NBS] (Namibia)
NEPALESE SIGN LANGUAGE [NSP] (Nepal)
NEW ZEALAND SIGN LANGUAGE [NZS] (New Zealand)
NICARAGUAN SIGN LANGUAGE [NCS] (Nicaragua)
NIGERIAN SIGN LANGUAGE [NSI] (Nigeria)
NORWEGIAN SIGN LANGUAGE [NSL] (Norway)
OLD KENTISH SIGN LANGUAGE [OKL] (United Kingdom)
PAKISTAN SIGN LANGUAGE [PKS] (Pakistan)
PENANG SIGN LANGUAGE [PSG] (Malaysia (Peninsular))
PERSIAN SIGN LANGUAGE [PSC] (Iran)
PERUVIAN SIGN LANGUAGE [PRL] (Peru)
PHILIPPINE SIGN LANGUAGE [PSP] (Philippines)
POLISH SIGN LANGUAGE [PSO] (Poland)
PORTUGUESE SIGN LANGUAGE [PSR] (Portugal)
PROVIDENCIA SIGN LANGUAGE [PRO] (Colombia)
PUERTO RICAN SIGN LANGUAGE [PSL] (Puerto Rico)
QUEBEC SIGN LANGUAGE [FCS] (Canada)
RENNELLESE SIGN LANGUAGE [RSI] (Solomon Islands)
ROMANIAN SIGN LANGUAGE [RMS] (Romania)
RUSSIAN SIGN LANGUAGE [RSL] (Russia (Europe))
SALVADORAN SIGN LANGUAGE [ESN] (El Salvador)
SAUDI ARABIAN SIGN LANGUAGE [SDL] (Saudi Arabia)
SINGAPORE SIGN LANGUAGE [SLS] (Singapore)
SLOVAKIAN SIGN LANGUAGE [SVK] (Slovakia)
SOUTH AFRICAN SIGN LANGUAGE [SFS] (South Africa)
SPANISH SIGN LANGUAGE [SSP] (Spain)
SRI LANKAN SIGN LANGUAGE [SQS] (Sri Lanka)
SWEDISH SIGN LANGUAGE [SWL] (Sweden)
SWISS-FRENCH SIGN LANGUAGE [SSR] (Switzerland)
SWISS-GERMAN SIGN LANGUAGE [SGG] (Switzerland)
SWISS-ITALIAN SIGN LANGUAGE [SLF] (Switzerland)
TAIWANESE SIGN LANGUAGE [TSS] (Taiwan)
TANZANIAN SIGN LANGUAGE [TZA] (Tanzania)
THAI SIGN LANGUAGE [TSQ] (Thailand)
TUNISIAN SIGN LANGUAGE [TSE] (Tunisia)
TURKISH SIGN LANGUAGE [TSM] (Turkey (Asia))
UGANDAN SIGN LANGUAGE [UGN] (Uganda)
UKRAINIAN SIGN LANGUAGE [UKL] (Ukraine)
URUBÚ-KAAPOR SIGN LANGUAGE [UKS] (Brazil)
URUGUAYAN SIGN LANGUAGE [UGY] (Uruguay)
VENEZUELAN SIGN LANGUAGE [VSL] (Venezuela)
YIDDISH SIGN LANGUAGE [YDS] (Israel)
YUCATEC MAYA SIGN LANGUAGE [MSD] (Mexico)
YUGOSLAVIAN SIGN LANGUAGE [YSL] (Yugoslavia)
ZAMBIAN SIGN LANGUAGE [ZSL] (Zambia)
ZIMBABWE SIGN LANGUAGE [ZIB] (Zimbabwe)

Dravidian (75)
Central (5)
Kolami-Naiki (2)
Parji-Gadaba (3)
Northern (5)
BRAHUI [BRH] (Pakistan)
KUMARBHAG PAHARIA [KMJ] (India)
KURUX [KVN] (India)
KURUX, NEPALI [KXL] (Nepal)
SAURIA PAHARIA [MJT] (India)
South Central (1)

- Telugu (1)
- South-Central (22)
  - Gondi-Kui (18)
  - Telugu (4)
- Southern (33)
  - Tamil-Kannada (27)
  - Tulu (5)
  - Unclassified (1)
- Unclassified (9)
  - ALLAR [ALL] (India)
  - BAZIGAR [BFR] (India)
  - BHARIA [BHA] (India)
  - KAMAR [KEQ] (India)
  - KANIKKARAN [KEV] (India)
  - KURICHIYA [KFH] (India)
  - MALANKURAVAN [MJO] (India)
  - MUTHUVAN [MUV] (India)
  - VISHAVAN [VIS] (India)

East Bird's Head (3)
- Meax (2)
  - MEYAH [MEJ] (Indonesia (Irian Jaya))
  - MOSKONA [MTJ] (Indonesia (Irian Jaya))
- MANIKION [MNX] (Indonesia (Irian Jaya))

East Papuan (36)
- Bougainville (13)
  - East (9)
  - West (4)
- Reef Islands-Santa Cruz (3)
  - NANGGU [NAN] (Solomon Islands)
  - AYIWO [NFL] (Solomon Islands)
  - SANTA CRUZ [STC] (Solomon Islands)
- Yele-Solomons-New Britain (20)
  - New Britain (12)
  - Yele-Solomons (8)

Eskimo-Aleut (11)
- Eskimo (10)
  - Inuit (5)
  - Yupik (5)
- ALEUT [ALW] (USA)

Geelvink Bay (33)
- East Geelvink Bay (11)
  - ANASI [BPO] (Indonesia (Irian Jaya))
  - BARAPASI [BRP] (Indonesia (Irian Jaya))
  - BURATE [BTI] (Indonesia (Irian Jaya))
  - DEMISA [DEI] (Indonesia (Irian Jaya))
  - KOFEI [KPI] (Indonesia (Irian Jaya))
  - NISA [NIC] (Indonesia (Irian Jaya))
  - BAUZI [PAU] (Indonesia (Irian Jaya))
  - SAURI [SAH] (Indonesia (Irian Jaya))
  - TEFARO [TFO] (Indonesia (Irian Jaya))
  - TUNGGARE [TRT] (Indonesia (Irian Jaya))
  - WORIA [WOR] (Indonesia (Irian Jaya))
- Lakes Plain (20)
  - Awera (1)
  - East Lakes Plain (2)
  - Rasawa-Saponi (2)
  - Tariku (15)
- Yawa (2)
  - SAWERU [SWR] (Indonesia (Irian Jaya))
  - YAWA [YVA] (Indonesia (Irian Jaya))

Guahiban (5)
- CUIBA [CUI] (Colombia)
- GUAHIBO [GUH] (Colombia)
- GUAYABERO [GUO] (Colombia)
- MACAGUÁN [MBN] (Colombia)
- PLAYERO [GOB] (Colombia)

Gulf (4)
- ATAKAPA [ALE] (USA)
- CHITIMACHA [CHM] (USA)
- NATCHEZ [NCZ] (USA)
- TUNICA [TUK] (USA)

Harakmbet (2)
- AMARAKAERI [AMR] (Peru)
- HUACHIPAERI [HUG] (Peru)

Hmong-Mien (32)
- Hmongic (26)
  - Bunu (5)
  - Chuanqiandian (16)
  - Qiandong (3)
  - Xiangxi (2)
- Ho Nte (1)
  - SHE [SHX] (China)
- Mienic (5)
  - Biao-Jiao (1)
  - Mian-Jin (3)
  - Zaomin (1)

Hokan (28)
- Esselen-Yuman (10)
  - Esselen (1)
  - Yuman (9)
- Northern (13)
  - CHIMARIKO [CID] (USA)
  - Karok-Shasta (4)
  - Pomo (7)
  - Yana (1)
- Salinan-Seri (2)
  - SALINAN [SAL] (USA)
  - SERI [SEI] (Mexico)
- Tequistlatecan (2)
  - CHONTAL DE OAXACA, SIERRA [CHD] (Mexico)
  - CHONTAL DE OAXACA, COSTA [CLO] (Mexico)
- Washo (1)
  - WASHO [WAS] (USA)

Huavean (4)
- HUAVE, SAN DIONISIO DEL MAR [HVE] (Mexico)
- HUAVE, SAN FRANCISCO DEL MAR [HUE] (Mexico)
- HUAVE, SAN MATEO DEL MAR [HUV] (Mexico)
- HUAVE, SANTA MARÍA DEL MAR [HVV] (Mexico)

Indo-European (443)
- Albanian (4)
  - Gheg (1)
  - Tosk (3)
- Armenian (2)
  - ARMENIAN [ARM] (Armenia)
  - LOMAVREN [RMI] (Armenia)
- Baltic (3)
  - Eastern (2)
  - Western (1)
- Celtic (7)
  - Insular (7)
- Germanic (58)
  - East (1)
  - North (14)
  - West (43)
- Greek (7)

- Attic (6)
- Doric (1)
- Indo-Iranian (296)
  - Indo-Aryan (210)
  - Iranian (84)
  - Unclassified (2)
- Italic (48)
  - Latino-Faliscan (1)
  - Romance (47)
- Slavic (18)
  - East (4)
  - South (6)
  - West (8)

Iroquoian (10)
- Northern iroquoian (8)
  - Five Nations (5)
  - Huron (1)
  - LAURENTIAN [LRE] (Canada)
  - Tuscarora-Nottoway (1)
- Southern iroquoian (1)
  - CHEROKEE [CER] (USA)
- SUSQUEHANNOCK [SQN] (USA)

Japanese (12)
- Japanese (1)
  - JAPANESE [JPN] (Japan)
- Ryukyuan (11)
  - Amami-Okinawan (8)
  - Sakishima (3)

Jivaroan (4)
- ACHUAR-SHIWIAR [ACU] (Peru)
- AGUARUNA [AGR] (Peru)
- HUAMBISA [HUB] (Peru)
- SHUAR [JIV] (Ecuador)

Katukinan (3)
- KANAMARÍ [KNM] (Brazil)
- KATAWIXI [QKI] (Brazil)
- KATUKÍNA [KAV] (Brazil)

Keres (2)
- KERES, EASTERN [KEE] (USA)
- KERES, WESTERN [KJQ] (USA)

Khoisan (29)
- Hatsa (1)
  - HADZA [HTS] (Tanzania)
- Sandawe (1)
  - SANDAWE [SBR] (Tanzania)
- Southern Africa (27)
  - Central (14)
  - Northern (7)
  - Southern (6)

Kiowa Tanoan (6)
- Kiowa-Towa (2)
  - Kiowa (1)
  - Towa (1)
- Tewa-Tiwa (4)
  - Tewa (1)
  - Tiwa (3)

Kwomtari-Baibai (6)
- Baibai (2)
  - BAIBAI [BBF] (Papua New Guinea)
  - NAI [BIO] (Papua New Guinea)
- Kwomtari (3)
  - FAS [FAS] (Papua New Guinea)
  - GURIASO [GRX] (Papua New Guinea)
  - KWOMTARI [KWO] (Papua New Guinea)
- Pyu (1)
  - PYU [PBY] (Papua New Guinea)

Language Isolate (30)
- ABINOMN [BSA] (Indonesia (Irian Jaya))
- AINU [AIN] (Japan)
- ANDOQUE [ANO] (Colombia)
- BURMESO [BZU] (Indonesia (Irian Jaya))
- BURUSHASKI [BSK] (Pakistan)
- BUSA [BHF] (Papua New Guinea)
- CAMSÁ [KBH] (Colombia)
- CAYUBABA [CAT] (Bolivia)
- GILYAK [NIV] (Russia (Asia))
- ITONAMA [ITO] (Bolivia)
- KARKAR-YURI [YUJ] (Papua New Guinea)
- KIBIRI [PRM] (Papua New Guinea)
- KOREAN [KKN] (Korea, South)
- KUTENAI [KUN] (Canada)
- NIHALI [NHL] (India)
- PANKARARÚ [PAZ] (Brazil)
- PUELCHE [PUE] (Argentina)
- PUINAVE [PUI] (Colombia)
- PURÉPECHA [TSZ] (Mexico)
- PURÉPECHA, SIERRA OCCIDENTAL [PUA] (Mexico)
- TICUNA [TCA] (Peru)
- TOL [JIC] (Honduras)
- TRUMAÍ [TPY] (Brazil)
- TUXÁ [TUD] (Brazil)
- WARAO [WBA] (Venezuela)
- YALE [NCE] (Papua New Guinea)
- YÁMANA [YAG] (Chile)
- YUCHI [YUC] (USA)
- YURACARE [YUE] (Bolivia)
- ZUNI [ZUN] (USA)

Left May (7)
- AMA [AMM] (Papua New Guinea)
- BO [BPW] (Papua New Guinea)
- ITERI [ITR] (Papua New Guinea)
- NAKWI [NAX] (Papua New Guinea)
- NIMO [NIW] (Papua New Guinea)
- OWINIGA [OWI] (Papua New Guinea)
- ROCKY PEAK [ROK] (Papua New Guinea)

Lower Mamberamo (2)
- WAREMBORI [WSA] (Indonesia (Irian Jaya))
- YOKE [YKI] (Indonesia (Irian Jaya))

Lule-Vilela (1)
- VILELA [VIL] (Argentina)

Macro-Ge (32)
- Bororo (3)
  - Bororo Proper (2)
  - Otuke (1)
- Botocudo (1)
  - KRENAK [KQQ] (Brazil)
- Chiquito (1)
  - CHIQUITANO [CAX] (Bolivia)
- Fulnio (1)
  - FULNIÔ [FUN] (Brazil)
- Ge-Kaingang (16)
  - Ge (13)
  - Kaingang (3)
- Guato (1)
  - GUATÓ [GTA] (Brazil)
- Kamakan (1)
  - KAMAKAN [VKM] (Brazil)
- Karaja (1)
  - KARAJÁ [KPJ] (Brazil)
- Maxakali (1)
  - MAXAKALÍ [MBL] (Brazil)
- Opaye (1)

- OPAYÉ [OPY] (Brazil)
- Puri (1)
  - PURI [PRR] (Brazil)
- Rikbaktsa (1)
  - RIKBAKTSA [ART] (Brazil)
- Yabuti (2)
  - ARIKAPÚ [ARK] (Brazil)
  - JABUTÍ [JBT] (Brazil)
- Oti (1)
  - OTI [OTI] (Brazil)

Maku (6)
- CACUA [CBV] (Colombia)
- HUPDË [JUP] (Brazil)
- KAMÃ [KWA] (Brazil)
- NADËB [MBJ] (Brazil)
- NUKAK MAKÚ [MBR] (Colombia)
- YUHUP [YAB] (Brazil)

Mascoian (5)
- EMOK [EMO] (Paraguay)
- GUANA [GVA] (Paraguay)
- LENGUA [LEG] (Paraguay)
- SANAPANÁ [SAP] (Paraguay)
- TOBA-MASKOY [TMF] (Paraguay)

Mataco-Guaicuru (11)
- Guaicuruan (4)
  - KADIWÉU [KBC] (Brazil)
  - MOCOVÍ [MOC] (Argentina)
  - PILAGÁ [PLG] (Argentina)
  - TOBA [TOB] (Argentina)
- Mataco (7)
  - CHULUPÍ [CAG] (Paraguay)
  - CHOROTE, IYO'WUJWA [CRQ] (Argentina)
  - CHOROTE, IYOJWA'JA [CRT] (Argentina)
  - WICHÍ LHAMTÉS VEJOZ [MAD] (Argentina)
  - MACA [MCA] (Paraguay)
  - WICHÍ LHAMTÉS NOCTEN [MTP] (Bolivia)
  - WICHÍ LHAMTÉS GÜISNAY [MZH] (Argentina)

Mayan (69)
- Cholan-Tzeltalan (12)
  - Cholan (4)
  - Tzeltalan (8)
- Huastecan (4)
  - CHICOMUCELTEC [COB] (Mexico)
  - HUASTECO, SAN FRANCISCO CHONTLA [HAU] (Mexico)
  - HUASTECO, TANTOYUCA [HUS] (Mexico)
  - HUASTECO, SAN LUÍS POTOSÍ [HVA] (Mexico)
- Kanjobalan-Chujean (8)
  - Chujean (3)
  - Kanjobalan (5)
- Quichean-Mamean (40)
  - Greater Mamean (11)
  - Greater Quichean (29)
- Yucatecan (5)
  - Mopan-Itza (2)
  - Yucatec-Lacandon (3)

Misumalpan (4)
- CACAOPERA [CCR] (El Salvador)
- MATAGALPA [MTN] (Nicaragua)
- MÍSKITO [MIQ] (Nicaragua)
- SUMO TAWAHKA [SUM] (Nicaragua)

Mixe-Zoque (16)
- Mixe (9)
  - Eastern Mixe (5)
  - Veracruz Mixe (2)
  - Western Mixe (2)
- Zoque (7)
  - Chiapas Zoque (3)
  - Oaxaca Zoque (1)
  - Veracruz Zoque (3)

Mixed Language (8)
- Cakchiquel-Quiche (1)
  - CAKCHIQUEL-QUICHE MIXED LANGUAGE [CKZ] (Guatemala)
- Chinese-Tibetan-Mongolian (1)
  - WUTUNHUA [WUH] (China)
- French-Cree (1)
  - MICHIF [CRG] (USA)
- German-Yiddish-Romani-Rotwelsch (1)
  - YENICHE [YEC] (Germany)
- Pare-Cushitic (1)
  - MBUGU [MHD] (Tanzania)
- Russian-Aleut (1)
  - MEDNYJ ALEUT [MUD] (Russia (Asia))
- Spanish-Quechua (1)
  - MEDIA LENGUA [MUE] (Ecuador)
- Zulu-Bantu (1)
  - CAMTHO [CMT] (South Africa)

Mosetenan (1)
- TSIMANÉ [CAS] (Bolivia)

Mura (1)
- MÚRA-PIRAHÃ [MYP] (Brazil)

Muskogean (6)
- Eastern (4)
  - ALABAMA [AKZ] (USA)
  - KOASATI [CKU] (USA)
  - MUSKOGEE [CRK] (USA)
  - MIKASUKI [MIK] (USA)
- Western (2)
  - CHOCTAW [CCT] (USA)
  - CHICKASAW [CIC] (USA)

Na-Dene (47)
- Haida (2)
  - HAIDA, NORTHERN [HAI] (Canada)
  - HAIDA, SOUTHERN [HAX] (Canada)
- Nuclear na-dene (45)
  - Athapaskan-Eyak (44)
  - Tlingit (1)

Nambiquaran (5)
- HALÓ TÉ SÚ [HLO] (Brazil)
- NAMBIKUÁRA, NORTHERN [MBG] (Brazil)
- NAMBIKUÁRA, SOUTHERN [NAB] (Brazil)
- SABANÊS [SAE] (Brazil)
- SARARÉ [SRR] (Brazil)

Niger-Congo (1489)
- Atlantic-Congo (1390)
  - Atlantic (64)
  - Ijoid (10)
  - Volta-Congo (1316)
- Kordofanian (31)
  - Kadugli (7)
  - Kordofanian Proper (24)
- Mande (68)
  - Eastern (18)
  - Western (50)

Nilo-Saharan (199)

- Berta (1)
  - BERTA [WTI] (Ethiopia)
- Central Sudanic (65)
  - East (22)
  - West (43)
- Eastern Sudanic (95)
  - Eastern (26)
  - Kuliak (3)
  - Nilotic (52)
  - Western (14)
- Komuz (6)
  - Gumuz (1)
  - Koman (5)
- Kunama (1)
  - KUNAMA [KUM] (Eritrea)
- Maban (9)
  - KARANGA [KTH] (Chad)
  - Mabang (8)
- Saharan (9)
  - Eastern (3)
  - Western (6)
- Songhai (9)
  - KORANDJE [KCY] (Algeria)
  - Northern (3)
  - Southern (5)
- Unclassified (1)
  - SHABO [SBF] (Ethiopia)
- Fur (3)
  - AMDANG [AMJ] (Chad)
  - FUR [FUR] (Sudan)
  - MIMI [MIV] (Chad)

North Caucasian (34)

- North Central (3)
  - Batsi (1)
  - Chechen-Ingush (2)
- Northeast (26)
  - Avaro-Andi-Dido (14)
  - Lak-Dargwa (2)
  - Lezgian (10)
- Northwest (5)
  - Abkhaz-Abazin (2)
  - Circassian (2)
  - Ubyx (1)

Oto-Manguean (172)

- Amuzgoan (3)
  - AMUZGO, GUERRERO [AMU] (Mexico)
  - AMUZGO, SAN PEDRO AMUZGOS [AZG] (Mexico)
  - AMUZGO, IPALAPA [AZM] (Mexico)
- Chiapanec-Mangue (2)
  - CHIAPANECO [CIP] (Mexico)
  - CHOROTEGA [CJR] (Costa Rica)
- Chinantecan (14)
  - CHINANTECO, COMALTEPEC [CCO] (Mexico)
  - CHINANTECO, OJITLÁN [CHJ] (Mexico)
  - CHINANTECO, QUIOTEPEC [CHQ] (Mexico)
  - CHINANTECO, VALLE NACIONAL [CHV] (Mexico)
  - CHINANTECO, OZUMACÍN [CHZ] (Mexico)
  - CHINANTECO, LEALAO [CLE] (Mexico)
  - CHINANTECO, LALANA [CNL] (Mexico)
  - CHINANTECO, TEPETOTUTLA [CNT] (Mexico)
  - CHINANTECO, PALANTLA [CPA] (Mexico)
  - CHINANTECO, CHILTEPEC [CSA] (Mexico)
  - CHINANTECO, SOCHIAPAN [CSO] (Mexico)
  - CHINANTECO, TEPINAPA [CTE] (Mexico)
  - CHINANTECO, TLACOATZINTEPEC [CTL] (Mexico)
  - CHINANTECO, USILA [CUS] (Mexico)
- Mixtecan (55)
  - Mixtec-Cuicatec (52)
  - Trique (3)
- Otopamean (17)
  - Chichimec (1)
  - Matlatzincan (2)
  - Otomian (11)
  - Pamean (3)
- Popolocan (17)
  - Chocho-Popolocan (8)
  - Ixcatecan (1)
  - Mazatecan (8)
- Zapotecan (64)
  - Chatino (6)
  - Zapotec (58)

Paezan (1)

- PÁEZ [PBB] (Colombia)

Panoan (30)

- Eastern (1)
  - KAXARARÍ [KTX] (Brazil)
- North-Central (7)
  - ATSAHUACA [ATC] (Peru)
  - ISCONAHUA [ISC] (Peru)
  - CAPANAHUA [KAQ] (Peru)
  - MARÚBO [MZR] (Brazil)
  - REMO [REM] (Peru)
  - SHIPIBO-CONIBO [SHP] (Peru)
  - SENSI [SNI] (Peru)
- Northern (3)
  - MATSÉS [MCF] (Peru)
  - MATÍS [MPQ] (Brazil)
  - PISABO [PIG] (Peru)
- South-Central (9)
  - Amahuaca (1)
  - Unclassified (1)
  - Yaminahua-Sharanahua (6)
  - Yora (1)
- Southeastern (2)
  - CASHINAHUA [CBS] (Peru)
  - KATUKÍNA, PANOAN [KNT] (Brazil)
- Southern (4)
  - CHÁCOBO [CAO] (Bolivia)
  - KARIPUNÁ [KUQ] (Brazil)
  - PACAHUARA [PCP] (Bolivia)
  - SHINABO [SHN] (Bolivia)
- Unclassified (2)
  - ARÁRA, ACRE [AXA] (Brazil)
  - PANOBO [PNO] (Peru)
- Western (2)
  - CASHIBO-CACATAIBO [CBR] (Peru)
  - NOCAMAN [NOM] (Peru)

Peba-Yaguan (2)

- YAGUA [YAD] (Peru)
- YAMEO [YME] (Peru)

Penutian (33)
- California penutian (1)
  - Wintuan (1)
- Chinookan (2)
  - CHINOOK [CHH] (USA)
  - WASCO-WISHRAM [WAC] (USA)
- Maiduan (4)
  - MAIDU, NORTHWEST [MAI] (USA)
  - MAIDU, NORTHEAST [NMU] (USA)
  - NISENAN [NSZ] (USA)
  - MAIDU, VALLEY [VMV] (USA)
- Oregon penutian (5)
  - Coast Oregon (3)
  - Kalapuyan (1)
  - Takelma (1)
- Plateau penutian (6)
  - Klamath-Modoc (1)
  - Sahaptin (5)
- Tsimshian (3)
  - GITXSAN [GIT] (Canada)
  - NISGA'A [NCG] (Canada)
  - TSIMSHIAN [TSI] (Canada)
- Unclassified (1)
  - MOLALE [MBE] (USA)
- Yok-Utian (11)
  - Utian (10)
  - Yokuts (1)

Pidgin (17)
- Amerindian (3)
  - CHINOOK WAWA [CRW] (Canada)
  - DELAWARE, PIDGIN [DEP] (USA)
  - MOBILIAN [MOD] (USA)
- English based (2)
  - Atlantic (1)
  - Pacific (1)
- French based (1)
  - TAY BOI [TAS] (Viet Nam)
- Hausa based (2)
  - BARIKANCHI [BXO] (Nigeria)
  - GIBANAWA [GIB] (Nigeria)
- Iha based (1)
  - IHA BASED PIDGIN [IHB] (Indonesia (Irian Jaya))
- Malay based (1)
  - BROOME PEARLING LUGGER PIDGIN [BPL] (Australia)
- Mascoian based (1)
  - MASKOY PIDGIN [MHH] (Paraguay)
- Motu based (1)
  - MOTU, HIRI [POM] (Papua New Guinea)
- Onin based (1)
  - ONIN BASED PIDGIN [ONX] (Indonesia (Irian Jaya))
- Romance-based (1)
  - LINGUA FRANCA [PML] (Tunisia)
- Swahili based (1)
  - SETTLA [STA] (Zambia)
- Zulu based (1)
  - FANAGOLO [FAO] (South Africa)
- NDYUKA-TRIO PIDGIN [NJT] (Suriname)

Quechuan (46)
- Quechua I (17)
  - QUECHUA, ANCASH, HUAYLAS [QAN] (Peru)
  - QUECHUA, ANCASH, CHIQUIAN [QEC] (Peru)
  - QUECHUA, ANCASH, CONCHUCOS, NORTHERN [QED] (Peru)
  - QUECHUA, ANCASH, CORONGO [QEE] (Peru)
  - QUECHUA, PASCO, SANTA ANA DE TUSI [QEF] (Peru)
  - QUECHUA, SAN RAFAEL-HUARIACA [QEG] (Peru)
  - QUECHUA, ANCASH, CONCHUCOS, SOUTHERN [QEH] (Peru)
  - QUECHUA, MARGOS-YAROWILCA-LAURICOCHA [QEI] (Peru)
  - QUECHUA, HUÁNUCO, HUAMALÍES-NORTHERN DOS DE MAYO [QEJ] (Peru)
  - QUECHUA, HUÁNUCO, PANAO [QEM] (Peru)
  - QUECHUA, ANCASH, SIHUAS [QES] (Peru)
  - QUECHUA, WANCA, JAUJA [QHJ] (Peru)
  - QUECHUA, WANCA, HUAYLLA [QHU] (Peru)
  - QUECHUA, NORTH JUNÍN [QJU] (Peru)
  - QUECHUA, NORTH LIMA, CAJATAMBO [QNL] (Peru)
  - QUECHUA, HUÁNUCO, HUALLAGA [QUB] (Peru)
  - QUECHUA, PASCO-YANAHUANCA [QUR] (Peru)
- Quechua II (29)
  - A (4)
  - B (14)
  - C (11)

Salishan (27)
- Bella Coola (1)
  - BELLA COOLA [BEL] (Canada)
- Central Salish (13)
  - Halkomelem (1)
  - Nooksack (1)
  - Northern (3)
  - Squamish (1)
  - Straits (2)
  - Twana (5)
- Interior Salish (8)
  - Northern (3)
  - Southern (5)
- Tillamook (1)
  - TILLAMOOK [TIL] (USA)
- Tsamosan (4)
  - Inland (3)
  - Maritime (1)

Salivan (2)
- PIAROA [PID] (Venezuela)
- SÁLIBA [SLC] (Colombia)

Sepik-Ramu (104)
- Gapun (1)
- Leonhard Schultze (6)
  - Papi (2)
  - Walio (4)
- Nor-Pondo (6)
  - Nor (2)
  - Pondo (4)
- Ramu (37)

- Ramu Proper (28)
- Yuat-Waibuk (9)
- Sepik (54)
  - Biksi (3)
  - Middle Sepik (20)
  - Ram (3)
  - Sepik Hill (14)
  - Tama (5)
  - Upper Sepik (6)
  - Yellow River (3)

Sign language (2)
- MONASTIC SIGN LANGUAGE [MZG] (Vatican State)
- PLAINS INDIAN SIGN LANGUAGE [PSD] (USA)

Sino-Tibetan (365)
- Chinese (14)
  - CHINESE, MIN DONG [CDO] (China)
  - CHINESE, MIN NAN [CFR] (China)
  - CHINESE, MANDARIN [CHN] (China)
  - CHINESE, JINYU [CJY] (China)
  - CHINESE, PU-XIAN [CPX] (China)
  - CHINESE, HUIZHOU [CZH] (China)
  - CHINESE, MIN ZHONG [CZO] (China)
  - DUNGAN [DNG] (Kyrgyzstan)
  - CHINESE, HAKKA [HAK] (China)
  - CHINESE, XIANG [HSN] (China)
  - CHINESE, GAN [KNN] (China)
  - CHINESE, MIN BEI [MNP] (China)
  - CHINESE, WU [WUU] (China)
  - CHINESE, YUE [YUH] (China)
- Tibeto-Burman (351)
  - Bai (1)
  - Himalayish (144)
  - Jingpho-Konyak-Bodo (24)
  - Karen (20)
  - Kuki-Chin-Naga (70)
  - Lolo-Burmese (49)
  - Meithei (1)
  - Mikir (1)
  - Mru (1)
  - North Assam (10)
  - Nungish (5)
  - Tangut-Qiang (15)
  - Tujia (2)
  - Unclassified (8)

Siouan (17)
- Catawba (1)
  - CATAWBA [CHC] (USA)
- sioux Proper (16)
  - Central (11)
  - Missouri Valley (2)
  - Southeastern (3)

Sko (7)
- Krisa (4)
  - KRISA [KRO] (Papua New Guinea)
  - PUARI [PUX] (Papua New Guinea)
  - RAWO [RWA] (Papua New Guinea)
  - WARAPU [WRA] (Papua New Guinea)
- Vanimo (3)
  - SKOU [SKV] (Indonesia (Irian Jaya))
  - VANIMO [VAM] (Papua New Guinea)
  - WUTUNG [WUT] (Papua New Guinea)

South Caucasian (5)
- Georgian (2)
  - GEORGIAN [GEO] (Georgia)
  - JUDEO-GEORGIAN [JGE] (Israel)
- Svan (1)
  - SVAN [SVA] (Georgia)
- Zan (2)
  - LAZ [LZZ] (Turkey (Asia))
  - MINGRELIAN [XMF] (Georgia)

Subtiaba-Tlapanec (4)
- SUBTIABA [SUT] (Nicaragua)
- TLAPANECO, ACATEPEC [TPX] (Mexico)
- TLAPANECO, AZOYÚ [TPC] (Mexico)
- TLAPANECO, MALINALTEPEC [TLL] (Mexico)

Tacanan (6)
- Araona-Tacana (5)
  - Araona (1)
  - Cavinena-Tacana (4)
- Tiatinagua (1)
  - ESE EJJA [ESE] (Bolivia)

Tai-Kadai (70)
- Hlai (2)
  - JIAMAO [JIO] (China)
  - HLAI [LIC] (China)
- Kadai (9)
  - Bu-Rong (1)
  - Ge-Chi (3)
  - Yang-Biao (5)
- Kam-Tai (59)
  - Be-Tai (49)
  - Kam-Sui (9)
  - Lakkja (1)

Torricelli (48)
- Kombio-Arapesh (9)
  - Arapesh (3)
  - Kombio (6)
- Maimai (6)
  - Beli (1)
  - Laeko-Libuat (1)
  - Maimai Proper (3)
  - Wiaki (1)
- Marienberg (7)
  - BUNGAIN [BUT] (Papua New Guinea)
  - BUNA [BVN] (Papua New Guinea)
  - ELEPI [ELE] (Papua New Guinea)
  - KAMASAU [KMS] (Papua New Guinea)
  - MUNIWARA [MWB] (Papua New Guinea)
  - MANDI [TUA] (Papua New Guinea)
  - URIMO [URX] (Papua New Guinea)
- Monumbo (2)
  - LILAU [LLL] (Papua New Guinea)
  - MONUMBO [MXK] (Papua New Guinea)
- Urim (1)
  - URIM [URI] (Papua New Guinea)
- Wapei-Palei (20)
  - Palei (7)
  - Urat (1)
  - Wapei (12)
- West Wapei (3)
  - AUNALEI [AUN] (Papua New Guinea)
  - SETI [SBI] (Papua New Guinea)
  - SETA [STF] (Papua New Guinea)

Totonacan (11)
- Tepehua (3)
  - TEPEHUA, HUEHUETLA [TEE] (Mexico)
  - TEPEHUA, PISA FLORES [TPP] (Mexico)

- TEPEHUA, TLACHICHILCO [TPT] (Mexico)
- Totonac (8)
  - TOTONACA, YECUATLA [TLC] (Mexico)
  - TOTONACA, FILOMENO MATA-COAHUITLÁN [TLP] (Mexico)
  - TOTONACA, COYUTLA [TOC] (Mexico)
  - TOTONACA, XICOTEPEC DE JUÁREZ [TOO] (Mexico)
  - TOTONACA, PAPANTLA [TOP] (Mexico)
  - TOTONACA, SIERRA [TOS] (Mexico)
  - TOTONACA, PATLA-CHICONTLA [TOT] (Mexico)
  - TOTONACA, OZUMATLÁN [TQT] (Mexico)

Trans-New Guinea (552)
- Eleman (7)
  - Eleman Proper (5)
  - Purari (1)
  - Tate (1)
- Inland Gulf (4)
  - Ipiko (1)
  - Minanibai (3)
- Kaure (4)
  - Kaure Proper (3)
  - KAPORI [KHP] (Indonesia (Irian Jaya))
- Kolopom (3)
  - KIMAAMA [KIG] (Indonesia (Irian Jaya))
  - NDOM [NQM] (Indonesia (Irian Jaya))
  - RIANTANA [RAN] (Indonesia (Irian Jaya))
- Madang-Adelbert Range (102)
  - Adelbert Range (44)
  - Madang (58)
- Main Section (308)
  - Central and Western (260)
  - East New Guinea Highlands (2)
  - Eastern (46)
- Molof (1)
  - MOLOF [MSL] (Indonesia (Irian Jaya))
- Morwap (1)
  - ELSENG [MRF] (Indonesia (Irian Jaya))
- Nimboran (5)
  - GRESI [GRS] (Indonesia (Irian Jaya))
  - MLAP [KJA] (Indonesia (Irian Jaya))
  - KEMTUIK [KMT] (Indonesia (Irian Jaya))
  - MEKWEI [MSF] (Indonesia (Irian Jaya))
  - NIMBORAN [NIR] (Indonesia (Irian Jaya))
- Northern (27)
  - Border (15)
  - Tor (12)
- Oksapmin (1)
  - OKSAPMIN [OPM] (Papua New Guinea)
- Pauwasi (4)
  - Eastern (2)
  - Western (2)
- Senagi (2)
  - ANGOR [AGG] (Papua New Guinea)
  - KAMBERATARO [KBV] (Indonesia (Irian Jaya))
- South Bird's Head-Timor-Alor-Pantar (32)
  - South Bird's Head (10)
  - Timor-Alor-Pantar (22)
- Teberan-Pawaian (3)
  - Pawaian (1)
  - Teberan (2)
- Tofanma (1)
  - TOFANMA [TLG] (Indonesia (Irian Jaya))
- Trans-Fly-Bulaka River (36)
  - Bulaka River (2)
  - Trans-Fly (34)
- Turama-Kikorian (3)
  - Kairi (1)
  - Turama-Omatian (2)
- Usku (1)
  - USKU [ULF] (Indonesia (Irian Jaya))
- Mek (7)
  - Eastern (3)
  - Western (4)

Tucanoan (25)
- Central Tucanoan (1)
  - CUBEO [CUB] (Colombia)
- Eastern Tucanoan (15)
  - Central (10)
  - Northern (4)
  - Unclassified (1)
- Miriti (1)
  - MIRITI [MMV] (Brazil)
- Western Tucanoan (8)
  - Northern (6)
  - Southern (1)
  - Tanimuca (1)

Tupi (70)
- Arikem (1)
  - KARITIÂNA [KTN] (Brazil)
- Aweti (1)
  - AWETÍ [AWE] (Brazil)
- Mawe-Satere (1)
  - SATERÉ-MAWÉ [MAV] (Brazil)
- Monde (6)
  - ARUÁ [ARX] (Brazil)
  - CINTA LARGA [CIN] (Brazil)
  - GAVIÃO DO JIPARANÁ [GVO] (Brazil)
  - MONDÉ [MND] (Brazil)
  - SURUÍ [SRU] (Brazil)
  - MEKEM [XME] (Brazil)
- Munduruku (2)
  - KURUÁYA [KYR] (Brazil)
  - MUNDURUKÚ [MYU] (Brazil)
- Purubora (1)
  - PURUBORÁ [PUR] (Brazil)
- Ramarama (2)
  - ARÁRA, RONDÔNIA [ARR] (Brazil)
  - ITOGAPÚK [ITG] (Brazil)
- Tupari (4)
  - KANOÉ [KXO] (Brazil)
  - MAKURÁP [MAG] (Brazil)
  - TUPARÍ [TUP] (Brazil)
  - WAYORÓ [WYR] (Brazil)
- tupi-Guarani (49)
  - Guarani (I) (10)
  - Guarayu-Siriono-Jora (II) (3)
  - Kamayura (VII) (1)
  - Kawahib (VI) (9)
  - Kayabi-Arawete (V) (3)
  - Oyampi (VIII) (8)

- Pauserna (1)
- Tenetehara (IV) (8)
- tupi (III) (6)

Unclassified (1)
- YUQUI [YUQ] (Bolivia)

Yuruna (2)
- JURÚNA [JUR] (Brazil)
- MARITSAUÁ [MSP] (Brazil)

Unclassified (96)
- AARIYA [AAR] (India)
- ABISHIRA [ASH] (Peru)
- AGAVOTAGUERRA [AVO] (Brazil)
- AGUANO [AGA] (Peru)
- AMERAX [AEX] (USA)
- AMIKOANA [AKN] (Brazil)
- ANDH [ANR] (India)
- ARÁRA, MATO GROSSO [AXG] (Brazil)
- BEOTHUK [BUE] (Canada)
- BETAF [BFE] (Indonesia (Irian Jaya))
- BETE [BYF] (Nigeria)
- BHATOLA [BTL] (India)
- BUNG [BQD] (Cameroon)
- CAGUA [CBH] (Colombia)
- CALLAWALLA [CAW] (Bolivia)
- CANDOSHI-SHAPRA [CBU] (Peru)
- CANICHANA [CAZ] (Bolivia)
- CARABAYO [CBY] (Colombia)
- CENTÚÚM [CET] (Nigeria)
- CHAK [CKH] (Myanmar)
- CHIPIAJES [CBE] (Colombia)
- CHOLON [CHT] (Peru)
- COXIMA [KOX] (Colombia)
- DOSO [DOL] (Papua New Guinea)
- GAIL [GIC] (South Africa)
- HAITIAN VODOUN CULTURE LANGUAGE [HVC] (Haiti)
- HIBITO [HIB] (Peru)
- HIMARIMÃ [HIR] (Brazil)
- HWLA [HWL] (Togo)
- IAPAMA [IAP] (Brazil)
- IMERAGUEN [IME] (Mauritania)
- KAIMBÉ [QKQ] (Brazil)
- KAMBA [QKZ] (Brazil)
- KAMBIWÁ [QKH] (Brazil)
- KAPINAWÁ [QKP] (Brazil)
- KARA [KAH] (Central African Republic)
- KARAHAWYANA [XKH] (Brazil)
- KARIPÚNA [KGM] (Brazil)
- KARIRI-XOCÓ [KZW] (Brazil)
- KEHU [KHH] (Indonesia (Irian Jaya))
- KEMBRA [XKW] (Indonesia (Irian Jaya))
- KIRIRÍ-XOKÓ [XOO] (Brazil)
- KOHOROXITARI [KOB] (Brazil)
- KORUBO [QKF] (Brazil)
- KUJARGE [VKJ] (Chad)
- KUNZA [KUZ] (Chile)
- KWAVI [CKG] (Tanzania)
- LAAL [GDM] (Chad)
- LECO [LEC] (Bolivia)
- LENCA [LEN] (Honduras)
- LEPKI [LPE] (Indonesia (Irian Jaya))
- LUFU [LDQ] (Nigeria)
- LUO [LUW] (Cameroon)
- MAJHWAR [MMJ] (India)
- MALAKHEL [MLD] (Afghanistan)
- MAWA [WMA] (Nigeria)
- MIARRÃ [XMI] (Brazil)
- MOLENGUE [BXC] (Equatorial Guinea)
- MONIMBO [MOL] (Nicaragua)
- MOVIMA [MZP] (Bolivia)
- MUKHA-DORA [MMK] (India)
- MUNICHE [MYR] (Peru)
- MURKIM [RMH] (Indonesia (Irian Jaya))
- MUTÚS [MUF] (Venezuela)
- NATAGAIMAS [NTS] (Colombia)
- PANKARARÉ [PAX] (Brazil)
- PAPAVÔ [PPV] (Brazil)
- PATAXÓ-HÃHAÃI [PTH] (Brazil)
- PIJAO [PIJ] (Colombia)
- POLARI [PLD] (United Kingdom)
- PUQUINA [PUQ] (Peru)
- QUINQUI [QUQ] (Spain)
- RER BARE [RER] (Ethiopia)
- SAKIRABIÁ [SKF] (Brazil)
- SHOBANG [SSB] (India)
- TAPEBA [TBB] (Brazil)
- TAUSHIRO [TRR] (Peru)
- TINGUI-BOTO [TGV] (Brazil)
- TRAVELLER SCOTTISH [TRL] (United Kingdom)
- TREMEMBÉ [TME] (Brazil)
- TRUKÁ [TKA] (Brazil)
- UAMUÉ [UAM] (Brazil)
- URARINA [URA] (Peru)
- URU-PA-IN [URP] (Brazil)
- WAKONÁ [WAF] (Brazil)
- WAORANI [AUC] (Ecuador)
- WARDUJI [WRD] (Afghanistan)
- WASU [WSU] (Brazil)
- WAXIANGHUA [WXA] (China)
- WEYTO [WOY] (Ethiopia)
- XINCA [XIN] (Guatemala)
- YARÍ [YRI] (Colombia)
- YARURO [YAE] (Venezuela)
- YAUMA [YAX] (Angola)
- YENI [YEI] (Cameroon)
- YUWANA [YAU] (Venezuela)

Uralic (38)

Finno-Ugric (32)
- Finno-Permic (29)
- Ugric (3)

Samoyedic (6)
- Northern Samoyedic (3)
- Southern Samoyedic (3)

Uru-Chipaya (2)
- CHIPAYA [CAP] (Bolivia)
- URU [URE] (Bolivia)

Uto-Aztecan (62)

Northern uto-aztèque (13)
- Hopi (1)
- Numic (7)
- Takic (4)
- Tubatulabal (1)

Southern uto-aztèque (49)
- Aztecan (29)
- Sonoran (20)

Wakashan (5)

Northern (3)
- HAISLA [HAS] (Canada)
- HEILTSUK [HEI] (Canada)
- KWAKIUTL [KWK] (Canada)

Southern (2)
- MAKAH [MYH] (USA)
- NOOTKA [NOO] (Canada)

West Papuan (26)
- Bird's Head (8)
  - North-Central Bird's Head (3)
  - West Bird's Head (5)
- Hattam (1)
  - HATAM [HAD] (Indonesia (Irian Jaya))
- Kebar (1)
  - MPUR [AKC] (Indonesia (Irian Jaya))
- North Halmahera (16)
  - North (14)
  - South (2)

Witotoan (6)
- Boran (2)
  - MUINANE [BMR] (Colombia)
  - BORA [BOA] (Peru)
- Witoto (4)
  - Ocaina (1)
  - Witoto Proper (3)

Yanomam (4)
- NINAM [SHB] (Brazil)
- SANUMÁ [SAM] (Brazil)
- YANOMÁMI [WCA] (Brazil)
- YANOMAMÖ [GUU] (Venezuela)

Yenisei Ostyak (2)
- KET [KET] (Russia (Asia))
- YUGH [YUU] (Russia (Asia))

Yukaghir (2)
- YUKAGHIR, NORTHERN [YKG] (Russia (Asia))
- YUKAGHIR, SOUTHERN [YUX] (Russia (Asia))

Yuki (2)
- WAPPO [WAO] (USA)
- YUKI [YUK] (USA)

Zamucoan (2)
- AYOREO [AYO] (Paraguay)
- CHAMACOCO [CEG] (Paraguay)

Zaparoan (7)
- ANDOA [ANB] (Peru)
- ARABELA [ARL] (Peru)
- AUSHIRI [AUS] (Peru)
- CAHUARANO [CAH] (Peru)
- IQUITO [IQU] (Peru)
- OMURANO [OMU] (Peru)
- ZÁPARO [ZRO] (Ecuador)

A.6 TABLEAU LANGUES - FAMILLES - NOMBRE DE LOCUTEURS - CODES ETHNOLOGUE

Le tableau ci-dessous a été construit à partir des données du site Ethnologue. Il contient, pour huit cent deux langues, des informations pouvant être utilisées dans le choix de langues à informatiser (voir le chapitre III.3.1). La norme 639-2 fournissant une liste de langues (voir cette liste en annexe A.2) destinée à des travaux de terminologie et de bibliographie, nous souhaitions l'utiliser à cette fin, les langues en question offrant potentiellement une littérature, des dictionnaires et une grammaire. Cependant, la liste de la norme 639-2 ne correspondait pas directement au besoin, ne contenant que 470 codes correspondant souvent à des langues mortes ou à des familles de langues.

La liste que nous proposons provient néanmoins de la norme 639-2. Après en avoir retiré les langues mortes, nous avons utilisé le lien proposé sur le site Ethnologue entre la norme 639-2 et les langues qui sont stockées dans sa base de données. Cela se fait très simplement, en envoyant la requête http://www.ethnologue.com/show_iso639.asp?code=xxx, avec le code 639-2 à la place de xxx. Des explications sont données sur la page de réponse lorsqu'il n'y a pas de correspondance nette, comme c'est le cas pour l'akan (aka), l'albanais (alb), le chinois (chi) ou le norvégien (nor). Lorsqu'un code 639-2 correspond à plusieurs langues dans la base Ethnologue, comme dans le cas des familles de langue, la liste des langues correspondantes est donnée. Les quelques codes 639-2 n'ayant pas de correspondance dans Ethnologue — le limbourgeois (lim), le kabardien (kbd)... — ont été ignorés, de même que les cas indiqués incohérents dans la base (norvégien...). Deux cas suspects n'ont cependant pas entraîné l'exclusion de la langue. Il s'agit de :

- ⇨ la langue *braj* (code 639-2 bra), pour laquelle Ethnologue indique que le nombre de locuteurs qu'il donne (44 000) est probablement faux et propose 17 000 000 à la place, valeur que nous avons prise,
- ⇨ la langue *gade lohar* (code 639-2 raj), qui est indiquée avec cinq cent locuteurs, valeur que nous avons prise malgré le très faible nombre pour une langue de l'Inde (classée sous *rajasthani*).

Quelques précisions sur notre tableau :

- ⇨ le nombre de locuteurs est celui des locuteurs de naissance,
- ⇨ lorsqu'une fourchette était donnée, la valeur basse a été retenue,
- ⇨ le pays est le pays d'origine de la langue,
- ⇨ le code Ethnologue identifie la langue dans la base Ethnologue.

Les informations sur une langue sont accessibles sur le site Ethnologue en envoyant la requête http://www.ethnologue.com/show_language.asp?code=xxx, avec le code Ethnologue à la place de xxx.

Nota : Nous nous sommes appuyé sur [Dubois et al. 1994] et sur [Breton 1995] pour traduire en français les noms des familles donnés par Ethnologue en anglais.

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
1	abkhaze	105 000	Géorgie	nord-caucasienne	abk	ABK
2	aceh	3 000 000	Indonésie	austronésienne	ace	ATJ
3	achi, cubulco	45 000	Guatemala	maya	myn	ACC
4	achi, rabinal	37 300	Guatemala	maya	myn	ACR
5	acoli	773 800	Ouganda	nilo-saharienne	ach	ACO
6	adangme	825 900	Ghana	nigéro-congolaise	ada	DGM
7	adygh	300 000	Russie	nord-caucasienne	ady	ADY
8	afar	1 579 000	Ethiopie	afro-asiatique	aar	AFR
9	afrikaans	6 381 000	Afrique du Sud	indo-européenne	afr	AFK
10	agariya	55 757	Inde	austro-asiatique	mun	AGI
11	aguacateco	18 000	Guatemala	maya	myn	AGU
12	akan (=fanti [fat] + twi [twi])	7 000 000	Ghana	nigéro-congolaise	aka	TWS
13	albanais (gheg)	2 000 000	Yougoslavie	indo-européenne	alb/sqi	ALS
14	albanais (tosk)	3 000 000	Albanie	indo-européenne	alb/sqi	ALN
15	aléoute	305	USA	esquimo-aléoute	ale	ALW

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
16	allemand	100 000 000	Allemagne	indo-européenne	ger/deu	GER
17	amharique	17 413 000	Ethiopie	afro-asiatique	amh	AMH
18	anglais	341 000 000	UK	indo-européenne	eng	ENG
19	apache (jicarilla)	812	USA	na-déné	apa	APJ
20	apache (kiowa)	18	USA	na-déné	apa	APK
21	apache (lipan)	2	USA	na-déné	apa	APL
22	apache (mescalero-chiricahua)	1 800	USA	na-déné	apa	APM
23	apache (occidental)	12 693	USA	na-déné	apa	APW
24	arabe (algérien)	22 400 000	Algérie	afro-asiatique	sem	ARQ
25	arabe (Bahrain)	310 000	Bahrain	afro-asiatique	sem	AFH
26	arabe (bedawi)	1 610 000	Egypte	afro-asiatique	sem	AVL
27	arabe (chypriote)	1 300	Chypre	afro-asiatique	sem	ACY
28	arabe (dhofari)	70 000	Oman	afro-asiatique	sem	ADF
29	arabe (égyptien)	46 306 000	Egypte	afro-asiatique	sem	ARZ
30	arabe (golfe)	2 440 000	Irak	afro-asiatique	sem	AFB
31	arabe (hadrami)	410 000	Yemen	afro-asiatique	sem	AYH
32	arabe (haute Egypte)	18 900 000	Egypte	afro-asiatique	sem	AEC
33	arabe (hizagi)	6 000 000	Arabie saoudite	afro-asiatique	sem	ACW
34	arabe (lybien)	4 505 000	Libye	afro-asiatique	sem	AYL
35	arabe (marocain)	19 542 000	Maroc	afro-asiatique	sem	ARY
36	arabe (mésopotamien)	13 900 000	Irak	afro-asiatique	sem	ACM
37	arabe (najdi)	9 800 000	Arabie saoudite	afro-asiatique	sem	ARS
38	arabe (nord-levantine)	15 000 000	Syrie	afro-asiatique	sem	APC
39	arabe (nord-mésopotamien)	6 300 000	Irak	afro-asiatique	sem	AYP
40	arabe (Oman)	1 010 000	Oman	afro-asiatique	sem	ACX
41	arabe (ouzbeki)	700	Ouzbekistan	afro-asiatique	sem	AUZ
42	arabe (Sahara algérien)	110 000	Algérie	afro-asiatique	sem	AAO
43	arabe (sanaani)	7 600 000	Yemen	afro-asiatique	sem	AYN
44	arabe (shihhi)	15 000	Emirats arabes unis	afro-asiatique	sem	SSH
45	arabe (soudanais)	16 000 000	Soudan	afro-asiatique	sem	APD
46	arabe (sud-levantine)	6 155 000	Jordanie	afro-asiatique	sem	AJP
47	arabe (tadjiki)	6 000	Tadjikistan	afro-asiatique	sem	ABH
48	arabe (ta'izzi-adeni)	6 840 000	Yemen	afro-asiatique	sem	ACQ
49	arabe (tchadien)	986 200	Tchad	afro-asiatique	sem	SHU
50	arabe (tunisien)	9 308 000	Tunisie	afro-asiatique	sem	AEB
51	aragonais	11 000	Espagne	indo-européenne	arg	APD
52	arapaho	1 038	USA	algique	arp	ARP
53	araucan	440 000	Chili	araucanienne	arn	ARU
54	arawak	2 400	Surinam	arawak	arw	ARW
55	argobba	10 860	Ethiopie	afro-asiatique	sem	AGJ
56	arménien	6 000 000	Armenie	indo-européenne	arm/hye	ARM
57	assamais	15 334 000	Inde	indo-européenne	asm	ASM
58	assyrien neo-araméen	210 000	Irak	afro-asiatique	sem	AII
59	asturien; bable	100 000	Espagne	indo-européenne	ast	AUB
60	asuri	5 819	Inde	austro-asiatique	mun	ASR
61	avar	601 000	Russie	nord-caucasienne	ava	AVR
62	awadhi	20 540 000	Inde	indo-européenne	awa	AWD
63	aymara	2 200 000	Bolivie	aymara	aym	AYM
64	azéri (nord)	7 059 000	Azerbaïdjan	altaïque	aze	AZE
65	azéri (sud)	24 364 000	Azerbaïdjan	altaïque	aze	AZB
66	bachkir	1 000 000	Russie	altaïque	bak	BXK
67	balinais	3 800 000	Indonésie	austronésienne	ban	BZC
68	baloutchi (est)	1 805 000	Pakistan	indo-européenne	bal	BGP
69	baloutchi (ouest)	1 800 000	Pakistan	indo-européenne	bal	BGN
70	baloutchi (sud)	3 400 000	Pakistan	indo-européenne	bal	BCC
71	bambara	2 777 400	Mali	nigéro-congolaise	bam	BRA
72	banda (bambari)	183 000	Centrafrique	nigéro-congolaise	bad	LIY
73	banda (banda)	102 000	Centrafrique	nigéro-congolaise	bad	BPD
74	banda (mbrès)	42 500	Centrafrique	nigéro-congolaise	bad	BQK

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
75	banda (mi-sud)	102 000	Centrafrique	nigéro-congolaise	bad	BJO
76	banda (ndélé)	35 500	Centrafrique	nigéro-congolaise	bad	BFL
77	banda (ouest centre)	7 500	Centrafrique	nigéro-congolaise	bad	BBP
78	banda (sud centre)	153 000	Centrafrique	nigéro-congolaise	bad	LNL
79	banda (togbo-vara)	24 000	Centrafrique	nigéro-congolaise	bad	TOR
80	banda (yangere)	26 500	Centrafrique	nigéro-congolaise	bad	YAJ
81	bas allemand; bas saxon; allemand, bas; saxon, bas	10 000 000	Allemagne	indo-européenne	nds	SXN
82	basa	230 000	Cameroun	nigéro-congolaise	bas	BAA
83	basque	580 000	Espagne	basque	baq/eus	BSQ
84	bas-sorabe	14 000	Allemagne	indo-européenne	dsb	WEE
85	batak (Indonésie) (alas-kluet)	80 000	Indonésie	austronésienne	btk	BTZ
86	batak (Indonésie) (angkola)	750 000	Indonésie	austronésienne	btk	AKB
87	batak (Indonésie) (dairi)	1 200 000	Indonésie	austronésienne	btk	BTD
88	batak (Indonésie) (karo)	600 000	Indonésie	austronésienne	btk	BTX
89	batak (Indonésie) (mandailing)	400 000	Indonésie	austronésienne	btk	BTM
90	batak (Indonésie) (simalungun)	800 000	Indonésie	austronésienne	btk	BTS
91	batak (Indonésie) (toba)	2 000 000	Indonésie	austronésienne	btk	BBC
92	bedja	1 148 000	Soudan	afro-asiatique	bej	BEI
93	bella coola	200	Canada	salish	sal	BEL
94	bemba	2 150 000	Zambie	nigéro-congolaise	bem	BEM
95	bengali	207 000 000	Bangladesh	indo-européenne	ben	BNG
96	bhattiyali	102 252	Inde	indo-européenne	him	BHT
97	bhojpuri	26 254 000	inde	indo-européenne	bho	BHJ
98	bichlamar	128 000	Vanuatu	créole	bis	BCY
99	biélorusse	10 200 000	Biélorussie	indo-européenne	bel	RUW
100	bihari (angika)	725 000	Inde	indo-européenne	bih	ANP
101	bihari (bhojpuri)	26 254 000	Inde	indo-européenne	bih	BHJ
102	bihari (kudmali)	37 000	Inde	indo-européenne	bih	KYW
103	bihari (magahi)	11 362 000	Inde	indo-européenne	bih	MQM
104	bihari (maithili)	24 191 900	Inde	indo-européenne	bih	MKP
105	bihari (panchpargania)	274 000	Inde	indo-européenne	bih	TDB
106	bihari (sadri)	1 965 000	Inde	indo-européenne	bih	SCK
107	bihari (surajpuri)	273 000	Inde	indo-européenne	bih	SJP
108	bijil neo-araméen	10	Israël	afro-asiatique	sem	BJF
109	bijori	2 391	Inde	austro-asiatique	mun	BIX
110	bikol	2 500 000	Philippines	austronésienne	bik	BKL
111	bilaspuri	295 387	Inde	indo-européenne	him	KFS
112	bini	1 000 000	Nigéria	nigéro-congolaise	bin	EDO
113	birhor	10 000	Inde	austro-asiatique	mun	BIY
114	birman	32 000 000	Myanmar	sino-tibétaine	bur/mya	BMS
115	blackfoot	5 800	Canada	algique	bla	BLC
116	blin; bilen	70 000	Erythrée	afro-asiatique	byn	BYN
117	bohtan neo-araméen	1 000	Géorgie	afro-asiatique	sem	BHN
118	bondo	8 000	Inde	austro-asiatique	mun	BFW
119	bouriate	318 000	Russie	altaïque	bua	MNB
120	braj	17 000 000	Inde	indo-européenne	bra	BFS
121	breton	500 000	France	indo-européenne	bre	BRT
122	bugi	3 500 000	Indonésie	austronésienne	bug	BPR
123	bulgare	9 000 000	Bulgarie	indo-européenne	bul	BLG
124	caddo	141	USA	caddo	cad	CAD
125	cakchiquel, central	132 200	Guatemala	maya	myn	CAK
126	cakchiquel, est	100 000	Guatemala	maya	myn	CKE
127	cakchiquel, nord	16 000	Guatemala	maya	myn	CKC
128	cakchiquel, ouest	77 000	Guatemala	maya	myn	CKW
129	cakchiquel, santa Maria de Jesus	15 000	Guatemala	maya	myn	CKI
130	cakchiquel, santo Domingo Xenacoj	5 200	Guatemala	maya	myn	CKJ
131	cakchiquel, sud	43 000	Guatemala	maya	myn	CKF
132	cakchiquel, sud central	43 000	Guatemala	maya	myn	CKD
133	cakchiquel, sud-ouest Yepocapa	8 000	Guatemala	maya	myn	CBM
134	caribe	10 000	Vénézuéla	carib	car	CRB

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
135	catalan; valencien	6 565 000	Espagne	indo-européenne	cat	CLN
136	cebuano	15 000 000	Philippines	austronésienne	ceb	CEB
137	chaldéen neo-araméen	200 000	Irak	afro-asiatique	sem	CLD
138	chambeali	129 654	Inde	indo-européenne	him	CDH
139	chamorro	78 000	Guam	austronésienne	cha	CJD
140	chan	3 000 000	Myanmar	taï-kadaï	shn	SJN
141	chehalis, lower	5	USA	salish	sal	CEA
142	chehalis, upper	2	USA	salish	sal	CJH
143	cherokee	14 000	USA	iroquoise	chr	CER
144	cheyenne	1 721	USA	algique	chy	CHY
145	chichewa; chewa; nyanja	5 622 000	Malawi	nigéro-congolaise	nya	NYJ
146	chicomuceltec	1 600	Mexique	maya	myn	COB
147	chinois (gan)	20 580 000	Chine	sino-tibétaine	chi/zho	KNN
148	chinois (hakka)	33 000 000	Chine	sino-tibétaine	chi/zho	HAK
149	chinois (jinyu)	45 000 000	Chine	sino-tibétaine	chi/zho	CJY
150	chinois (mandarin)	874 000 000	Chine	sino-tibétaine	chi/zho	CHN
151	chinois (min bei)	10 537 000	Chine	sino-tibétaine	chi/zho	MNP
152	chinois (min dong)	247 000	Chine	sino-tibétaine	chi/zho	CDO
153	chinois (min nan)	45 000 000	Chine	sino-tibétaine	chi/zho	CFR
154	chinois (pu-xian)	6 000	Chine	sino-tibétaine	chi/zho	CPX
155	chinois (wu)	77 175 000	Chine	sino-tibétaine	chi/zho	WUU
156	chinois (xiang)	36 015 000	Chine	sino-tibétaine	chi/zho	HSN
157	chinois (yue)	71 000 000	Chine	sino-tibétaine	chi/zho	YUH
158	chinook, jargon	100	Canada	pidgine	chn	CRW
159	chipewyan	4 000	Canada	na-dené	chp	CPW
160	chippewa	103 826	USA	algique	oji	CIW
161	choctaw	17 890	USA	muskogéenne	cho	CCT
162	ch'ol tila	35 000	Mexique	maya	myn	CTI
163	ch'ol tumbala	90 000	Mexique	maya	myn	CTU
164	chontal, tabasco	55 000	Mexique	maya	myn	CHF
165	chorti	31 500	Guatemala	maya	myn	CAA
166	chuj, san Mateo Ixtatan	22 130	Guatemala	maya	myn	CNM
167	chuj, san Sebastian Coatan	18 458	Guatemala	maya	myn	CAC
168	churahi	110 552	Inde	indo-européenne	him	CDJ
169	chuuk	38 341	Micronésie	austronésienne	chk	TRU
170	clallam	5	USA	salish	sal	CLM
171	cœur d'alene	5	USA	salish	sal	CRD
172	columbia-wenatchi	75	USA	salish	sal	COL
173	comox	400	Canada	salish	sal	COO
174	coréen	78 000 000	Corée	langue isolée	kor	KKN
175	cornique	1 000	UK	indo-européenne	cor	CRN
176	corse	341 000	France	indo-européenne	cos	COI
177	cowlitz	2	USA	salish	sal	COW
178	cree (moose)	4 500	Canada	algique	cre	CRM
179	cree (nord-est)	5 308	Canada	algique	cre	CRL
180	cree (plaines)	34 000	Canada	algique	cre	CRP
181	cree (sud-est)	7 306	Canada	algique	cre	CRE
182	cree (swampi)	4 500	Canada	algique	cre	CSW
183	dair	1 000	Soudan	nilo-saharienne	nub	DRB
184	dakota	20 355	USA	sioux	dak	DHG
185	danois	5 326 000	Danemark	indo-européenne	dan	DNS
186	dargwa	371 000	Russie	nord-caucasienne	dar	DAR
187	dayak	250 000	Indonésie	austronésienne	day	NIJ
188	delaware (munsee)	7	Canada	algique	del	UMU
189	delaware (unami)	5	USA	algique	del	DEL
190	dilling	5 295	Soudan	nilo-saharienne	nub	DIL
191	dinka (centre-sud)	250 000	Soudan	nilo-saharienne	din	DIB
192	dinka (nord-est)	320 000	Soudan	nilo-saharienne	din	DIP
193	dinka (nord-ouest)	80 000	Soudan	nilo-saharienne	din	DIW
194	dinka (sud-est)	250 000	Soudan	nilo-saharienne	din	DIN
195	dinka (sud-ouest)	450 000	Soudan	nilo-saharienne	din	DIK
196	dioula	2 520 000	Burkina Faso	nigéro-congolaise	dyu	DYU

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
197	dogri	2 200 000	Inde	indo-européenne	doi	DOJ
198	dogrib	2 085	Canada	na-déné	dgr	DGB
199	dogri-kangri	2 200 000	Inde	indo-européenne	him	DOJ
200	douala	87 700	Cameroun	nigéro-congolaise	dua	DOU
201	dzongkha	160 000	Bouthan	sino-tibétaine	dzo	DZO
202	écossais	100 000	UK	indo-européenne	sco	SCO
203	efik	400 000	Nigéria	nigéro-congolaise	efi	EFK
204	ekajuk	30 000	Nigéria	nigéro-congolaise	eka	EKA
205	el hugeirat	1 000	Soudan	nilo-saharienne	nub	ELH
206	erza	500 000	Russie	ouralienne	myv	MYV
207	esclave (athapascan) (nord)	290	Canada	na-déné	den	SCS
208	esclave (athapascan) (sud)	2 620	Canada	na-déné	den	SLA
209	espagnol; castillan	322 200 000	Espagne	indo-européenne	spa	SPN
210	espéranto	1 000	France	langues artificielles	epo	ESP
211	estonien	1 100 000	Estonie	ouralienne	est	EST
212	éwé	2 477 600	Ghana	nigéro-congolaise	ewe	EWE
213	éwondo	577 700	Cameroun	nigéro-congolaise	ewo	EWO
214	fang	858 000	Gabon	nigéro-congolaise	fan	FNG
215	féroïen	47 000	Danemark	indo-européenne	fao	FAE
216	fidjien	350 000	Fidji	austronésienne	fij	FJI
217	finnois	6 000 000	Finlande	ouralienne	fin	FIN
218	fon	1 436 000	Bénin	nigéro-congolaise	fon	FOA
219	français	77 000 000	France	indo-européenne	fre/fra	FRN
220	frioulan	600 000	Italie	indo-européenne	fur	FRL
221	frison	730 000	Pays-Bas	indo-européenne	fry	FRI
222	ga	300 000	Ghana	nigéro-congolaise	gaa	GAC
223	gadaba, bodo	32 500	Inde	austro-asiatique	mun	GBJ
224	gaddi	120 000	Inde	indo-européenne	him	GBK
225	gaélique; gaélique écossais	94 000	UK	indo-européenne	gla	GLS
226	galicien	4 000 000	Espagne	indo-européenne	glg	GLN
227	galla (Borana-Arsi-Guji)	3 786 000	Ethiopie	afro-asiatique	orm	GAX
228	galla (est)	4 526 000	Ethiopie	afro-asiatique	orm	HAE
229	galla (ouest-centre)	8 920 000	Ethiopie	afro-asiatique	orm	GAZ
230	gallois	580 000	UK	indo-européenne	wel/cym	WLS
231	ganda	3 025 000	Ouganda	nigéro-congolaise	lug	LAP
232	gata'	3 055	Inde	austro-asiatique	mun	GAQ
233	gayo	180 000	Indonésie	austronésienne	gay	GYO
234	gbaya (bokoto)	130 000	Centrafrique	nigéro-congolaise	gba	BDT
235	gbaya (bossangoa)	176 000	Centrafrique	nigéro-congolaise	gba	GBP
236	gbaya (bozoum)	35 000	Centrafrique	nigéro-congolaise	gba	GBQ
237	gbaya (ngbaka)	1 005 000	Congo	nigéro-congolaise	gba	NGA
238	gbaya (nord-ouest)	267 000	Centrafrique	nigéro-congolaise	gba	GYA
239	gbaya (sud-ouest)	177 000	Centrafrique	nigéro-congolaise	gba	MDO
240	géorgien	4 103 000	Géorgie	sud-caucasienne	geo/kat	GEO
241	ghulfan	16 000	Soudan	nilo-saharienne	nub	GHL
242	gobu	12 000	Centrafrique	nigéro-congolaise	bad	GOX
243	gond (nord)	2 632 000	Inde	dravidienne	gon	GON
244	gond (sud)	600 000	Inde	dravidienne	gon	GGO
245	gorontalo	900 000	Indonésie	austronésienne	gor	GRL
246	goudjrati	46 100 000	Inde	indo-européenne	guj	GJR
247	grebo (Barclayville)	23 700	Liberia	nigéro-congolaise	grb	GRY
248	grebo (gboloo)	56 300	Liberia	nigéro-congolaise	grb	GEC
249	grebo (glio-oubi)	6 000	Liberia	nigéro-congolaise	grb	OUB
250	grebo (krumen, pye)	20 000	Côte d'Ivoire	nigéro-congolaise	grb	PYE
251	grebo (krumen, yepo)	28 300	Côte d'Ivoire	nigéro-congolaise	grb	TED
252	grebo (nord)	84 500	Liberia	nigéro-congolaise	grb	GRB
253	grebo (sud)	28 700	Liberia	nigéro-congolaise	grb	GRJ
254	grec moderne (après 1453)	12 000 000	Grèce	indo-européenne	gre/ell	GRK
255	groenlandais	47 000	Groënland	esquimo-aléoute	kal	ESG
256	guarani (branche principale)	5 000 000	Paraguay	tupi	grn	GUG
257	gurage est	827 764	Ethiopie	afro-asiatique	sem	GRE
258	gurage ouest	798 202	Ethiopie	afro-asiatique	sem	GUY

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
259	gurage soddo	254 682	Ethiopie	afro-asiatique	sem	GRU
260	gwich´in	700	Canada	na-déné	gwi	KUC
261	haida (nord)	45	Canada	na-déné	hai	HAI
262	haida (sud)	10	Canada	na-déné	hai	HAX
263	haïtien; créole haïtien	7 800 000	Haïti	créole	hat	HAT
264	halkomelem	500	Canada	salish	sal	HUR
265	haoussa	24 200 000	Nigéria	afro-asiatique	hau	HUA
266	harari	21 283	Ethiopie	afro-asiatique	sem	HAR
267	harsusi	1 000	Oman	afro-asiatique	sem	HSS
268	hassaniyya	2 511 000	Mauritanie	afro-asiatique	sem	MEY
269	haut-sorabe	15 000	Allemagne	indo-européenne	hsb	WEN
270	hawaïen	1 000	USA	austronésienne	haw	HWI
271	hébreu	5 150 000	Israël	afro-asiatique	heb	HBR
272	herero	144 000	Namibie	nigéro-congolaise	her	HER
273	hértevin	1 000	Turquie	afro-asiatique	sem	HRT
274	hiligaynon	7 000 000	Philippines	austronésienne	hil	HIL
275	hindi	366 000 000	Inde	indo-européenne	hin	HND
276	hinduri	138	Inde	indo-européenne	him	HII
277	ho	1 077 000	Inde	austro-asiatique	mun	HOC
278	hobyot	100	Oman	afro-asiatique	sem	HOH
279	hongrois	14 500 000	Hongrie	ouralienne	hun	HNG
280	huasteco, san Luis Potosi	70 000	Mexique	maya	myn	HVA
281	huasteco, tantoyuca	50 000	Mexique	maya	myn	HUS
282	hulaula	10 000	Israël	afro-asiatique	sem	HUY
283	hupa	8	USA	na-déné	hup	HUP
284	iakoute	363 000	Russie	altaïque	sah	UKT
285	iban	415 000	Malaisie	austronésienne	iba	IBA
286	igbo	18 000 000	Nigéria	nigéro-congolaise	ibo	IGR
287	ijo	1 770 000	Nigéria	nigéro-congolaise	ijo	IJC
288	ilocano	8 000 000	Philippines	austronésienne	ilo	ILO
289	indonésien	17 050 000	Indonésie	austronésienne	ind	INZ
290	inga	10 000	Colombie	quetchua	que	INB
291	inga, jungle	5 000	Colombie	quetchua	que	INJ
292	ingouche	230 315	Russie	nord-caucasienne	inh	INH
293	inuktitut (est)	14 000	Canada	esquimo-aléoute	iku	ESB
294	inuktitut (ouest)	4 000	Canada	esquimo-aléoute	iku	ESC
295	inupiaq (nord Alaska)	3 500	USA	esquimo-aléoute	ipk	ESI
296	inupiaq (nord-ouest Alaska)	4 000	USA	esquimo-aléoute	ipk	ESK
297	irlandais	260 000	Irlande	indo-européenne	gle	GLI
298	islandais	250 000	Islande	indo-européenne	ice/isl	ICE
299	italien	62 000 000	Italie	indo-européenne	ita	ITN
300	itza	12	Guatemala	maya	myn	ITZ
301	ixil, chajul	18 000	Guatemala	maya	myn	IXJ
302	ixil, nebaj	35 000	Guatemala	maya	myn	IXI
303	ixil, san Juan cotzal	16 000	Guatemala	maya	myn	IXL
304	jacalteco, est	11 000	Guatemala	maya	myn	JAC
305	jacalteco, ouest	77 700	Guatemala	maya	myn	JAI
306	japonais	125 000 000	Japon	japonaise	jpn	JPN
307	jaunsari	97 000	Inde	indo-européenne	him	JNS
308	javanais	75 500 800	Indonésie	austronésienne	jav	JAN
309	jibbali	25 000	Oman	afro-asiatique	sem	SHV
310	juang	40 000	Inde	austro-asiatique	mun	JUN
311	judéo-arabe (judéo-irakien)	105 000	Israël	afro-asiatique	jrb	YHD
312	judéo-arabe (judéo-marocain)	254 000	Israël	afro-asiatique	jrb	AJU
313	judéo-arabe (judéo-tripolitain)	35 000	Israël	afro-asiatique	jrb	YUD
314	judéo-arabe (judéo-tunisien)	45 500	Israël	afro-asiatique	jrb	AJT
315	judéo-arabe (judéo-yéménite)	51 000	Israël	afro-asiatique	jrb	JYE
316	judéo-espagnol	160 000	Israël	indo-européenne	lad	SPJ
317	judéo-persan	60 000	Israël	indo-européenne	jpr	DZH
318	kabardien	443 000	Russie	nord-caucasienne	kbd	KAB
319	kabyle	3 074 000	Algérie	afro-asiatique	kab	KYL
320	kachin	645 000	sino-tibétaine	sino-tibétaine	kac	CGP

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
321	kachoube	3 000	Pologne	indo-européenne	csb	CSB
322	kadaru	7 000	Soudan	nilo-saharienne	nub	KDU
323	kalispel-pend d'oreille	200	USA	salish	sal	FLA
324	kalmouk	518 000	Russie	altaïque	xal	KGZ
325	kamba	2 448 302	Kenya	nigéro-congolaise	kam	KIK
326	kanjobal, est	77 700	Guatemala	maya	myn	KJB
327	kanjobal, ouest	48 500	Guatemala	maya	myn	KNJ
328	kannada	35 346 000	Inde	dravidienne	kan	KJV
329	kanouri	3 500 000	Nigéria	nilo-saharienne	kau	KPH
330	karakalpak	409 000	Ouzbekistan	altaïque	kaa	KAC
331	karatcha balkar	241 000	Russie	altaïque	krc	KRC
332	karen (brek)	16 600	Myanmar	sino-tibétaine	kar	KVL
333	karen (bwe)	15 700	Myanmar	sino-tibétaine	kar	BWE
334	karen (geba)	40 100	Myanmar	sino-tibétaine	kar	KVQ
335	karen (geko)	9 500	Myanmar	sino-tibétaine	kar	GHK
336	karen (manumanaw)	3 000	Myanmar	sino-tibétaine	kar	KXF
337	karen (padaung)	41 000	Myanmar	sino-tibétaine	kar	PDU
338	karen (paku)	5 300	Myanmar	sino-tibétaine	kar	KPP
339	karen (pa'o)	560 600	Myanmar	sino-tibétaine	kar	BLK
340	karen (pwo est)	1 050 000	Myanmar	sino-tibétaine	kar	KJP
341	karen (pwo nord)	60 000	Myanmar	sino-tibétaine	kar	PWW
342	karen (pwo ouest)	210 000	Myanmar	sino-tibétaine	kar	PWO
343	karen (s'gaw)	2 000 000	Myanmar	sino-tibétaine	kar	KSW
344	karen (yinbaw)	7 300	Myanmar	sino-tibétaine	kar	KVU
345	karen (zayein)	9 300	Myanmar	sino-tibétaine	kar	KXK
346	karko	12 986	Soudan	nilo-saharienne	nub	KKO
347	kashmiri	4 511 000	Inde	indo-européenne	kas	KSH
348	kayah (est)	77 900	Myanmar	sino-tibétaine	kar	EKY
349	kayah (ouest)	210 000	Myanmar	sino-tibétaine	kar	KYU
350	kazakh	8 000 000	Kazakhstan	altaïque	kaz	KAZ
351	kekchi	400 000	Guatemala	maya	myn	KEK
352	kenuzi-dongola	180 000	Soudan	nilo-saharienne	nub	KNC
353	kharia	278 500	Inde	austro-asiatique	mun	KHR
354	khasi	950 000	Inde	austro-asiatique	kha	KHI
355	khmer	7 039 200	Cambodge	austro-asiatique	khm	KMR
356	kikuyu	5 347 000	Kenya	nigéro-congolaise	kik	KIU
357	kimbundu	3 000 000	Angola	nigéro-congolaise	kmb	MLO
358	kinnauri	6 331	Inde	indo-européenne	him	KJO
359	kirghize	2 631 420	Kirghizistan	altaïque	kir	KDO
360	kiribati	67 500	Kiribati	austronésienne	gil	GLB
361	kom (permiak)	116 000	Russie	ouralienne	kom	KOI
362	kom (zyrian)	262 200	Russie	ouralienne	kom	KPV
363	kongo	3 217 000	Congo	nigéro-congolaise	kon	KON
364	konkani	4 000 000	Inde	indo-européenne	kok	KNK
365	korku	478 000	Inde	austro-asiatique	mun	KFQ
366	korwa	66 000	Inde	austro-asiatique	mun	KFP
367	kosrae	6 900	Micronésie	austronésienne	kos	KSI
368	koumyk	282 500	Russie	altaïque	kum	KSK
369	koy sanjaq surat	800	Irak	afro-asiatique	sem	KQD
370	kpagua	3 000	Centrafrique	nigéro-congolaise	bad	KUW
371	kpellé (Guinée)	308 000	Guinée	nigéro-congolaise	kpe	GKP
372	kpellé (Libéria)	487 400	Libéria	nigéro-congolaise	kpe	KPE
373	kuanyama; kwanyama	421 000	Angola	nigéro-congolaise	kua	KUY
374	kullu pahari	109 000	Inde	indo-européenne	him	KFX
375	kurde (kurdi)	6 036 000	Irak	indo-européenne	kur	KDB
376	kurde (kurmanji)	7 000 000	Turquie	indo-européenne	kur	KUR
377	kurukh	2 053 000	Inde	dravidienne	kru	KVN
378	kutenai	222	Canada	langue isolée	kut	KUN
379	lacandon	700	Mexique	maya	myn	LAC
380	lahnda	30 000 000	Pakistan	indo-européenne	lah	PNB
381	lamba	211 000	Zambie	nigéro-congolaise	lam	LAB
382	langbashe	43 000	Centrafrique	nigéro-congolaise	bad	LNA

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
383	laotien	3 188 000	Laos	taï-kadaï	lao	NOL
384	letton	1 500 000	Lettonie	indo-européenne	lav	LAT
385	lezghien	451 000	Russie	nord-caucasienne	lez	LEZ
386	lillooet	300	Canada	salish	sal	LIL
387	lingala	309 100	Congo	nigéro-congolaise	lin	LIN
388	lishan didan	3 050	Israël	afro-asiatique	sem	TRG
389	lishana deni	7 000	Israël	afro-asiatique	sem	LSD
390	lishanid noshan	2 000	Israël	afro-asiatique	sem	AIJ
391	lituanien	4 000 000	Lituanie	indo-européenne	lit	LIT
392	lodhi	75 000	Inde	austro-asiatique	mun	LBM
393	lozi	557 000	Zambie	nigéro-congolaise	loz	LOZ
394	luba-katanga	1 505 000	Congo	nigéro-congolaise	lub	LUH
395	luba-lulua	6 300 000	Congo	nigéro-congolaise	lua	LUB
396	luiseno	43	Etats-Unis	uto-aztèque	lui	LUI
397	lunda	310 000	Zambie	nigéro-congolaise	lun	LVN
398	luo (Kenya et Tanzanie)	3 408 000	Kenya et Tanzanie	nilo-saharienne	luo	LUO
399	lushai	542 500	Inde	sino-tibétaine	lus	LSH
400	lushootseed	60	USA	salish	sal	LUT
401	luxembourgeois	300 000	Luxembourg	indo-européenne	ltz	LUX
402	macédonien	2 000 000	Macédoine	indo-européenne	mac/mkd	MKJ
403	madourais	13 694 000	Indonésie	austronésienne	mad	MHJ
404	magahi	11 362 000	Inde	indo-européenne	mag	MQM
405	mahali	66 000	Inde	austro-asiatique	mun	MJX
406	mahasu pahari	500 000	Inde	indo-européenne	him	BFZ
407	maithili	24 191 900	Inde	indo-européenne	mai	MKP
408	makassar	1 600 000	Indonésie	austronésienne	mak	MSR
409	malais	18 000 000	Malaisie	austronésienne	may/msa	MLI
410	malayalam	35 706 000	Inde	dravidienne	mal	MJS
411	maldivien	220 000	Maldives	indo-européenne	div	SNM
412	malgache	9 398 700	Madagascar	austronésienne	mlg	MEX
413	maltais	330 000	Malte	afro-asiatique	mlt	MLS
414	mam, centre	100 000	Guatemala	maya	myn	MVC
415	mam, nord	180 000	Guatemala	maya	myn	MAM
416	mam, sud	125 000	Guatemala	maya	myn	MMS
417	mam, tajumulco	35 000	Guatemala	maya	myn	MPF
418	mam, todos santos cuchumatan	50 000	Guatemala	maya	myn	MVJ
419	mandaic	800	Iran	afro-asiatique	sem	MID
420	mandar	200 000	Indonésie	austronésienne	mdr	MHN
421	mandchou	20	Chine	altaïque	mnc	MJF
422	mandeali	776 372	Inde	indo-européenne	him	MJL
423	mandingue	1 178 500	Sénégal	nigéro-congolaise	man	MNK
424	manipuri	1 648 000	Inde	sino-tibétaine	mni	MNR
425	manx; mannois	77 000	UK	indo-européenne	glv	MJD
426	maori	50 000	Nouvelle Zélande	austronésienne	mao/mri	MBF
427	marathe	68 022 000	Inde	indo-européenne	mar	MRT
428	mari	534 000	Russie	ouralienne	chm	MAL
429	marshall	43 900	Iles Marshall	austronésienne	mah	MZM
430	marvari	12 963 000	Inde	indo-européenne	mwr	MKD
431	massaï	883 000	Kenya	nilo-saharienne	mas	MET
432	maya, chan santa Cruz	40 000	Mexique	maya	myn	YUS
433	maya, yucatan	700 000	Mexique	maya	myn	YUA
434	mazahua central	350 000	Mexique	otomangue	oto	MAZ
435	mazahua, michoacan	15 000	Mexique	otomangue	oto	QMN
436	mbandja	200 000	Centrafrique	nigéro-congolaise	bad	ZMZ
437	mehri	100 000	Yemen	afro-asiatique	sem	MHR
438	mendé	1 480 000	Sierra Leone	nigéro-congolaise	men	MFY
439	micmac	8 500	Canada	algique	mic	MIC
440	midob	50 000	Soudan	nilo-saharienne	nub	MEI
441	minangkabau	6 500 000	Indonésie	austronésienne	min	MPU
442	mocho	168	Mexique	maya	myn	MHC

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
443	mohawk	2 000	Canada	iroquoise	moh	MOH
444	moksa	428 333	Russie	ouralienne	mdf	MDF
445	moldave	26 000 000	Roumanie	indo-européenne	mol	RUM
446	mongo	400 000	Congo	nigéro-congolaise	lol	MOM
447	mongol (halh)	2 330 000	Mongolie	altaïque	mon	KHK
448	mongol (périphérique)	3 381 000	Chine	altaïque	mon	MVF
449	mono	65 000	Centrafrique	nigéro-congolaise	bad	MNH
450	mopan maya	7 000	Belize	maya	myn	MOP
451	moré	5 050 000	Burkina Faso	nigéro-congolaise	mos	MHM
452	mundari	2 069 000	Inde	austro-asiatique	mun	MUW
453	muskogee	6 213	USA	muskogéenne	mus	CRK
454	nahualt, central	40 000	Mexique	uto-aztèque	nah	NHN
455	nahualt, coapetec	1 400	Mexique	uto-aztèque	nah	NAZ
456	nahualt, durango	1 000	Mexique	uto-aztèque	nah	NLN
457	nahualt, guerrego	150 000	Mexique	uto-aztèque	nah	NAH
458	nahualt, huasteca, est	410 000	Mexique	uto-aztèque	nah	NAI
459	nahualt, huasteco, ouest	400 000	Mexique	uto-aztèque	nah	NHW
460	nahualt, huaxcaleca	7 000	Mexique	uto-aztèque	nah	NHQ
461	nahualt, istmo-cosoleacaque	5 144	Mexique	uto-aztèque	nah	NHK
462	nahualt, istmo-mecayapan	20 000	Mexique	uto-aztèque	nah	NAU
463	nahualt, istmo-pajapan	7 000	Mexique	uto-aztèque	nah	NHP
464	nahualt, ixhuatlancillo	4 000	Mexique	uto-aztèque	nah	NHX
465	nahualt, michoacan	3 000	Mexique	uto-aztèque	nah	NCL
466	nahualt, morelos	15 000	Mexique	uto-aztèque	nah	NHM
467	nahualt, oaxaca nord	9 000	Mexique	uto-aztèque	nah	NHY
468	nahualt, ometepec	433	Mexique	uto-aztèque	nah	NHT
469	nahualt, orizaba	120 000	Mexique	uto-aztèque	nah	NLV
470	nahualt, puebla central	16 000	Mexique	uto-aztèque	nah	NCX
471	nahualt, puebla nord	60 000	Mexique	uto-aztèque	nah	NCJ
472	nahualt, puebla sierra	125 000	Mexique	uto-aztèque	nah	AZZ
473	nahualt, puebla sud-est	130 000	Mexique	uto-aztèque	nah	NHS
474	nahualt, santa Maria la alta	2 000	Mexique	uto-aztèque	nah	NHZ
475	nahualt, temascaltepec	311	Mexique	uto-aztèque	nah	NHV
476	nahualt, tenango	1 500	Mexique	uto-aztèque	nah	NHI
477	nahualt, tetelcingo	3 500	Mexique	uto-aztèque	nah	NHG
478	nahualt, tlalitzlipa	108	Mexique	uto-aztèque	nah	NHJ
479	nahualt, tlamacazapa	1 548	Mexique	uto-aztèque	nah	NUZ
480	napolitain	7 047 399	Italie	indo-européenne	nap	NPL
481	nauruan	6 000	Nauru	austronésienne	nau	NRU
482	navaho	148 530	USA	na-déné	nav	NAV
483	ndébélé du nord	1 502 000	Zimbabwe	nigéro-congolaise	nde	NDF
484	ndébélé du sud	588 000	Afrique du Sud	nigéro-congolaise	nbl	NEL
485	ndonga	240 000	Namibie	nigéro-congolaise	ndo	NDG
486	néerlandais; flamand	20 000 000	Pays-Bas	indo-européenne	dut/nld	DUT
487	neo-araméen occidental	15 000	Syrie	afro-asiatique	sem	AMW
488	népalais	16 056 000	Népal	indo-européenne	nep	NEP
489	newari	690 000	Népal	sino-tibétaine	new	NEW
490	ngbundu	16 000	Centrafrique	nigéro-congolaise	bad	NUU
491	nias	480 000	Indonésie	austronésienne	nia	NIP
492	niué	8 000	Niue	austronésienne	niu	NIQ
493	nobiin	295 000	Soudan	nilo-saharienne	nub	FIA
494	nogaï; nogay	68 000	Russie	altaïque	nog	NOG
495	nooksack	350	USA	salish	sal	NOK
496	norvégien bokmål; bokmål, norvégien	5 000 000	Norvège	indo-européenne	nob	NRR
497	norvégien nynorsk; nynorsk, norvégien	?	Norvège	indo-européenne	nno	NRN
498	nyamwezi	926 000	Tanzanie	nigéro-congolaise	nym	NYZ
499	nyankolé	1 643 193	Ouganda	nigéro-congolaise	nyn	NYN
500	nyoro	495 443	Ouganda	nigéro-congolaise	nyo	NYR
501	nzema	352 500	Ghana	nigéro-congolaise	nzi	NZE

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
502	occitan (après 1500); provençal (auvergnat)	?	France	indo-européenne	oci	AUV
503	occitan (après 1500); provençal (gascon)	254 800	France	indo-européenne	oci	GSC
504	occitan (après 1500); provençal (languedocien)	?	France	indo-européenne	oci	LNC
505	occitan (après 1500); provençal (limousin)	?	France	indo-européenne	oci	LMS
506	occitan (après 1500); provençal (provençal)	354 500	France	indo-européenne	oci	PRV
507	occitan (après 1500); provençal (shuadit)	?	France	indo-européenne	oci	SDT
508	ojibwa, ouest	35 000	Canada	algique	oji	OJI
509	ojibwa, severn	8 000	Canada	algique	oji	OJS
510	okanagan	500	Canada	salish	sal	OKA
511	oriya	32 000 000	Inde	indo-européenne	ori	ORY
512	osage	5	USA	sioux	osa	OSA
513	ossète	593 000	Géorgie	indo-européenne	oss	OSE
514	otomi, état de Mexico	10 000	Mexique	otomangue	oto	OTS
515	otomi, ixtenco	736	Mexique	otomangue	oto	OTA
516	otomi, mezquital	100 000	Mexique	otomangue	oto	OTE
517	otomi, nord ouest	33 000	Mexique	otomangue	oto	OTQ
518	otomi, sierra oriental	20 000	Mexique	otomangue	oto	OTM
519	otomi, temoaya	37 000	Mexique	otomangue	oto	OTT
520	otomi, tenango	10 000	Mexique	otomangue	oto	OTN
521	otomi, texcatepec	12 000	Mexique	otomangue	oto	OTX
522	otomi, tilapa	400	Mexique	otomangue	oto	OTL
523	ottawa	8 000	Canada	algique	oji	OTW
524	ouïgour	7 595 512	Chine	altaïque	uig	UIG
525	ourdou	60 290 000	Pakistan	indo-européenne	urd	URD
526	ouszbek	18 466 000	Ouzbekistan	altaïque	uzb	UZB
527	pachto (nord)	9 685 000	Pakistan	indo-européenne	pus	PBU
528	pachto (sud)	9 204 000	Afghanistan	indo-européenne	pus	PBT
529	palau	15 000	Palau	austronésienne	pau	PLU
530	pampangan	1 897 378	Philippines	austronésienne	pam	PMP
531	pangasinan	1 164 586	Philippines	austronésienne	pag	PNG
532	pangwali	17 000	Inde	indo-européenne	him	PGG
533	papiamento	329 000	Antilles Néerlandaises	créole	pap	PAE
534	parenga	4 281	Inde	austro-asiatique	mun	PCJ
535	penjabi (est)	27 125 000	Inde	indo-européenne	pan	PNJ
536	penjabi (ouest)	30 000 000	Pakistan	indo-européenne	pan	PNB
537	persan (est)	7 000 000	Afghanistan	indo-européenne	per/fas	PRS
538	persan (ouest)	24 280 000	Iran	indo-européenne	per/fas	PES
539	peul (adamawa)	760 000	Cameroun	nigéro-congolaise	ful	FUB
540	peul (bagirmi)	180 000	Tchad	nigéro-congolaise	ful	FUI
541	peul (Bénin-Togo)	328 000	Bénin-Togo	nigéro-congolaise	ful	FUE
542	peul (centre-est Niger)	450 000	Niger	nigéro-congolaise	ful	FUQ
543	peul (fuuta jalon)	2 900 000	Guinée	nigéro-congolaise	ful	FUF
544	peul (maasina)	919 700	Mali	nigéro-congolaise	ful	FUL
545	peul (Nigéria)	7 611 000	Nigéria	nigéro-congolaise	ful	FUV
546	peul (ouest Niger)	1 150 000	Niger	nigéro-congolaise	ful	FUH
547	peul (pulaar)	2 921 300	Sénégal	nigéro-congolaise	ful	FUC
548	pohnpei	27 000	Micronésie	austronésienne	pon	PNF
549	pokomam, central	8 600	Guatemala	maya	myn	POC
550	pokomam, est	12 500	Guatemala	maya	myn	POA
551	pokomam, sud	27 912	Guatemala	maya	myn	POU
552	pokomchi, est	35 000	Guatemala	maya	myn	POH
553	pokomchi, ouest	50 000	Guatemala	maya	myn	POB
554	polonais	44 000 000	Pologne	indo-européenne	pol	PQL
555	portugais	176 000 000	Portugal	indo-européenne	por	POR
556	quetchua, ancash, chiquian	25 000	Pérou	quetchua	que	QEC

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
557	quetchua, ancash, conchucos, nord	200 000	Pérou	quetchua	que	QED
558	quetchua, ancash, conchucos, sud	250 000	Pérou	quetchua	que	QEH
559	quetchua, ancash, corongo	8 500	Pérou	quetchua	que	QEE
560	quetchua, ancash, huaylas	300 000	Pérou	quetchua	que	QAN
561	quetchua, ancash, sihuas	10 000	Pérou	quetchua	que	QES
562	quetchua, arequipa-la union	18 000	Pérou	quetchua	que	QAR
563	quetchua, ayacuchos	1 000 000	Pérou	quetchua	que	QUY
564	quetchua, cajamarca	35 000	Pérou	quetchua	que	QNT
565	quetchua, chachapoyas	3 000	Pérou	quetchua	que	QUK
566	quetchua, Cuzco	1 500 000	Pérou	quetchua	que	QUZ
567	quetchua, highland, calderon	25 000	Equateur	quetchua	que	QUD
568	quetchua, highland, cañar	100 000	Equateur	quetchua	que	QQC
569	quetchua, highland, chimborazo	1 000 000	Equateur	quetchua	que	QUG
570	quetchua, highland, imbabura	300 000	Equateur	quetchua	que	QHO
571	quetchua, highland, loja	10 000	Equateur	quetchua	que	QQU
572	quetchua, highland, tungurahua	7 000	Equateur	quetchua	que	QQS
573	quetchua, huanuco, huallaga	40 000	Pérou	quetchua	que	QUB
574	quetchua, huanuco, huamalies nord dos de mayo	60 000	Pérou	quetchua	que	QEJ
575	quetchua, huanuco, panao	17 540	Pérou	quetchua	que	QEM
576	quetchua, lambayeque	20 000	Pérou	quetchua	que	QUF
577	quetchua, lowland, napo	5 000	Equateur	quetchua	que	QLN
578	quetchua, lowland, tena	5 000	Equateur	quetchua	que	QUW
579	quetchua, margos-yarowilca-lauricocha	120 000	Pérou	quetchua	que	QEI
580	quetchua, nord Bolivie	116 500	Bolivie	quetchua	que	QUL
581	quetchua, nord Junin	60 000	Pérou	quetchua	que	QJU
582	quetchua, nord Lima, cajatambo	16 525	Pérou	quetchua	que	QNL
583	quetchua, nord-ouest Jujuy	5 000	Argentine	quetchua	que	QUO
584	quetchua, pacaroas	250	Pérou	quetchua	que	QCP
585	quetchua, pasco, santa Ana de Tusi	10 000	Pérou	quetchua	que	QEF
586	quetchua, pasco-yanahuanca	20 500	Pérou	quetchua	que	QUR
587	quetchua, pastaza, nord	4 000	Equateur	quetchua	que	QLB
588	quetchua, pastaza, sud	1 000	Pérou	quetchua	que	QUP
589	quetchua, san Martin	40 000	Pérou	quetchua	que	QSA
590	quetchua, san Rafael-Huariaca	90 000	Pérou	quetchua	que	QEG
591	quetchua, Sandiego del Estero	60 000	Argentine	quetchua	que	QUS
592	quetchua, sud Bolivie	3 632 500	Bolivie	quetchua	que	QUH
593	quetchua, wanca, huaylla	300 000	Pérou	quetchua	que	QHU
594	quetchua, wanca, jauja	14 549	Pérou	quetchua	que	QHJ
595	quetchua, yauyos	18 950	Pérou	quetchua	que	QUX
596	quiché, central	216 910	Guatemala	maya	myn	QUC
597	quiché, cunen	6 500	Guatemala	maya	myn	CUN
598	quiché, est, chichicastenango	100 000	Guatemala	maya	myn	QUU
599	quiché, joyabaj	54 298	Guatemala	maya	myn	QUJ
600	quiché, ouest central	250 000	Guatemala	maya	myn	QUT
601	quiché, san Andrés	19 728	Guatemala	maya	myn	QIE
602	quinault	6	USA	salish	sal	QUN
603	rajasthani (bagri)	2 007 000	Inde	indo-européenne	raj	BGQ
604	rajasthani (gade lohar)	500	Inde	indo-européenne	raj	GDA
605	rajasthani (gujari)	1 400 000	Inde	indo-européenne	raj	GJU
606	rajasthani (harauti)	572 000	Inde	indo-européenne	raj	HOJ
607	rajasthani (malvi)	1 102 000	Inde	indo-européenne	raj	MUP
608	rajasthani (marwari, Inde)	12 963 000	Inde	indo-européenne	raj	MKD
609	rajasthani (marwari, Pakistan)	220 000	Pakistan	indo-européenne	raj	MRI
610	rajasthani (mewari)	1 220 000	Inde	indo-européenne	raj	MTR
611	rapanui	2 400	Chili	austronésienne	rap	PBA
612	rarotonga	43 000	Iles Cook	austronésienne	rar	RRT
613	rhéto-roman	40 000	Suisse	indo-européenne	roh	RHE
614	roumain	26 000 000	Roumanie	indo-européenne	rum/ron	RUM
615	rundi	6 000 000	Burundi	nigéro-congolaise	run	RUD
616	russe	167 000 000	Russie	indo-européenne	rus	RUS

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
617	rwanda	7 362 800	Rwanda	nigéro-congolaise	kin	RUA
618	sacapulteco	36 823	Guatemala	maya	myn	QUV
619	salish, southern puget sound	107	USA	salish	sal	SLH
620	salish, straits	30	Canada	salish	sal	STR
621	sami de Lule	2 000	Suède	ouralienne	smj	LPL
622	sami d'Inari	250	Finlande	ouralienne	smn	LPI
623	sami du nord	21 000	Norvège	ouralienne	sme	LPR
624	sami du sud	600	Suède	ouralienne	sma	LPC
625	sami skolt	320	Finlande	ouralienne	sms	LPK
626	samoan	426 394	Samoa occidental	austronésienne	smo	SMY
627	sandawe	70 000	Tanzanie	khoï-san	sad	SBR
628	sango	400 000	Centrafrique	créole	sag	SAJ
629	sanskrit	6 106	Inde	indo-européenne	san	SKT
630	santal	6 050 000	Inde	austro-asiatique	sat	SNT
631	sarde (variante logudoraise)	1 500 000	Italie	indo-européenne	srd	SRD
632	sasak	2 100 000	Indonésie	austronésienne	sas	SAS
633	sechelt	40	Canada	salish	sal	SEC
634	selkoupe	1 570	Russie	ouralienne	sel	SAK
635	senaya	500	Iran	afro-asiatique	sem	SYN
636	serbo-croate (=bosniaque [bos] + serbe [scc/srp] + croate [scr/hrv])	10 200 000	Yougoslavie	indo-européenne	bos-scc/srp-scr/hrv	SRC
637	sérère	1 051 000	Sénégal	nigéro-congolaise	srr	SES
638	shona	7 000 000	Zimbabwe	nigéro-congolaise	sna	SHD
639	shuswap	745	Canada	salish	sal	SHS
640	sidamo	1 876 329	Ethiopie	afro-asiatique	sid	SID
641	sindhi	19 720 000	Pakistan	indo-européenne	snd	SND
642	singhalais	13 220 000	Sri Lanka	indo-européenne	sin	SNH
643	sipacapense	6 000	Guatemala	maya	myn	QUM
644	sirmauri	14 542	Inde	indo-européenne	him	SRX
645	skagit	100	USA	salish	sal	SKA
646	slovaque	5 606 000	Slovaquie	indo-européenne	slo/slk	SLO
647	slovène	2 000 000	Slovénie	indo-européenne	slv	SLV
648	snohomish	10	USA	salish	sal	SNO
649	somali	9 472 000	Somalie	afro-asiatique	som	SOM
650	songhai	400 000	Mali	nilo-saharienne	son	SON
651	soninké	1 067 000	Mali	nigéro-congolaise	snk	SNN
652	soqotri	70 000	Yemen	afro-asiatique	sem	SQT
653	sora	288 000	Inde	austro-asiatique	mun	SRB
654	sotho du nord	3 851 000	Afrique du Sud	nigéro-congolaise	nso	SRT
655	sotho du Sud	4 197 000	Lesotho	nigéro-congolaise	sot	SSO
656	soundanais	27 000 000	Indonésie	austronésienne	sun	SUO
657	soussou	923 500	Guinée	nigéro-congolaise	sus	SUD
658	spokane	50	USA	salish	sal	SPO
659	squamish	20	Canada	salish	sal	SQU
660	suédois	9 000 000	Suède	indo-européenne	swe	SWD
661	sukuma	5 000 000	Tanzanie	nigéro-congolaise	suk	SUA
662	swahili	5 000 000	Tanzanie	nigéro-congolaise	swa	SWA
663	swati	1 670 000	Swaziland	nigéro-congolaise	ssw	SWZ
664	tacaneco	20 000	Guatemala	maya	myn	MTZ
665	tadjik	4 380 000	Tajikistan	indo-européenne	tgk	PET
666	tagalog	17 000 000	Philippines	austronésienne	tgl	TGL
667	tahitien	125 000	Polynésie Française	austronésienne	tah	THT
668	tamatchek	270 000	Mali	afro-asiatique	tmh	TAQ
669	tamatchek (tahaggart)	62 000	Algérie	afro-asiatique	tmh	THV
670	tamatchek (tawallammat)	640 000	Niger	afro-asiatique	tmh	TTQ
671	tamatchek (tayart)	250 000	Niger	afro-asiatique	tmh	THZ
672	tamoul	66 000 000	Inde	dravidienne	tam	TCV
673	tatar	7 000 000	Russie	altaïque	tat	TTR

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
674	tchèque	12 000 000	République tchèque	indo-européenne	cze/ces	CZC
675	tchétchène	1 000 000	Russie	nord-caucasienne	che	CJC
676	tchouvache	1 809 000	Russie	altaïque	chv	CJU
677	tectiteco	1 200	Guatemala	maya	myn	TTC
678	télougou	69 666 000	Inde	dravidienne	tel	TCW
679	temne	1 200 000	Sierra Leone	nigéro-congolaise	tem	TEJ
680	tereno	15 000	Brésil	arawak	ter	TEA
681	tetum	450 000	Indonésie	austronésienne	tet	TTM
682	thaï	20 047 000	Thaïlande	taï-kadaï	tha	THJ
683	thompson	595	Canada	salish	sal	THP
684	tibétain	1 254 000	Chine	sino-tibétaine	tib/bod	TIC
685	tigré	800 000	Erythrée	afro-asiatique	tig	TIC
686	tigrigna	5 135 000	Ethiopie	afro-asiatique	tir	TGN
687	tiv	2 212 000	Nigéria	nigéro-congolaise	tiv	TIV
688	tlingit	775	USA	na-déné	tli	TLI
689	tojolabal	36 000	Guatemala	maya	myn	TOJ
690	tok pisin	50 000	Papouasie Nouvelle-Guinée	créole	tpi	PDG
691	tokelau	4 500	Tokelau	austronésienne	tkl	TOK
692	tonga (nyasa)	220 000	Malawi	nigéro-congolaise	tog	TOG
693	tongan (îles Tonga)	123 000	Tonga	austronésienne	ton	TOV
694	touva	233 400	Russie	altaïque	tyv	TUN
695	tsigane (Balkan)	1 000 000	Yougoslavie	indo-européenne	rom	RMN
696	tsigane (Baltique)	100 000	Pologne	indo-européenne	rom	ROM
697	tsigane (Carpathes)	241 000	République tchèque	indo-européenne	rom	RMC
698	tsigane (kalo finnois)	5 000	Finlande	indo-européenne	rom	RMF
699	tsigane (sinte)	200 000	Yougoslavie	indo-européenne	rom	RMO
700	tsigane (vlax)	1 500 000	Roumanie	indo-européenne	rom	RMY
701	tsimshian	500	Canada	pénutienne	tsi	TSI
702	tsonga	3 165 000	Afrique du Sud	nigéro-congolaise	tso	TSO
703	tswana	4 000 000	Botswana	nigéro-congolaise	tsn	TSW
704	tumbuka	2 000 000	Malawi	nigéro-congolaise	tum	TUW
705	turc (= turc [tur] + turc ottoman [ota], différenciés uniquement par l'écriture selon Ethnologue)	61 000 000	Turquie	altaïque	tur-ota	TRK
706	turi	2 000	Inde	austro-asiatique	mun	TRD
707	turkmène	6 400 000	Turkmenistan	altaïque	tuk	TCK
708	turoyo	70 000	Turquie	afro-asiatique	sem	SYR
709	tuvalu	11 000	Tuvalu	austronésienne	tvl	ELL
710	twana	350	USA	salish	sal	TWA
711	tzeltal, bachajon	100 000	Mexique	maya	myn	TZB
712	tzeltal, oxchuc	200 000	Mexique	maya	myn	TZH
713	tzolzil, chamula	130 000	Mexique	maya	myn	TZC
714	tzolzil, ch'enalho	35 000	Mexique	maya	myn	TZE
715	tzolzil, huixtan	20 000	Mexique	maya	myn	TZU
716	tzolzil, san Andrés Larrainzar	50 000	Mexique	maya	myn	TZS
717	tzolzil, venustiano carranza	4 226	Mexique	maya	myn	TZO
718	tzolzil, zinacantan	25 000	Mexique	maya	myn	TZZ
719	tzutujil, est	50 000	Guatemala	maya	myn	TZJ
720	tzutujil, ouest	33 000	Guatemala	maya	myn	TZT
721	ukrainien	47 000 000	Ukraine	indo-européenne	ukr	UKR
722	umbundu	4 003 000	Angola	nigéro-congolaise	umb	MNF
723	uspanteco	3 000	Guatemala	maya	myn	USP
724	vaï	105 000	Libéria	nigéro-congolaise	vai	VAI
725	venda	750 000	Afrique du Sud	nigéro-congolaise	ven	VEN
726	vietnamien	68 000 000	Vietnam	austro-asiatique	vie	VIE
727	vote	25	Russie	ouralienne	vot	VOD

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
728	walamo	1 231 673	Ethiopie	afro-asiatique	wal	WBC
729	wali	487	Soudan	nilo-saharienne	nub	WLL
730	waray	2 437 688	Philippines	austronésienne	war	WRY
731	washo	10	USA	hoka	was	WAS
732	wolof	3 215 000	Sénégal	nigéro-congolaise	wol	WOL
733	xhosa	6 876 000	Afrique du Sud	nigéro-congolaise	xho	XOS
734	yao	1 597 000	Malawi	nigéro-congolaise	yao	YAO
735	yapois	6 592	Micronésie	austronésienne	yap	YPS
736	yi de Sichuan	1 600 000	Chine	sino-tibétaine	iii	III
737	yiddish	3 000 000	Israël	indo-européenne	yid	YDD
738	yoruba	20 000 000	Nigéria	nigéro-congolaise	yor	YOR
739	zandé	1 142 000	Congo	nigéro-congolaise	znd	ZAN
740	zapotèque, albarradas	5 500	Mexique	otomangue	zap	ZAS
741	zapotèque, amatlan	5 000	Mexique	otomangue	zap	ZPO
742	zapotèque, asuncion	100	Mexique	otomangue	zap	ZOO
743	zapotèque, ayoquesco	876	Mexique	otomangue	zap	ZAF
744	zapotèque, cajonos	5 000	Mexique	otomangue	zap	ZAD
745	zapotèque, chichicapan	4 000	Mexique	otomangue	zap	ZPV
746	zapotèque, choapan	24 000	Mexique	otomangue	zap	ZPC
747	zapotèque, coatecas altas	5 000	Mexique	otomangue	zap	ZAP
748	zapotèque, coatlan	500	Mexique	otomangue	zap	ZPS
749	zapotèque, el alto	900	Mexique	otomangue	zap	ZPP
750	zapotèque, elotepec	200	Mexique	otomangue	zap	ZTE
751	zapotèque, gueva de Humbolt	7 000	Mexique	otomangue	zap	ZPG
752	zapotèque, güila	9 500	Mexique	otomangue	zap	ZTU
753	zapotèque, istmo	85 000	Mexique	otomangue	zap	ZAI
754	zapotèque, ixtlan sud-est	6 000	Mexique	otomangue	zap	ZPD
755	zapotèque, juarez , sierra	4 000	Mexique	otomangue	zap	ZAA
756	zapotèque, lachiguiri	5 000	Mexique	otomangue	zap	ZPA
757	zapotèque, lachirioag	2 000	Mexique	otomangue	zap	ZTC
758	zapotèque, loxicha	50 000	Mexique	otomangue	zap	ZTP
759	zapotèque, mazaltepec	2 200	Mexique	otomangue	zap	ZPY
760	zapotèque, miahuatlan central	80 000	Mexique	otomangue	zap	ZAM
761	zapotèque, mitla	19 500	Mexique	otomangue	zap	ZAW
762	zapotèque, mixtepec	7 000	Mexique	otomangue	zap	ZPM
763	zapotèque, ocotlan ouest	15 000	Mexique	otomangue	zap	ZAC
764	zapotèque, ozolotepec	6 500	Mexique	otomangue	zap	ZAO
765	zapotèque, petapa	8 000	Mexique	otomangue	zap	ZPE
766	zapotèque, quiavicuzas	4 000	Mexique	otomangue	zap	ZPJ
767	zapotèque, quioquitani y quieri	4 000	Mexique	otomangue	zap	ZTQ
768	zapotèque, rincon	23 000	Mexique	otomangue	zap	ZAR
769	zapotèque, rincon sud	12 000	Mexique	otomangue	zap	ZSR
770	zapotèque, san Augustin mixtepec	59	Mexique	otomangue	zap	ZTM
771	zapotèque, san Baltazar loxicha	1 500	Mexique	otomangue	zap	ZPX
772	zapotèque, san Juan guelavia	28 000	Mexique	otomangue	zap	ZAB
773	zapotèque, san Pedro quiatoni	12 000	Mexique	otomangue	zap	ZPF
774	zapotèque, san Vicente coatlan	2 430	Mexique	otomangue	zap	ZPT
775	zapotèque, santa catarina albarradas	1 000	Mexique	otomangue	zap	ZTN
776	zapotèque, santa Ines yatzechi	2 235	Mexique	otomangue	zap	ZPN
777	zapotèque, santa Maria quiegolani	3 000	Mexique	otomangue	zap	ZPI
778	zapotèque, santiago lapaguia	4 200	Mexique	otomangue	zap	ZTL
779	zapotèque, santiago xanica	2 500	Mexique	otomangue	zap	ZPR
780	zapotèque, solo de vega est	6 500	Mexique	otomangue	zap	ZPL
781	zapotèque, tabaa	2 000	Mexique	otomangue	zap	ZAT
782	zapotèque, tejalapan	124	Mexique	otomangue	zap	ZTT
783	zapotèque, texmelucan	4 100	Mexique	otomangue	zap	ZPZ
784	zapotèque, tilquiapan	2 700	Mexique	otomangue	zap	ZTS
785	zapotèque, tlacolulita	135	Mexique	otomangue	zap	ZPK
786	zapotèque, totomachapan	259	Mexique	otomangue	zap	ZPH
787	zapotèque, xadani	338	Mexique	otomangue	zap	ZAX

	Nom de la langue	Locuteurs	Pays	Famille	639-2	Ethno.
788	zapotèque, xanaguia	2 500	Mexique	otomangue	zap	ZTG
789	zapotèque, yalalag	5 000	Mexique	otomangue	zap	ZPU
790	zapotèque, yareni	6 000	Mexique	otomangue	zap	ZAE
791	zapotèque, yatee	3 000	Mexique	otomangue	zap	ZTY
792	zapotèque, yatzachi	2 500	Mexique	otomangue	zap	ZAV
793	zapotèque, yautepec	314	Mexique	otomangue	zap	ZPB
794	zapotèque, zaachila	550	Mexique	otomangue	zap	ZTX
795	zapotèque, zanina	770	Mexique	otomangue	zap	ZPW
796	zapotèque, zoogocho	1 000	Mexique	otomangue	zap	ZPQ
797	zay	4 880	Ethiopie	afro-asiatique	sem	ZWA
798	zenaga	200	Mauritanie	afro-asiatique	zen	ZEN
799	zhuang; chuang (nord)	10 000 000	Chine	taï-kadaï	zha	CCX
800	zhuang; chuang (sud)	4 000 000	Chine	taï-kadaï	zha	CCY
801	zoulou	9 142 000	Afrique du Sud	nigéro-congolaise	zul	ZUU
802	zuni	6 413	USA	langue isolée	zun	ZUN
	TOTAL	**5 568 409 621**				

A.7 PARTIES DU DISCOURS UTILISÉES DANS LAODICT (NIVEAU CATÉGORIE)

1) NOMS

Nom

- Nom commun
- Locution nominale
- Nom propre

Mots entrant dans la constitution des locutions nominales

- Générique de classification
- Générique d'abstraction
- Générique de réalisation
- Générique de concrétisation
- Générique de titre

2) PRONOMS

Pronom

- Pronom personnel
- Pronom impersonnel neutre
- Pronom réciproque
- Pronom réfléchi
- Pronom indéfini-interrogatif
- Locution pronominale indéfinie-interrogative
- Pronom relatif
- Spécificatif

Mots entrant dans la constitution des locutions pronominales

- Adjectif indéfini
- Locution adjectivale indéfinie
- Adnominal

3) VERBES

Verbe

- Verbe d'action
- Verbe d'état
- Verbe désidératif

Auxiliaire

- Auxiliaire de voie passive
- Auxiliaire de temps
- Auxiliaire d'aspect
- Auxiliaire secondaire
- Auxiliaire modal de l'impératif
- Auxiliaire modal du subjonctif dubitatif
- Auxiliaire royal

4) PRÉDICATIFS

Adjectifs-adverbes

- Adjectif-adverbe qualificatif
- Locution adjectivo-adverbiale qualificative
- Adjectif-adverbe quantitatif indéfini

Adjectifs purs

- Adjectif démonstratif
- Adjectif interrogatif
- Adjectif indéfini
- Adjectif numéral
- Locution adjectivale numérale

Adverbes purs

- Adverbe d'affirmation
- Adverbe de négation
- Adverbe modal postverbal de l'affirmatif
- Adverbe modal postverbal de l'interrogatif
- Adverbe modal postverbal du volitif et de l'exhortatif
- Adverbe modal postverbal de l'impératif
- Adverbe modal préverbal du conditionnel
- Adverbe modal postverbal du subjonctif dubitatif
- Adverbe quantitatif
- Adverbe conjonctif
- Adverbe de restriction

Mots entrant dans la constitution de locutions prédicatives

- Nom entrant dans la constitution de locutions prédicatives

Mots marquant le nombre et le genre

- Collectif
- Qualificatif

5) PRÉPOSITIONS

- Préposition

6) CONJONCTIONS

- Conjonction de coordination
- Conjonction de subordination

7) INTERJECTIONS

- Interjectif direct
- Interjectif indirect
- Exclamatif

A.8 PARTIES DU DISCOURS, AVEC EXEMPLES, DU LAOTIEN (NIVEAU SOUS-CATÉGORIE)

Groupe des noms (ຄໍານາມ)

Ce groupe contient les noms : « *mot désignant soit une espèce ou un représentant de l'espèce (noms communs) soit un individu particulier (noms propres)* ».

Outre ces classiques noms communs (ສາມັນນາມ), locutions nominales et noms propres (ວິສາມັນນາມ), nous faisons aussi apparaître dans ce groupe des éléments entrant dans la constitution de locutions nominales[1] (les génériques) :

- **génériques de classification** (ou classificateur - ຄໍາປັດໄຈ) : nom de classe qui peut être suivi par un déterminant pour la préciser (enchâssement possible),
- **génériques d'abstraction** : mot clé qui doit être accompagné par un déterminant pour former un nom abstrait,
- **génériques de réalisation** : mot clé qui doit être accompagné par un déterminant (verbe) pour le nominaliser,
- **génériques de concrétisation** : générique qui doit être suivi par un déterminant pour former un nom concret,
- **génériques de titre** : rang social d'un individu qui doit être suivi par un déterminant (qui peut être un nom propre) pour désigner une personne,

Catégorie	Sous-catégorie	Exemples
Noms		
Nom commun	Nom simple quantifiable	ຕັກແຕນ (sauterelle)
	Nom simple non quantifiable	ຝົນ (pluie)
Locution nominale	Locution nominale quantifiable	ນົກເຄົ້າ (hibou)
	Locution nominale non quantifiable	ນ້ຳຕານຊາຍ (sucre en poudre)
Nom propre	Nom individuel	ສີລາ (pierre), ແກ້ວ (gemme), ເພັດ (diamant), ຄໍາ (or), ບົວ (lotus), ສີ (lion), ອິນ (Indra), ສວນ (Içvara), ສີດາ (Sita), ພຸດ (Bouddha), ຊັຍ (victoire), ແພງ (chéri), ພອນ (voeu, bénédiction), ຂີ້ (excrément), ຢูด (grognement de porc)
	Nom patronymique	ວິຣะอ๋าง (Viravong)
	Nom de règne	สົมเด็ดพระเจ้าสุริยะอ๋างสาทัมมิกะราด (Souligna Vongsa)
Génériques		
Générique de classification	Classificateur d'agent de formation laotienne	ผู้ (personne), ຄົນ (personne), ไท (ethnie, habitant, gens), ชาว (gens d'une communauté ou ayant une activité commune), นัก (celui qui a choisi ou qui pratique une activité), ช่าง (artisan, ouvrier), หมໍ (praticien), พ่ (père, chef), แม่ (mère, chef), ลูก (enfant, subordonné), นาย (maître), ข้า (esclave, serviteur), ผี (génie, esprit), พระ (divinité), เจ้า (maître, seigneur, propriétaire)

[1] Procédé appelé ຄໍາຕັດທິດ. Il s'agit bien ici uniquement de locutions nominales et non de groupes nominaux. Les adjectifs sont traités dans le groupe des prédicatifs. Par ailleurs, les spécificatifs sont traités dans le groupe des pronoms du fait de leur fonction.

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Classificateur d'agent de formation indo-européenne	ກອນ (agent qui opère, qui produit), ຊ້ຳ (personne dont l'état est ...), ຊ້ຳ (personne dont l'état est ...)
	Classificateur de lieu	ບ່ອນ (lieu, endroit), ທີ່ (lieu, endroit), ເມືອງ (pays, ville), ບ້ານ (village, bourg, quartier), ຊຽງ (ville), ທ່າ (port), ປາກ (embouchure), ບໍ່ (source, mine), ນາ (rizière), ທົ່ງ (plaine), ດອນ (île, îlot), ໂພນ (butte), ພູ (mont, montagne), ຜາ (rocher), ນ້ຳ (cours d'eau), ເຊ (rivière), ຫ້ວຍ (ruisseau), ໜອງ (lac, étang, mare), ບຶງ (étang, ancien méandre), ທາງ (route, chemin)
	Classificateur de végétaux	ຕົ້ນ (plante), ກົກ (tronc, arbre), ກຳ (touffe, bouquet), ເຄືອ (liane), ຫຽ້ງ (herbe), ຜັກ (légume), ມັນ (tubercule), ເຂົ້າ (céréale), ຫົວ (tête, tubercule, rhizome, oignon), ໝາກ (fruit), ດອກ (fleur), ໝໍ່ (pousse, bourgeon), ໃບ (feuille)
	Classificateur de matière	ດິນ (terre), ຫີນ (pierre), ແກ້ວ (verre, gemme), ເຫຼັກ (fer), ໄມ້ (bois), ຂີ້ (excrément, résidu, produit), ນ້ຳ (eau, liquide, jus), ຢາງ (gomme, glue), ຢາ (médicament)
	Classificateur d'instrument	ໄມ້ (instrument), ຜ້າ (tissu, linge), ສາຍ (ligne, fil)
	Classificateur de locaux	ເຮືອນ (maison), ໂຮງ (édifice, bâtiment), ຫໍ (temple, autel), ຮ້ານ (boutique), ຫ້າງ (magasin),
	Classificateur d'animaux	ຕົວ (corps), ໂຕ (corps), ນົກ (oiseau), ງູ (serpent), ແມງ (insecte), ປາ (poisson), ປູ (crustacé de type crabe), ກຸ້ງ (crustacé de type crevette), ຫອຍ (coquillage)
Générique d'abstraction	Générique d'abstraction de formation laotienne	ຄວາມ (mot, idée), ອັນ (chose, objet), ທີ່ (situation, lieu)
	Générique d'abstraction de formation indo-européenne	ກັມ (acte), ກິດ (tâche, devoir), ສາດ (science), ບານ (garde), ພາບ (état, fait d'être)
Générique de réalisation	Générique de réalisation	ການ (travail, réalisation)
Générique de concrétisation	Générique de concrétisation	ເຄື່ອງ (instrument), ຂອງ (chose, objet), ອັນ (chose, objet), ແນວ (sorte, espèce)
Générique de titre	Titre populaire	ບັກ (gars), ອີ່ (fille), ນາຍ (maître), ສາວ (femme)
	Titre bourgeois	ທ້າວ (sieur), ນາງ (dame, madame, demoiselle, mademoiselle), ທ່ານ (monsieur)
	Titre de mandarinat	ພຍາ (seigneur, haut dignitaire), ອາດຍາ (seigneur, fonctionnaire d'autorité)
	Titre de bourgeois allié à la noblesse	ໝ່ອມ (monsieur ou madame), ໝ່ອມນາງ (madame), ເຈົ້າໝ່ອມ (madame), ໝ່ອມຍິງ (madame, la princesse)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Titre nobiliaire de l'ancien régime	ເຈົ້າ (noble, prince, princesse), ເຈົ້າຊາຍ (prince de sang), ເຈົ້າຍິງ (princesse royale), ເຈົ້າເຮືອນຍິງ (princesse, de la Maison de Champassac), ເຈົ້າຟ້າ (roi, littéralement maître du ciel), ເຈົ້າຟ້າຊາຍ (prince héritier), ສເດັດເຈົ້າ (altesse), ສເດັດເຈົ້າຟ້າ (altesse royale), ສເດັດເຈົ້າຟ້າຊາຍ (altesse royale, prince héritier), ເຈົ້າຍັງຂ່ອນ (altesse, de la Maison de Muang Phouan, Xieng-Khouang), ສາທຸ ou ທຸ (seigneur, noble, gentil, de Louang Phrabang), ພຣະບາດສົມເດັດພຣະເຈົ້າ (majesté)
	Titre nobiliaire ancien	ຂຸນ (seigneur), ທ້າວ (seigneur), ນາງ (princesse), ບ ou ບາຄານ (seigneur), ພຍາ (seigneur)
	Titre de mandarinat militaire ancien	ເພັຽ (chevalier), ເມືອງ (gouverneur), ໝື່ນ (général de brigade), ແສນ (général de division)
	Titre de religieux bouddhiste	ຈົ (novice, bonzillon), ສາມະເນນ ou ເນນ (novice bonzillon), ພຣະ ou ພະ (bonze, vénérable), ພຣະຄຣູ (vénérable maître), ພຣະຄຣູຫຼັກຄຳ (vénérable maître-pilier d'or), ພຣະຄຣູຫຼັກຄຳແກ້ວ (vénérable maître-pilier d'or et de gemme), ພຣະສົມເດັດ (très vénérable)
	Titre d'ancien religieux	ຊຽງ (ancien novice), ທິດ (ancien bonze non gradé), ຈານ (ancien bonze gradé), ຈານຄຣູ (ancien bonze gradé enseignant), ມະຫາ (ancien bonze ayant fait des études de pali)
	Titre ancien de citation en justice	ແດ (sieur, dame)
	Titre familial	ປູ່ (grand-père paternel), ຍ່າ (grand-mère paternelle), ພໍ່ເຖົ້າ ou ພໍ່ຕູ້ (grand-père maternel), ແມ່ເຖົ້າ ou ແມ່ຕູ້ ou ແມ່ຍາຍ (grand-mère maternelle), ພໍ່ (père), ແມ່ (mère), ລຸງ (oncle, ainé des parents), ຕາ (oncle maternel), ປ້າ (tante, ainée des parents), ອາວ (oncle, cadet du père), ອາ (tante, cadette du père), ນ້າບ່າວ (oncle, cadet de la mère), ນ້າສະວ (tante, cadette de la mère), ອ້າຍ (frère aîné), ເອື້ອຍ (soeur aînée), ນ້ອງ (cadet, cadette), ລູກ (fils, fille), ຫຼານ (neveu, nièce, cousin, cousine, petit-fils, petite-fille), ວ (sieur),

Groupe des pronoms (ຄຳສັບພະນາມ / ຄຳສັພນາມ)
Ce groupe contient les pronoms : « *mot représentant un nom, un adjectif, une phrase et dont les fonctions syntaxiques sont identiques à celles d'un nom* », ainsi que des mots entrant dans la constitution des locutions pronominales.

Catégorie	Sous-catégorie	Exemples
Pronoms		
Pronom personnel	Pronom personnel réel	ກັນ (je : populaire et emphatique), ກູ (je : populaire, péjoratif, archaïque, dialectal), ...
	Titre d'étiquette commune	ພໍ່ເຖົ້າ (grand-père), ແມ່ເຖົ້າ (grand-mère, belle-mère), ພໍ່ (père), ແມ່ (mère), ລຸງ (oncle), ປ້າ (tante), ລູກ (enfant), ຫຼານ (neveu, nièce, cousin, cousine, petit-fils, petite-fille), ອ້າຍ (aîné), ເອື້ອຍ (aînée), ນ້ອງ (cadet, cadette)
	Titre d'étiquette bourgeoise	ຍາ (indique l'autorité)
	Titre d'étiquette familiale noble de Luang-Phrabang	ທຸ (noble, gentil)
	Titre d'étiquette familiale d'autres noblesses	ຍາເຈົ້າ (indique l'autorité du prince)
	Titre d'étiquette royale	ພຣະອົງ (Majesté)
	Titre d'étiquette princière	ສາທຸ (noble, de Louang Phrabang), ອາດຍາເຈົ້າ (autorité, noble, autres que Louang Phrabang), ສະເດດ (Prince), ສະເດດເຈົ້າ (Altesse)
	Titre d'étiquette de bourgeoisie alliée à la noblesse	ໝ່ອມ ou ເຈົ້າໝ່ອມ (monsieur, madame)
	Titre d'étiquette de dignitaires	ພຍາ (seigneur), ອາດຍາຫຼວງ (Haute Autorité, gouverneur, préfet), ອາດຍາ (autorité, fonctionnaire d'autorité), ພະນະທ່ານ (Excellence)
	Titre d'étiquette civile ou militaire	ທ່ານ (monsieur, cadre supérieur), ນາຍ (maître, cadre moyen), ທ່ານນາຍພົນ (général), ທ່ານນາຍພັນ (commandant), ທ່ານນາຍຮ້ອຍ (capitaine)
	Titre d'étiquette pour les enseignants	ນາຍຄຫ (maître, instituteur), ອາຈານ (professeur, du secondaire), ສາສຕຣາຈານ (professeur, du supérieur), ນາງ (madame, mademoiselle)
	Titre d'étiquette bouddhiste de religieux à laïcs	ພໍ່ອຍກ (homme), ແມ່ນອຍກ (femme)
	Titre d'étiquette bouddhiste de laïcs à religieux	ຈົວ ou ເນນ (novice, bonzillon), ພຣະ (bonze), ຄູບາ (bonze), ສົມເດັດ (vénérable), ແມ່ຊີ (religieuse)
	Titre d'étiquette catholique	ຄຸນພໍ່ (Révérend Père), ຄຸນແມ່ (Soeur, Révérende Mère), ພຣະສັງຄະຣາດ (Monseigneur, Evêque)
	Titre d'étiquette populaire commune	ໝູ່ (ami, camarade), ຊຽງ (ancien bonzillon), ທິດ (ancien bonze), ຈານ (ancien bonze gradé), ມະຫາ (ancien bonze diplômé de pâli)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Titre d'étiquette mixte socio-familiale	ອ້າຍທິດ (de jeune à aîné), ລຸງຈາບ (de jeune à homme mur), ທຸເອີ້ອຍ (à dame noble plus âgée), ຍາລູກ (de personne âgée à jeune homme de haute situation), ຍາແມ่ (dame respectable)
Pronom impersonnel neutre	Pronom impersonnel neutre	ມັນ (il) Nota : Théorique, n'est pas employé (Marc Reinhorn)
Pronom réciproque	Pronom réciproque	ກັນ (se, l'un-l'autre, les uns-les autres)
Pronom réfléchi	Pronom réfléchi	ໃຕ (soi, soi-même), ເອງ (soi, soi-même)
Pronom indéfini-interrogatif	Pronom indéfini-interrogatif d'objet	ຫຽ້ງ (quoi, que, rien), ສັ່ງ (quoi, que, rien (Laos du sud))
	Pronom indéfini-interrogatif de personne	ໃຜ (quelqu'un, qui)
	Pronom indéfini-interrogatif de lieu	ໃສ (où)
Locution pronominale indéfinie-interrogative	Locution pronominale indéfinie-interrogative de temps	ຍາມໃດ (quand, quel moment, un moment quelconque)
	Locution pronominale indéfinie-interrogative de manière	ຢ่າງໃດ (comment, quelle manière)
	Locution pronominale indéfinie-interrogative d'évaluation de grandeur	ເທົ่າໃດ (combien, comment), ຫໍ่ໃດ (combien, comment)
Pronom relatif	Pronom relatif de personne	ผู้ (qui, que, lequel, laquelle, ...)
	Pronom relatif d'animal	ໃຕ (qui, que, lequel, laquelle, ...)
	Pronom relatif d'objet	ອັນ (qui, que, lequel, laquelle, dont, ...), ເຊ่ິ = ຊ ่ ิ (qui, que, lequel, laquelle, dont, ...)
	Pronom relatif de lieu et d'objet	ที่ = บ่อบ (où, qui, que, lequel, laquelle, dont, ...)
Spécificatif	Spécificatif	ກັນ (unité d'aumône de participation), ก้าบ (nervure, palme, lettre, navette, cuiller), ...
Mots entrant dans la constitution des locutions pronominales		
Adjectif indéfini formant un pronom indéfini avec un classificateur	Adjectif indéfini formant un pronom indéfini avec un classificateur placé avant	ນึ่ง (un), ... ນึ่ง ... ນึ่ง (un ... un autre),... ใด ... ນึ่ง (un quelconque), ອื่ม (autre)
	Adjectif indéfini formant un pronom indéfini avec un classificateur placé après	ทุก = ฆู่ (chaque, tout), จัก (quelque, quelques), ຫຼาย (nombreux, beaucoup)
Locution adjectivale indéfinie formant un pronom indéfini avec un classificateur	Locution adjectivale indéfinie formant un pronom indéfini avec un classificateur placé après	ทุกๆ (tout, tous), ฆ็ดทุก (tout, tous), จัก (quelque, quelques), ຫຼาย (nombreux, beaucoup), ทั๊ ... (locutions diverses)
Adnominal formant un pronom démonstratif avec un spécificatif	Adnominal formant un pronom démonstratif avec un spécificatif	ນี้ (-ci), นั้น (-là)

Groupe des verbes (ຄໍາກິຣິຍາ)

Ce groupe contient les verbes : « *mot qui, dans une proposition, exprime l'action ou l'état du sujet* », un certain nombre d'auxiliaires permettant de marquer l'aspect, le temps, le mode, le mouvement, l'aptitude et le sujet royal. Notons que le mode peut être rendu par ces auxiliaires modaux mais aussi par des adverbes modaux (voir le groupe des prédicatifs).

Catégorie	Sous-catégorie	Exemples
Verbes		
Verbe d'action	Verbe d'action transitif	ຂຽນ (écrire)
	Verbe d'action intransitif	ນອນ (dormir)
Verbe d'état	Verbe d'état dynamique	ມີ (avoir, posséder, exister), ໄດ້ (avoir, gagner, acquérir, réussir, pouvoir), ແລ້ວ (finir, avoir terminé, être terminé, fait)
	Verbe d'état statique	ກົງ (toucher, atteindre, être conforme, juste), ແນ່ນ (être (vrai, juste, exact)), ເປັນ (être (en état)), ..., ຄື (sembler, ressembler, être comme, être pareil à, être seyant), ເຄີຍ (être habitué à, avoir l'habitude de, être capable de, être expert, connaître, savoir), ຊ່ຽ (être apte à, être expert, habile, avoir la pratique, savoir), ຢູ່ (être (en mesure, en situation), rester), ຍັງ (en rester, il en reste)
Verbe désidératif	Verbe désidératif	ຂໍ (demander, prier, solliciter), ທຽນ (faire don, aumône), ໃຫ=ໄປຣດ (accorder la liberté, la faveur, aider, secourir)
Auxiliaires		
Auxiliaire de voie passive	Auxiliaire de voie passive	ກົງ (être (concerné, touché))
Auxiliaire de temps	Auxiliaire du futur	ຈະ (forme littéraire), ຊິ (forme courante), ຈັກ (forme archaïque), ຈິ (forme intermédiaire), ອິ (forme Louang-Phrabang)
	Auxiliaire du passé	ໄດ້ (avoir, gagner, acquérir, réussir)
	Auxiliaire du futur à accomplir	ຈະໄດ້ (avoir à)
	Syntagme postverbal de procès terminé	ແລ້ວ (passé révolu)
	Syntagme postverbal de procès ultérieurement terminé	ຈະ ... ແລ້ວ (futur accompli), ຈະໄດ້ ... ແລ້ວ (futur accompli et antérieurement à un autre futur à accomplir)
Auxiliaire d'aspect	Syntagme préverbal de procès en cours	ກໍາລັງ (en cours de, en train de, en ce moment)
	Syntagme préverbal de procès venant de commencer	ຫຽກ່ (venir de)
	Syntagme postverbal de procès qui dure	ຢູ່ (rester, demeurer (duratif))
	Syntagme postverbal de procès engagé sans délai	ໂລດ (d'un coup, immédiatement)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
Auxiliaire secondaire	Auxiliaire secondaire de mouvement	ໄປ (aller, indique l'éloignement, de non-arrêt par rapport au locuteur ou à l'actant), ມາ (venir, indique le rapprochement par rapport au locuteur ou à l'actant), ອອກ (sortir, indique le mouvement de sortie, d'éloignement par rapport au locuteur ou à l'actant), ເຂົ້າ (entrée, indique le mouvement d'entrée, de pénétration, d'introduction, d'approche), ຂຶ້ນ (monter, indique le mouvement d'ascension, de progrès, d'augmentation), ລົງ (descendre, indique le mouvement de descente, de régression, de diminution), ໄວ້ (garder, mettre en réserve, indique la garde, la mise en réserve), ໃສ່ (placer, mettre, porter sur, indique la mise en place), ເອົາ (prendre, indique la prise pour son profit), ໃຫ້ (donner, indique l'attribution, le don), ຮອດ = ເຖິງ (atteindre, indique le but atteint ou à atteindre)
	Auxiliaire secondaire d'aptitude	ເປັນ (être en état de, être apte à, savoir), ໄດ້ (pouvoir, être possible)
Auxiliaire modal de l'impératif	Préverbe modal de l'impératif	ຕ້ອງ (devoir, catégorique), ຄວນ (devoir, convenance), ຄົງ (devoir, logique), ຈົ່ງ (impératif optatif), ໃຫ້ (impératif conatif)
Auxiliaire modal du subjonctif dubitatif	Préverbe modal du subjonctif dubitatif	ຄື (il semble)
Auxiliaire royal	Préverbe royal	ຫຣົ່ງ= ທົ່ງ (daigner), ຊົ່ງ (daigner)

Groupe des prédicatifs (ຄຳວິເສດ)

Ce groupe contient les déterminants (ຄຳກິຣິຍະຫຽ່ຍ) pris dans le sens élargi : « *élément linguistique qui en détermine un autre (le déterminé - ຄຳຂະຫຽ່ຍ)* ». Ils peuvent déterminer des noms ou des spécificatifs (adjectifs purs), des verbes ou des adverbes (adverbes purs), ou les deux à la fois (adjectifs-adverbes). En plus des adjectifs et des adverbes, il contient les constituants de base de certaines locutions prédicatives ainsi que des mots permettant de marquer le genre (les qualitatifs) et le nombre (les collectifs) :

- **collectifs** : mot clé qui doit être suivi par un déterminant pour marquer le pluriel,
- **qualificatifs** : mot clé qui doit être précédé par un générique pour marquer le genre.

Notons que le mode peut être rendu par des adverbes modaux mais aussi par des préverbes (voir le groupe des verbes).

Catégorie	Sous-catégorie	Exemples
Adjectifs-adverbes		
Adjectif-adverbe qualificatif	Adjectif-adverbe qualificatif de temps	ດົນ (longtemps), ນານ (longtemps), ເທິງ (longtemps), ບຶດໜຶ່ງ (un instant), ບຶດດຽວ (un instant), ບາດ(ບັດ)ນີ້ (à présent), ຈັກໜ່ອຍ (tout à l'heure), ກ່ອນ (avant), ແຕ່ກ່ອນ (autrefois), ກີ້ () : s'emploie dans les locutions ເມື່ອກີ້, ແຕ່ກີ້ (autrefois), ນີ່ກີ້ (à l'instant), ແຕ່ກີ້ (autrefois), ເດີນ (début), ກົກ (début), ເຄິ້ງ (début), ແຕ່ນີ້ (dès maintenant), ແຕ່ນີ້ເນີ້ອໜ້າ (dorénavant), ແຕ່ນັ້ນ (depuis lors), ດຽວນີ້ (maintenant), ພ້ອມກັນ (en même temps), ຊ ານຊາ (en attendant), ຢ่ູບ่ຢ่ູ (inopinément), ໄລດ (immédiatement), ລົດ (immédiatement après), ເລີຍ (sans s'arrêter), ດຶກ (tard), ນີ້ ... (plusieurs locutions), ຍານ ... (plusieurs locutions), ເວລາ ... (plusieurs locutions), ເມື່ອ ... (plusieurs locutions), ຄາວ ... (plusieurs locutions), ພາຍ ... (plusieurs locutions), ເຊົ້າ (le matin), ສວາຍ (la journée), ງາຍ (la fin de matinée), ທ່ຽງ (midi), ຄ่ຳ (le soir, la nuit), ຄືນ (le soir, la nuit), ກາງຄືນ (pendant la nuit), ທ່ຽງຄືນ (minuit), ຕອນ (plusieurs locutions), ເກົ່າ (ancien), ໃໝ່ (nouveau), ລຸ່ນ (après), ຫຼ້າ (dernier), ຫຼັງ (derrière), ໜ້າ (devant)
	Adjectif-adverbe qualificatif de situation, de lieu, d'orientation	ໄກ (loin), ໃກ້ (près), ສູງ (haut), ເທິງ (en haut), ບົນ (en haut), ເໜືອ (au nord, en amont), ຕ่ຳ (en bas), ລ่ຸມ (en dessous), ກ້ອງ (dessous), ໃຕ້ (au sud, en aval, en dessous), ໃນ (à l'intérieur), ນອກ (à l'extérieur), ຊ້າຍ (à gauche), ຂວາ (à droite), ໜ້າ (devant), ຫຼັງ (derrière), ຫັ້ນ (là), ໜີ້ (ici), ພ้ຸນ (là-bas), ພີ້ (ici)
	Adjectif-adverbe qualificatif de manière ordinaire	ດີ (bien, bon), ຈົບ (bien), ຊື่ (droit, honnête, franc, sincère), ...
	Adjectif-adverbe qualificatif de manière onomatopée	ແຊ່ວ (bruyant), ฮี่ (vrombissant), ວາກໆ (d'un braillement), ก๊าบๆ (cri du canard), ...

Catégorie	Sous-catégorie	Exemples
	Adjectif-adverbe qualificatif autre	ແຂ໋ງ (dur, rigide), ຄັກ (adapté, juste, vrai), ຄ່ອຍ (doux, lent, précautionneux), ງ່າຍ (facile, aisé), ໄງ່ (sot), ຈິງ (vrai), ສະອ້ານ (propre, net), ແຊ່ບ (bon, savoureux), ຄ້ຶ (dissipé, incongru), ທຸກ (pauvre), ໜາວ (froid)
Locution adjectivo-adverbiale qualificative	Locution adjectivo-adverbiale qualificative péjorative	ຂີ້ຄ້ານ (paresseux), ຂີ້ລັກ (voleur), ຂີ້ຢາ (opiomane, drogué), ຂີ້ເຫຼົ້າ (alcoolique, ivrogne)
	Locution adjectivo-adverbiale qualificative affective ou morale	ໃຈນ້ອຍ (mesquin), ໃຈດຳ (rancunier), ໃຈຮ້ອນ (impatient)
	Locution adjectivo-adverbiale d'appréciation	ເປັນຕາຮັກ (charmant), ເປັນຕາຊັງ (détestable), ເປັນຕາກິນ (mangeable, comestible ; d'une manière profitable), ເປັນຕາຢ້ານ (effroyable), ເປັນຕາຢາກໄດ້ (désirable ; de façon à inspirer le désir), ເປັນຕາຢາກຮົ້ວ (risible ; drôle, amusant ; de façon à donner envie de rire)
	Locution adjectivo-adverbiale d'impression nuancée	(ເປັນ)ສີເຈັບ (d'une certaine douleur), (ເປັນ)ສີຄັນໆ (d'une légère démangeaison), (ເປັນ)ສີເມື່ອຍໆ (d'une certaine fatigue), (ເປັນ)ສີຮ້າຍໆ (d'un certain air méchant)
Adjectif-adverbe quantitatif indéfini	Adjectif-adverbe quantitatif indéfini superlatif relatif, comparatif	ກວ່າ (plus que)
	Adjectif-adverbe quantitatif indéfini de comparaison	ປານ (comme), ເທົ່າ (comme), ຄື (semblable), ພໍ (suffisamment), ພຽງ (à hauteur de), ສເມີ (régulièrement), ຕ່າງ (différemment)
Adjectifs purs		
Adjectif déterminatif démonstratif	Adjectif démonstratif	ນັ້ນ (-là), ນີ້ (-ci)
Adjectif déterminatif interrogatif	Adjectif interrogatif d'objet	ໃດ (quel ?)
	Adjectif interrogatif quantitatif	ຈັກ (combien ?)
Adjectif déterminatif indéfini	Adjectif indéfini	ໃດ (quel, quelque, quelconque, un), ນຶ່ງ (un quelconque), ...ໃດ...ນຶ່ງ (un quelconque), ...ໃດ...ນັ້ນ (?), ອື່ນ (autre), ຈັກ (quelques), ..., ທຸກ (chaque, tout), ຊ່ (chaque, tout), ຊ (chaque, tout), ທຸກໆ (tous, forme intensive), ໝົດທຸກ (tous, forme collective), ທັ (?), sert à former les locutions : ທັ້ງຫຼາຍ (les), ທັ້ງມວນ (l'ensemble des), ທັ້ງໝົດ (la totalité des)
Adjectif déterminatif numéral	Adjectif numéral cardinal	ນຶ່ງ (un), ສອງ (deux), ...
	Adjectif numéral ordinal d'origine indienne	ເອກ (premier), ໂທ (deuxième, second), ຕຣີ (troisième), ຈັຕວາ (quatrième)
	Adjectif numéral ordinal d'origine sino-thaïe	ຈຽງ / ກຽງ (premier), ຍີ່ (deuxième)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Adjectif numéral autre	ເຄິ່ງ (demi), ...
Locution adjectivale déterminative numérale	Locution adjectivale numérale ordinale	ທີ່ນຶ່ງ (premier), ທີ່ສອງ (deuxième)
Adverbes purs		
Adverbe d'affirmation	Affirmation, adverbe de base	ເອີ (oui), ເຈົ້າ (oui) , ອື (oui, forme populaire),
	Affirmation nuancée	ໂດຍ (oui, forme respectueuse), ໂດຍຂະນ້ອຍ (oui, forme respectueuse)
Adverbe de négation	Négation, adverbe de base	ບໍ່ (ne ... pas, sans, non), ບໍ (ne ... pas, sans, non)
	Négation nuancée	ອີ (non, forme populaire), ຫຼື (non), ບໍ່ໃຊ່ (ne ... pas, assurément : négation renforcée), ບໍ่ຫ่ອນ (ne ... pas, naturellement : négation ?)
	Négation, forme vétative	ຢ่າ (ne faîtes pas ...), ຢ่າຊู(ສู) (ne faîtes pas ..., forme renforcée)
Adverbe modal postverbal de l'affirmatif	Adverbe modal postverbal affirmatif soutenu	ຕວ່ (marque de l'affirmatif), ຕວ่ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif emphatique	ລະ(ແ)ໝ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif péremptoire	ລະບໍ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif de constat	ແລ้ວ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif de désignation	ແຫ຺ະ (marque de l'affirmatif), ລະ (marque de l'affirmatif), ຫ຺ະ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif démonstratif, intensif	ເດ (marque de l'affirmatif), ເນ (marque de l'affirmatif), ນະ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif d'avertissement	ໃດ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif de délibération	ຕີ้ (marque de l'affirmatif), ລະຕີ้ (marque de l'affirmatif)
	Adverbe modal postverbal affirmatif de satisfaction	ເວ้ີຍ (marque de l'affirmatif)
Adverbe modal postverbal de l'interrogatif	Adverbe modal postverbal interrogatif ordinaire	ບໍ่ (marque d'interrogation), ບໍ (marque d'interrogation)
	Adverbe modal postverbal interrogatif archaïque	ຫຼື (marque d'interrogation), ຫឺ (marque d'interrogation), ຮឺ (marque d'interrogation), ฮึ (marque d'interrogation)
	Adverbe modal postverbal interrogatif familier	ເຫ໋ (marque d'interrogation)
	Adverbe modal postverbal interrogatif d'incitation	ເບ໋ະ (marque d'interrogation)
	Adverbe modal postverbal interrogatif insistant	ຫວ໋ະ (marque d'interrogation)
	Adverbe modal postverbal interrogatif de consentement	ເນ໋ະ (marque d'interrogation)

Catégorie	Sous-catégorie	Exemples
	Adverbe modal postverbal interrogatif incisif	ເລີ່ອະ (marque d'interrogation), ກໍ (marque d'interrogation)
	Adverbe modal postverbal interrogatif amène	ນໍ (marque d'interrogation)
	Adverbe modal postverbal d'auto-interrogation	ໝໍ (marque d'interrogation)
	Adverbe modal postverbal interrogatif hypothétique	ສະບໍ (marque d'interrogation)
	Adverbe modal postverbal interrogatif d'objet	ຫຽ (que, quoi ?), ສ (que, quoi ?), ...
Adverbe modal postverbal du volitif et de l'exhortatif	Adverbe modal postverbal volitif	ດອ (marque du volitif ?), + ດອນ, ດອຕິ, ສະດອ
	Adverbe modal postverbal exhortatif pressant	ເທອະ (marque de l'exhortatif ?)
Adverbe modal postverbal de l'impératif	Adverbe modal postverbal impératif de relâchement	ຊ (marque du relachement ?), ...
	Adverbe modal postverbal impératif d'invite	ແມ (marque de l'impératif), ແມ້ (marque de l'impératif)
	Adverbe modal postverbal impératif ou affirmatif de provocation	ແໝ (marque de l'impératif)
	Adverbe modal postverbal impératif de sollicitation	ດຸ (marque de l'impératif), ດູ (marque de l'impératif)
	Adverbe modal postverbal impératif de recommandation	ເປ (marque de l'impératif), ເປ້ (marque de l'impératif), ເດ້ (marque de l'impératif)
	Adverbe modal postverbal impératif de courtoisie	ແດ່ (marque de l'impératif)
	Adverbe modal postverbal impératif de voeu	ທ້ອນ (marque de l'impératif), ເກີດ (marque de l'impératif)
Adverbe modal préverbal du conditionnel	Adverbe modal préverbal du conditionnel potentiel	ຄັນ (si, quand), ກ້ (si, quand)
	Adverbe modal préverbal du conditionnel d'implication	ຫ (si (naturellement))
	Adverbe modal préverbal du conditionnel potentiel renforcé	ເກ່ (même si), ແມນ່ (même si), ...
	Adverbe modal préverbal du conditionnel potentiel de possibilité	ອດຈະ (pouvoir éventuellement)
Adverbe modal postverbal du subjonctif dubitatif	Adverbe modal postverbal du subjonctif dubitatif	ບຸ (marque du dubitatif ?)
Adverbe quantitatif	Adverbe quantitatif indéfini superlatif	ຫຼ (beaucoup), ໜ້ອຍ (peu), ໝົດ (entièrement, tout), ພໍ (suffisamment), ຄົບ (complètement), ເຕັມ (pleinement)
	Adverbe quantitatif indéfini superlatif absolu	ເຫຼືອ (trop, excédant), ໂຜດ (trop, excessivement), ແທ້ (vraiment), ເກີນ (exagérément), ກ (au-delà), ທະວີຂຶ້ນ (plus, en augmentation), ທີ່ສຸດ (extrêmement), ຍິ່ (très, extrêmement)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Adverbe quantitatif indéfini augmentatif préverbal	ແຮ່ງ (de degré supérieur), ແສນ (extrêmement)
	Adverbe quantitatif indéfini superlatif progressif	ຊັກໃຊ້ (de plus en plus)
	Adverbe quantitatif indéfini réitératif	ๆ (signe de répétition augmentatif p.e. ๆๆນๆ =très joli ou diminutif p.e. ເຈັບๆ =légèrement douloureux)
	Adverbe quantitatif indéfini de répétition ou d'ajout	ຊ້ຳ (en plus), ອີກ (de nouveau), ຕື່ມ (encore), ໃໝ່ (de nouveau)
Adverbe conjonctif	Adverbe conjonctif de cause et de conséquence	ກໍ (marque de la cause), ..., ຈຶ່ງ (marque de la conséquence), ...
	Adverbe conjonctif de temps	ລົດ (marque du temps ?), ເລີຍ (marque du temps ?), ທັງ (marque du temps ?)
Adverbe de restriction	Adverbe de restriction	ແຕ່ (ne ... que ...),
Nom entrant dans la constitution de locutions prédicatives		
Nom entrant dans la constitution de locutions prédicatives	Nom entrant dans la constitution de locutions adjectivo-adverbiales qualificatives	ຂີ້ (formation des locutions péjoratives), ໃຈ (formation des locutions affectives ou morales), ເປັນໜ້າ / ເປັນຕາ (formation des locutions d'appréciation), ສີ / ເປັນສີ (formation des locutions d'impression nuancée)
Mots marquant le nombre et le genre		
Collectif	Collectif général	ໝູ່ (troupe), ຝູງ (troupe)
	Collectif de noms de personnes (emploi populaire)	ໝວດ (groupe)
	Collectif de noms de personnes et de pronoms	ພວກ (groupe),
	Collectif général pour les êtres et les choses	ບັນດາ (collectif),
	Collectif pour les personnes	ຈໍາພວກ (sous-groupe),
	Collectif pour les objets	ຈຸ (collectif),
	Collectif pour les pronoms	ຕຸ (collectif),
	Collectif pour les personnes	ຄະນະ (collectif),
	Collectif pour les personnes d'un même cercle	ຊາວ (collectif),
	Collectif pour les objets ou les personnes	ເຫຼົ່າ (collectif),
Qualificatif	Qualificatif pour les personnes	ຍິງ (féminin), ຊາຍ (masculin)
	Qualificatif pour les animaux et certains objets qui s'emboîtent	ແມ່ (femelle), ຜູ້ (male)
	Qualificatif pour les grands animaux	ພັງ (femelle d'éléphant), ສີດໍ (male d'éléphant sans défense), ສານ (male d'éléphant à défense)

Groupe des prépositions (ຄໍາບຸບພະບົດ / ຄໍາບຸພບົດ)
Ce groupe contient les prépositions :
« *mot qui, placé devant un complément, explicite le rapport entre celui-ci et l'élément complété* ».

Catégorie	Sous-catégorie	Exemples
Préposition	Préposition établissant un rapport de temps et d'espace	ຈົນ (jusqu'à : aboutissement), ເທິ້ງ (jusqu'à : aboutissement), ເກີ່ງ (jusqu'à : aboutissement), ຮອດ (jusqu'à : aboutissement), ທັນ (à : atteinte, rattrapage), ໃນ (dans, pendant), ຕາມ (dans, parmi, le long de, selon, suivant, en à), ນອກ (hors de, en dehors de, à l'extérieur de), ກາງ (au milieu de, au centre de, dans, en), ເທິງ (sur, en haut de, au-dessus de, à, dans), ເທັ້ງ (sur, en haut de, au-dessus de, à, dans), ລຸ່ມ (au-dessous de, en bas de), ຂ້າງ=ກໍ້ າ, ເປີ້ ອງ, ທາງ, ຟາກ, ດ້ ານ, ສ ່ ຽງ (à, du côté, vers), ໜ ້ າ (devant), ຫ ຼ ັ ງ (derrière), ລຸ ນ (après), ກ ່ ອນ (avant), ໃກ ້ =ນ ໍ ່ , ກ ິ ດ, ຊ ິ ດ, ແປະ (près de), ໄ ກ (loin de)
	Préposition établissant un rapport de position et de destination	ຕ ໍ ່ (à), ຕ ໍ ່ ໜ ້ າ (vis-à-vis de, en face de, devant)
	Préposition établissant un rapport de voie, de moyen	ທ າ ງ (par, en)
	Préposition établissant un rapport de place, de mise	ໃ ສ ່ (à, sur, dans)
	Préposition établissant un rapport de cheminement, de but	ຫ າ (à, vers)
	Préposition établissant un rapport d'origine	ແ ຕ ່ (de), ຕ ັ ້ ງ ແ ຕ ່ (depuis), ເ ວ ັ ້ ນ ແ ຕ ່ (sauf, hormis, à l'exception de), ຈ າ ກ (de), ນ ອ ກ ຈ າ ກ (en dehors de), ຕ ໍ ່ ຈ າ ກ (à la suite de), ເ ນ ື ່ ອ ງ ຈ າ ກ (à propos de), ຫ ຼ ັ ງ ຈ າ ກ (à la suite de, après)
	Préposition établissant un rapport d'accompagnement, d'objet	ນ ໍ າ (avec, à, de), ກ ັ ບ (avec, à), ພ ້ ອ ມ ou ທ ັ ້ ງ (avec), ດ ອ ມ (avec ; archaïque)
	Préposition établissant un rapport de but, d'attribution	ໃ ຫ ້ (à, pour, afin de), ເ ພ ື ່ ອ (à, pour, afin de)
	Préposition établissant un rapport de contact, d'attribution	ສ ູ ່ (à), ແ ກ ່ ou ແ ດ ່ (à, envers ; archaïque), ຍ ັ ງ (à)
	Préposition établissant un rapport de cause	ຍ ້ ອ ນ ou ຄ ອ ບ (à cause de, grâce à, par)
	Préposition établissant un rapport de moyen	ດ ້ ວ ຍ (à, en, par)
	Préposition établissant un rapport de moyen, d'agent, de manière	ໂ ດ ຍ (par, avec), ໂ ດ ຍ ດ ັ ່ ງ (selon ; archaïque)
	Préposition établissant un rapport d'appartenance	ຂ ອ ງ (de, à)

Annexe 8 Parties du discours, avec exemples, du laotien (niveau sous-catégorie)

Catégorie	Sous-catégorie	Exemples
	Préposition établissant un rapport de réserve, de garde, de destination	ໄວ້ (pour), ສຳລັບ ou ສຳຫຼັບ (pour), ເພື່ອ (à, pour, afin de)
	Préposition établissant un rapport d'établissement, de translation	ເປັນ (en, comme)
	Préposition établissant un rapport de discrimination, de proportion	ສ່ວນ (quant à, pour)
	Préposition établissant un rapport de distribution	ລະ (à, par)
	Préposition établissant un rapport d'évaluation	ປະມານ (aux environs de, environ, vers), ລະຫວ່າງ (environ)
	Préposition établissant un rapport de manière	ຢ່າງ (de manière à, da la manière)
	Préposition établissant un rapport de remplacement	ຕາງ (à la place de)
	Préposition établissant un rapport d'opposition	ກົງກັນຂ້າມ (à l'opposé de)
	Préposition établissant un rapport d'exception	ເວັ້ນແຕ່ ou ນອກຈາກ (sauf, excepté)

Groupe des conjonctions (ຄຳສັນທານ)

Ce groupe contient les conjonctions :

« *mot servant à réunir deux mots, deux groupes de mots ou des propositions de même nature (conjonction de coordination) ou à relier une proposition subordonnée à une principale (conjonction de subordination)* ».

Catégorie	Sous-catégorie	Exemples
Conjonction de coordination	Conjonction de coordination	ແລະ (et), ຫຼື (ou), ແຕ່ (mais), ເພາະ (car), ສະນັ້ນ (donc), ບໍ່ສະນັ້ ນ (sinon), ຄື=ປານ, ເໝືອນ, ດັ່ງ, ສົມ (comme)
	Adverbe conjonctif	ກໍ (alors), ... ຈຶ່ງ (alors), ...
Conjonction de subordination	Conjonction de subordination	ວ່າ (que), ໃຫ້ (que), ຄັນ (si), ກ້າ (si), ຫາກ (si), ເມື່ອ (quand), ເວລາ (quand)
	Locution conjonctive de subordination	ເພາະວ່າ (parce que), ດ້ວຍວ່າ (du fait que), ຍ້ອນວ່າ (grâce à quoi, puisque), ເຫັນວ່າ (vu que), ເຖິງວ່າ (même si), ຄືວ່າ (il semble que), ເພື່ອໃຫ້ (pour que), ຈົນໃຫ້ (jusqu'à ce que)

Groupe des interjections (ຄໍາອຸທານ)
Ce groupe contient les interjections : « *mot isolé qui exprime un sentiment violent, une émotion ou un ordre* »,
et les exclamations : « *mot ou phrase exprimant une émotion vive ou un jugement affectif* ».

Catégorie	Sous-catégorie	Exemples
Interjectif direct	Titre social courant	ທ່ານຍິງແລະທ່ານຊາຍ (mesdames, messieurs), ແມ່ (maman !)
	Expression archaïque	ຂ້ານ້ອມ(ແຕ)ພໍ່ເປັນເຈົ້າ (mon seigneur !), ເຈົ້າ (seigneur !), ...
	Interjectif spécifique	ເອີ້ຍ (hé ! : appel direct), ນີ້ນະ (dîtes ! : pour attirer l'attention), ສາທຸ (ô ... ! : évocation divine)
Interjectif indirect	Titre social courant et pronom personnel	ຍາແມ່ dans ສໍາບາຍຍາແມ່ (bonjour maman), ລູກ dans ສໍາບາຍຍາລູກ (bonjour mon enfant), ຂະນ້ອຍ dans ແມ່ນແລ້ວ ຂະນ້ອຍ (c'est vrai, monsieur ou madame : archaïque ?), ...
	Interjectif spécifique	ເອີຍ (ô ...), ອຶ (! : de colère, de dépit, de provocation)
Exclamatif	Terme simple	ໂອ (ha : surprise, réserve), ໂອລະນໍ (introduction à un chant), ໂອຍ (ah ! : surprise), ໂອ່ຍ (aïe : douleur), ອຸຍ (ah ! : frayeur, répulsion), ເນ (là ! : démonstratif), ແນ (là ! : démonstratif), ອ້ວຍ (eh bien ! : surprise), ອ້ວາຍ (eh bien ! : surprise), ໂຍ (oh ! : réprobation), ...
	Expression	ກະນັ້ນລະຕີ (c'est cela même !), ຕາຍກູ (je suis perdu), ແລ້ວກັນ (c'est fichu), ເຕຊິ່ເຕບັ (formidable !), ໂຮ່ຮິ (hourra !), ເຍີ້ກ (invite à agir ensemble), ເຍີ້ກປີ້ (attaquons ferme : cri scandé par les piroguiers à la course)

A.9 SCHÉMA XML PAPILLON POUR LE LAOTIEN

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!-- XML Schema for the Papillon Lao lexies
Includes the language specific elements (parts-of-speech, etc.)
Namespace: http://www-clips.imag.fr/geta/services/dml
Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd
$Author: Vincent $ Vincent BERMENT Vincent.Berment@imag.fr
$Date: 2003/07/04 02:39:32 $
$Revision: 1.0 $ -->
<schema
  xmlns:d='http://www-clips.imag.fr/geta/services/dml'
  xmlns='http://www.w3.org/2001/XMLSchema'
  xmlns:xlink='http://www.w3.org/1999/xlink'
  targetNamespace='http://www-clips.imag.fr/geta/services/dml'>
<annotation>
  <documentation xml:lang="en">
    XML Schema for the Papillon Lao lexies.
    Includes the language specific elements (parts-of-speech, etc.).
    Namespace: http://www-clips.imag.fr/geta/services/dml.
    Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd.
    Schema location:
http://www-clips.imag.fr/geta/services/dml/papillon_lao.xsd.
  </documentation>
</annotation>
<!--======= Elements of the Papillon common schema redefinition ======-->
<redefine schemaLocation="http://www-clips.imag.fr/geta/services/dml/papillon.xsd">
 <!— ** The parts-of-speech we will use for the Lao language ** -->
 <simpleType name='posType'>
  <!-- Examples are proposed in Vincent Berment's paper for Papillon 03 -->
  <restriction base='d:posType'>
    <!-- Noms -->
    <!-- Nom commun -->
        <enumeration value="nc"/>
    <!-- Locution nominale -->
        <enumeration value="locn"/>
    <!-- Nom propre -->
        <enumeration value="np"/>
    <!-- Générique de classification -->
        <enumeration value="gcl"/>
    <!-- Générique d'abstraction -->
        <enumeration value="ga"/>
    <!-- Générique de réalisation -->
        <enumeration value="gr"/>
    <!-- Générique de concrétisation -->
        <enumeration value="gco"/>
    <!-- Générique de titre -->
        <enumeration value="gt"/>
    <!-- Pronoms -->
    <!-- Pronom personnel -->
        <enumeration value="pp"/>
    <!-- Pronom impersonnel neutre -->
        <enumeration value="pin"/>
    <!-- Pronom réciproque -->
        <enumeration value="prq"/>
    <!-- Pronom réfléchi -->
        <enumeration value="prf"/>
    <!-- Pronom indéfini-interrogatif -->
        <enumeration value="pii"/>
    <!-- Locution pronominale indéfinie-interrogative -->
        <enumeration value="lpii"/>
    <!-- Pronom relatif -->
        <enumeration value="prl"/>
    <!-- Spécificatif -->
        <enumeration value="spec"/>
```

```
<!-- Adjectif indéfini -->
    <enumeration value="ai"/>
<!-- Locution adjectivale indéfinie -->
    <enumeration value="lai"/>
<!-- Adnominal -->
    <enumeration value="adn"/>
<!-- Verbes -->
<!-- Verbe d'action -->
    <enumeration value="va"/>
<!-- Verbe d'état -->
    <enumeration value="ve"/>
<!-- Verbe désidératif -->
    <enumeration value="vd"/>
<!-- Auxiliaire de voie passive -->
    <enumeration value="avp"/>
<!-- Auxiliaire de temps -->
    <enumeration value="at"/>
<!-- Auxiliaire d'aspect -->
    <enumeration value="aa"/>
<!-- Auxiliaire secondaire -->
    <enumeration value="as"/>
<!-- Auxiliaire modal de l'impératif -->
    <enumeration value="ami"/>
<!-- Auxiliaire modal du subjonctif dubitatif -->
    <enumeration value="amsd"/>
<!-- Auxiliaire royal -->
    <enumeration value="ar"/>
<!-- Predicatifs -->
<!-- Adjectif-adverbe qualificatif -->
    <enumeration value="aaq"/>
<!-- Locution adjectivo-adverbiale qualificative -->
    <enumeration value="laaq"/>
<!-- Adjectif-adverbe quantitatif indéfini -->
    <enumeration value="aaqi"/>
<!-- Adjectif démonstratif -->
    <enumeration value="adem"/>
<!-- Adjectif interrogatif -->
    <enumeration value="aint"/>
<!-- Adjectif indéfini -->
    <enumeration value="aind"/>
<!-- Adjectif numéral -->
    <enumeration value="anum"/>
<!-- Locution adjectivale numérale -->
    <enumeration value="lan"/>
<!-- Adverbe d'affirmation -->
    <enumeration value="aaff"/>
<!-- Adverbe de négation -->
    <enumeration value="aneg"/>
<!-- Adverbe modal postverbal de l'affirmatif -->
    <enumeration value="ampa"/>
<!-- Adverbe modal postverbal de l'interrogatif -->
    <enumeration value="locn"/>
<!-- Adverbe modal postverbal du volitif et de l'exhortatif -->
    <enumeration value="ampve"/>
<!-- Adverbe modal postverbal de l'impératif -->
    <enumeration value="ampi"/>
<!-- Adverbe modal préverbal du conditionnel -->
    <enumeration value="ampc"/>
<!-- Adverbe modal postverbal du subjonctif dubitatif -->
    <enumeration value="ampsd"/>
<!-- Adverbe quantitatif -->
    <enumeration value="aquan"/>
<!-- Adverbe conjonctif -->
    <enumeration value="acon"/>
<!-- Adverbe de restriction -->
    <enumeration value="arestr"/>
<!-- Nom entrant dans la constitution de locutions prédicatives -->
    <enumeration value="neclp"/>
```

```
    <!-- Collectif -->
        <enumeration value="coll"/>
    <!-- Qualificatif -->
        <enumeration value="qual"/>
    <!-- Prépositions -->
    <!-- Préposition -->
        <enumeration value="prep"/>
    <!-- Conjonctions -->
    <!-- Conjonction de coordination -->
        <enumeration value="ccoord"/>
    <!-- Conjonction de subordination -->
        <enumeration value="csubord"/>
    <!-- Interjections -->
    <!-- Interjectif direct -->
        <enumeration value="idir"/>
    <!-- Interjectif indirect -->
        <enumeration value="iindir"/>
    <!-- Exclamatif -->
        <enumeration value="excl"/>
   </restriction>
  </simpleType>
 <!— ** The levels of language we will use for the Lao language ** -->
 <simpleType name="leveloflanguageType">
   <restriction base="d:leveloflanguageType">
    <enumeration value="courant" />
    <enumeration value="respectueux" />
    <enumeration value="familier" />
    <enumeration value="argotique" />
    <enumeration value="spécialisé" />
    <enumeration value="recherché" />
    <enumeration value="bonze" />
    <enumeration value="royal" />
    <enumeration value="littéraire" />
    <enumeration value="parlé" />
    <enumeration value="archaïque" />
   </restriction>
  </simpleType>
</redefine>
</schema>
```

A.10 EXEMPLES D'ARTICLES (LEXIES) DE LAODICT

Attribut	Valeur
Lexie	ດີ
Orthographes anciennes	
Orthographes alternatives	
Définition (laotien)	ບໍ່ຊົ່ວ,ລັກສະນະອັນຈົບງາມ ປະພຶດກິກຕ້ອງ
Définition (traduction en français)	caractéristique de ce qui est joli ou qui se comporte correctement
Partie du discours	Adjectif-adverbe qualificatif (prédicatif généralisé)
Spécificatif (noms quantifiables)	
Traduction	gentil, gentille, bon, bonne (adj.), bien, gentiment (adv.), être gentil (verbe)
Traduction mot unique	gentil (adj.), gentiment (adv.), être gentil (verbe)
Exemples (laotien)	ນິໄສດີ,ຄົນຜູ້ນີ້ດີ
Exemples (traduction en français)	bon comportement, cette personne est gentille
Exemple court (laotien)	ນິໄສດີ,ຄົນຜູ້ນີ້ດີ
Exemple court (traduction en français)	bon comportement, cette personne est gentille
Expression idiomatique (laotien)	ທຳດີໄດ້ດີ
Expression idiomatique (traduction en français)	en faisant le bien on reçoit le bien
Commentaires	

Attribut	Valeur
Lexie	ຢູ່
Orthographes anciennes	
Orthographes alternatives	
Définition (laotien)	ອາໄສ, ບໍ່ໄປຈາກບ່ອນພັກ
Définition (traduction en français)	demeurer, habiter
Partie du discours	Verbe d'état
Spécificatif (noms quantifiables)	
Traduction	être, habiter, demeurer
Traduction mot unique	habiter
Exemples (laotien)	ຂ້ອຍຢູ່ເຮືອນ
Exemples (traduction en français)	Je suis à la maison
Exemple court (laotien)	ຂ້ອຍຢູ່ເຮືອນ
Exemple court (traduction en français)	Je suis à la maison
Expression idiomatique (laotien)	ບໍ່ໄປຈາກບ່ອນພັກ
Expression idiomatique (traduction en français)	habiter
Commentaires	

Attribut	**Valeur**
Lexie	ຢູ່
Orthographes anciennes	
Orthographes alternatives	
Définition (laotien)	ກິຣິຍາທີ່ກຳລັງສືບຕໍ່
Définition (traduction en français)	marque l'action qui continue
Partie du discours	Auxiliaire d'aspect
Spécificatif (noms quantifiables)	
Traduction	gramm. duratif
Traduction mot unique	
Exemples (laotien)	ລາວນອນຢູ່
Exemples (traduction en français)	Il dort encore
Exemple court (laotien)	ລາວນອນຢູ່
Exemple court (traduction en français)	Il dort encore
Expression idiomatique (laotien)	
Expression idiomatique (traduction en français)	
Commentaires	Syntagme postverbal de procès qui dure

Attribut	**Valeur**
Lexie	ປາ
Orthographes anciennes	
Orthographes alternatives	
Définition (laotien)	ຊື່ລວມຂອງສັດຢູ່ນ້ຳ.
Définition (traduction en français)	Nom générique des animaux vivant dans l'eau.
Partie du discours	Générique de classification
Spécificatif (noms quantifiables)	ໂຕ
Traduction	poisson
Traduction mot unique	poisson
Exemples (laotien)	ມີປາຫຼາຍໃນນ້ຳຂອງ
Exemples (traduction en français)	Il y a beaucoup de poissons dans le Mékong.
Exemple court (laotien)	ມີປາຫຼາຍໃນນ້ຳຂອງ
Exemple court (traduction en français)	Il y a beaucoup de poissons dans le Mékong.
Expression idiomatique (laotien)	ໃນນ້ຳມີປາໃນນາມີເຂົ້າ
Expression idiomatique (traduction en français)	À l'eau les poissons, à la rizière le riz.
Commentaires	

Attribut	**Valeur**
Lexie	ກະໂປ່ງ
Orthographes anciennes	
Orthographes alternatives	
Définition (laotien)	ສິ້ນນຸ່ງແບບເອີຣົບ
Définition (traduction en français)	Robe de style européen
Partie du discours	Nom commun
Spécificatif (noms quantifiables)	ໂຕ
Traduction	robe de style occidental
Traduction mot unique	robe
Exemples (laotien)	ລາວນຸ່ງກະໂປ່ງງາມ
Exemples (traduction en français)	Elle porte une belle robe
Exemple court (laotien)	ລາວນຸ່ງກະໂປ່ງງາມ
Exemple court (traduction en français)	Elle porte une belle robe
Expression idiomatique (laotien)	
Expression idiomatique (traduction en français)	
Commentaires	

A.11 GRAMMAIRE DES GRAMMAIRES SYLLABIQUES

La grammaire ci-dessous est utilisée par le compilateur de grammaire saint-jean (Del Vigna 2000) pour générer le compilateur des grammaires de syllabes.

```
nom cGGSyllabes;

// Grammaire : 'RÈGLE_1 RÈGLES_SUIVANTES *'.
axiome = syllabes . reecritures ;
reecritures = reecriture . R1 + {};
R1 = reecritures + {};

// Règle 1 : 'syllabes = POLYNÔME * ;'.
(syllabes) = gaucheSyl . {=} . droite * {;} ;
(gaucheSyl) = {syllabes};

// Règles suivantes : 'PARTIE_GAUCHE = POLYNÔME * ;'.
 (reecriture) = gauche . {=} . droite * {;} ;

// Partie gauche : Non terminal.
 (gauche) = nonTerminal ;

// Polynôme : Disjonction de termes (séparés par '+').
 (droite) = polynome;
polynome = terme P1;
P1 = . plus P1  +  {};
(plus)     = {+} . terme;

// Terme : Conjonction de facteurs (juxtaposés).
terme    = facteur  T1;
T1 = conc  T1  +  {};
(conc) = * facteur;

// Facteur : Non terminal, terminal ou polynôme entre parenthèses.
facteur    = nonTerminal + terminal + {(} . polynome . {)};

// Non terminal : Chaîne d'au moins un caractère pouvant contenir
   des lettres de l'alphabet, des chiffres et les caractères - et _.
 (nonTerminal)   = [A..Z,a..z,0..9,-,_]~+;

// Terminal :
• Soit une chaîne ne contenant pas d'espace, précédée par ':'[1] et
éventuellement par des espaces[2],
• Soit la chaîne vide « {} ».
terminal = {:} [#32,#160]~* [#33..#255]~+ (terminal) + {{} []}~1 {}
(terminal) ;
/
```

[1] Le caractère ':' permet d'identifier le début d'un terminal.

[2] Ces espaces permettent d'éviter que les terminaux sans avance (voir II.1.2.2) se mêlent aux caractères précédents (lisibilité).

A.12 ARTICLE « AMBIGUÏTÉS IRRÉDUCTIBLES DANS LES MONOÏDES DE MOTS »

Claude DEL VIGNA
CAMS (1) – Paris
delvigna@ivry.cnrs.fr

Vincent BERMENT
GETA (2) – Grenoble
Vincent.Berment@imag.fr

RÉSUMÉ – Le point de départ de l'étude présentée ici est cette « malice » de certaines langues du Sud-Est asiatique qui s'écrivent sans que des espaces séparent les mots. Les traitements automatiques de ces langues s'en trouvent compliqués d'autant que, dès le premier niveau, celui des syllabes, le découpage des textes n'est en général pas unique. Autrement dit, rapporté à la combinatoire des mots, le système syllabique de ces langues n'est pas un code. On s'intéresse ici à l'origine des ambiguïtés de découpage, plus précisément au recensement de celles qu'on appelle irréductibles, en ce sens qu'elles sont à l'origine de toutes les autres. On montre que le langage des ambiguïtés irréductibles est rationnel et on présente le moyen d'en calculer une expression régulière en l'étayant de l'expérience de son application à la langue laotienne.

INTRODUCTION

Parmi les langues de l'Asie du Sud-Est, vingt à trente d'entre elles s'écrivent sans que les mots soient séparés par des espaces. C'est le cas du khmer, du thaï, du laotien et du birman pour ne citer qu'elles. Le texte laotien ci-contre en est l'illustration. Le traitement automatique de ces langues se complique de cette caractéristique. Ainsi, dès le niveau syllabique, le découpage d'un texte n'est-il en général pas unique. Formellement, pour une langue donnée, l'ensemble des concaténations ambiguës de syllabes est un idéal (3) du monoïde sur les syllabes de cette langue. L'article établit que cet idéal est engendrable par un sous-langage rationnel, dit des ambiguïtés irréductibles, dont il propose une méthode pour en calculer une expression régulière. Aussi, la finitude de ce sous-langage est-elle une question décidable. En termes plus linguistiques, l'article montre la possibilité de recenser de façon exhaustive les suites ambiguës de syllabes et comment les construire toutes à partir des motifs de base que sont les ambiguïtés irréductibles.

ບ້ານຂ້ອຍມີໂຮງຮຽນຫຼັງນຶ່ງ

Dans mon village, il y a une école.

Le travail présenté ici vient à la suite de la réalisation logicielle *LaoWord* (Berment, 1998, 2002), (*LaoWord*, 1998). *LaoWord*, qui s'inscrit dans ce qu'il est convenu d'appeler l'*informatisation des langues minoritaires* (4), est une librairie dynamique qui adapte le traitement de texte *Word* de Microsoft au laotien. Elle est fondée sur les syllabes en ce sens que lorsqu'on double-clique sur un caractère, la syllabe autour de ce caractère est sélectionnée. Lorsque plusieurs syllabes sont candidates, l'examen de la syllabe à droite et de celle à gauche suffit, en général, pour choisir l'une d'elles. C'est dans le but de cerner formellement cette technique heuristique, qu'on s'intéresse ici au recensement des ambiguïtés propres à un dictionnaire de syllabes.

---

(1) Centre d'Analyse et de Mathématique Sociales, CNRS, Paris, **delvigna@ivry.cnrs.fr**.

(2) Groupe d'Etude pour la Traduction Automatique, Grenoble, **Vincent.Berment@imag.fr**.

(3) Un idéal d'un monoïde M est un sous-ensemble I de M tel que $M\ I\ M \subseteq I$ (Berstel & Perrin, 1985).

(4) Voir le site http://isl.ntflex.uni-lj.si/SALTMIL/

L'étude est naturellement liée à la théorie des codes (Berstel & Perrin, 1985), (Lothaire, 2002). Mais « en creux », puisqu'aucun des dictionnaires des syllabes des langues concernées n'est un code et qu'on s'intéresse précisément ici aux raisons qui font qu'ils n'en sont pas. Ceci étant, la méthode présentée relève de la famille des algorithmes qui permettent de tester si un dictionnaire est ou n'est pas un code, dont celui de Sardinas et Patterson, 1953, (Berstel & Perrin, 1985), et celui de Spehner, 1975, (Lallement, 1979). On souligne, à propos du dernier, que la notion d'équation irréductible qu'il utilise correspond, au niveau découpage, à celle d'ambiguïté irréductible. On note enfin la parenté entre la notion de recouvrement sur laquelle se fonde la méthode présentée ici et celle de domino développée dans (Weber & Head, 1994) et (Head & Weber, 1995) pour les codes MSD [1].

1. AMBIGUÏTÉS IRRÉDUCTIBLES

Etant donné un alphabet fini V, on note V^* le monoïde sur V, ε la chaîne vide, V^+ l'ensemble $V^* - \{\varepsilon\}$, $|x|$ la longueur d'une chaîne x sur V et, si A et B sont des langages sur V, $A\,B$ leur langage produit (ou concaténation). On appelle *dictionnaire* tout ensemble fini non vide W de V^+ . Les éléments de W sont appelés *mots*. Ce sont les correspondants formels des syllabes.

Si $w = (w_1, w_2, \ldots, w_n)$, $n \geq 1$, est une suite non vide de mots, on note $\gamma(w)$ la concaténation $w_1\, w_2 \ldots w_n$. De plus, si la suite w est vide, on pose $\gamma(w) = \varepsilon$. On désigne par W^* le sous-monoïde $\{\gamma(w) \mid w \text{ est une suite sur } W\}$, dit *monoïde de mots*, et par W^+ la différence $W^* - \{\varepsilon\}$. Inversement, pour tout p dans W^*, une W-*factorisation* de p est une suite w sur W telle que $p = \gamma(w)$.

Un élément p de W^+ est *ambigu* s'il admet deux W-factorisations différentes. On note $A(W)$, plus simplement A, le sous-ensemble des éléments ambigus de W^+. Un élément p de A est dit *ambigu irréductible* s'il ne peut se factoriser d'aucune des manières $x\,p'\,y$, $x\,p'$ et $p'\,y$ dans lesquelles p' est dans A, x et y dans W^+. On note $AI(W)$, plus simplement AI, l'ensemble des éléments ambigus irréductibles.

LEMME 1.A – Etant donné un dictionnaire W sur un alphabet V, on a :

(1.1a) A est un idéal de W^*, i.e. $W^*\,A\,W^* \subseteq A$;

(1.1b) $A = W^*\,AI\,W^*$;

(1.1c) $AI = A - (W^+\,A \cup A\,W^+ \cup W^+A\,W^+)$. □

Exemple 1 – Le dictionnaire $W = \{ cv, cvc, ccv, ccvc \}$ est celui des syllabes du laotien, rapportées aux deux caractères c (pour consonne) et v (pour voyelle) (Berment, 1997). L'ensemble AI correspondant est fini et égal à $\{ cvccv, cvccvc, ccvccv, ccvccvc \}$ et toute concaténation ambiguë d'éléments de W contient nécessairement un de ces motifs.

Exemple 2 – Si $W = \{ aa, aabb, bbc, cc, cdd, dd \}$, AI est infini, égal au langage rationnel que décrit l'expression régulière $aabbcc^*dd$.

---

[1] Un code MSD (*multiset decipherable*) est un dictionnaire W pour lequel toutes les factorisations d'un quelconque élément de W^+ en éléments de W conduisent au même tableau de fréquences.

2. RECOUVREMENTS ET FEUILLURES

Un *recouvrement* sur un dictionnaire W est un quadruplet (a, b, x, y) dans lequel a et b sont des suites non vides d'éléments de W, x et y des chaînes sur V et tel que

(2.1a) $x\,\gamma(b) = \gamma(a)\,y$

(2.1b) $|\gamma(a)| + |\gamma(b)| > |x| + |y|$.

L'inégalité stricte de la condition (2.1b) impose que a et b se chevauchent. Les chaînes x et y sont les *feuillures* du recouvrement, respectivement *gauche* et *droite*. La figure 1 est celle d'un recouvrement.

LEMME 2.A – Si deux recouvrements r et r' sur un dictionnaire W sont tels que la feuillure droite de r est égale à la feuillure gauche de r', i.e. si $r = (a, b, x, z)$ et $r' = (a', b', z, y)$, alors le quadruplet $(a\,a', b\,b', x, y)$ est un recouvrement sur W. □

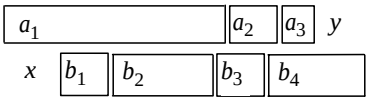


Fig. 1 : recouvrement

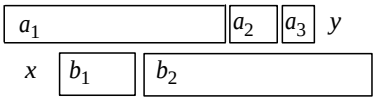


Fig. 2 : 1-recouvrement

Eu égard au lemme précédent, deux recouvrements r et r' de la forme $r = (a, b, x, z)$ et $r' = (a', b', z, y)$, sont dits *emboîtables* et le recouvrement $(a\,a', b\,b', x, y)$ est appelé leur *emboîtement*, noté $r \circ r'$. L'opération s'étend canoniquement à des ensembles de recouvrements.

On appelle 1-*recouvrement* tout recouvrement (a, b, x, y) dans lequel $a = (a_i\,;\, 1 \le i \le m)$ et $b = (b_j\,;\, 1 \le j \le n)$, et tel que

(2.2a) le quadruplet $((a_1), (b_n), x\,\gamma(b_1, \ldots, b_{n-1})), \gamma(a_2, \ldots, a_m)\,y)$ est un recouvrement,

(2.2b) $(n \ne 1)$ <u>ou</u> $(m \ne 1)$ <u>ou</u> $(x\,y \ne \varepsilon)$.

La condition (2.2a) impose que le premier élément de la suite a chevauche le dernier de la suite b, tandis que la disjonction (2.2b) interdit les quadruplets de la forme $((c),(c),\varepsilon,\varepsilon)$, $c \in W$, dans lesquels les feuillures sont vides et les deux suites sont égales et de longueur 1. La figure 2 est celle d'un 1-recouvrement. On note $R^1(W)$, ou R^1, l'ensemble des 1-recouvrements du dictionnaire W et F^1 l'ensemble de leurs feuillures. Lorsque $W = \{ aa, aabb, bbc, cc, cdd, dd \}$, l'ensemble R^1 est celui de la figure 3 et $F^1 = \{ \varepsilon, a, aa, aab, abb, bb, bc, c, cd, d, dd \}$.

LEMME 2.B – L'ensemble R^1 des 1-recouvrements d'un dictionnaire W et celui F^1 de leurs feuillures sont finis. □

((*aa*), (*aa*), *a*, *a*)	((*aabb*), (*bbc*), *aab*, *bc*)	((*bbc*), (*cdd*), *bb*, *dd*)	((*cc*, *dd*), (*cdd*), *c*, ε)
((*aa*), (*aabb*), ε, *bb*)	((*cc*), (*cc*), *c*, *c*)	((*cdd*), (*dd*), *c*, ε)	((*bbc*, *dd*), (*cdd*), *bb*, ε)
((*aa*), (*aabb*), *a*, *abb*)	((*cc*), (*cdd*), *c*, *dd*)	((*cdd*), (*dd*), *cd*, *d*)	((*aabb*), (*aa*, *bbc*), ε, *c*)
((*aabb*), (*bbc*), *aa*, *c*)	((*bbc*), (*cc*), *bb*, *c*)	((*dd*), (*dd*), *d*, *d*)	

Fig. 3 : 1-recouvrements de $W = \{ aa, aabb, bbc, cc, cdd, dd \}$

3. AMBIGUÏTÉS IRRÉDUCTIBLES ET 1-RECOUVREMENTS

À partir de l'ensemble F^1 des feuillures d'un dictionnaire W, on considère l'ensemble $\underline{F} = (F^1 - \{\varepsilon\}) \cup \{\varepsilon_g, \varepsilon_d\}$ obtenu en remplaçant la feuillure vide ε par deux éléments ε_g et ε_d n'appartenant pas à V^*, de façon à pouvoir distinguer dans la suite les cas où la feuillure vide est à gauche ou bien à droite. À l'ensemble R^1 des 1-recouvrements, on associe alors le plus petit sous-ensemble T du produit $\underline{F} \times R^1 \times \underline{F}$ tel que pour toutes suites non vides a et b de mots, pour tout $x \in F^1$, $x \neq \varepsilon$, et pour pour tout $y \in F^1$, $y \neq \varepsilon$,

(3.1a) $(a, b, x, y) \in R^1$, alors $(x, (a, b, x, y), y) \in T$;

(3.1b) $(a, b, x, \varepsilon) \in R^1$, alors $(x, (a, b, x, \varepsilon), \varepsilon_d) \in T$;

(3.1c) $(a, b, \varepsilon, y) \in R^1$, alors $(\in_g, (a, b, \varepsilon, y), y) \in T$ et $(\in_d, (a, b, \varepsilon, y), y) \in T$;

(3.1d) $(a, b, \varepsilon, \varepsilon) \in R^1$, alors $(\in_g, (a, b, \varepsilon, \varepsilon), \varepsilon_d) \in T$ et $(\varepsilon_d, (a, b, \varepsilon, \varepsilon), \varepsilon_d) \in T$.

Soit alors le quintuplet $(R^1, \underline{F}, \varepsilon_g, T, \varepsilon_d)$. Puisque R^1 et F^1 , et donc $\underline{F}$, sont finis (lemme 2.B), ce quintuplet est un automate d'états finis dont R^1 est l'alphabet, $\underline{F}$ l'ensemble des états, ε_g l'état initial, T la relation de transition et ε_d l'état final. Soit L le langage reconnu. Chaque élément l de L est une suite non vide de 1-recouvrements qui, par construction, sont emboîtables dans l'ordre de la suite et dont l'emboîtement « tombe juste » à gauche et à droite, autrement dit, est de la forme $(a, b, \varepsilon, \varepsilon)$. Soit $\underline{l}$ l'élément $\gamma(a)$, ou $\gamma(b)$ de W^+. On note K le langage $\{ \underline{l} \mid l \in L \}$ sur V.

On considère, par ailleurs, le quintuplet $(V, \underline{F}, \varepsilon_g, T', \varepsilon_d)$ dans lequel T' est le sous-ensemble $\{ (x, \gamma(b), y) \mid (x, (a, b, x, y), y) \in T \}$ du produit $\underline{F} \times V^* \times \underline{F}$. Ce quintuplet n'est pas *stricto sensu* un automate fini puisque T' n'est pas un sous-ensemble du produit $\underline{F} \times V \times \underline{F}$. Selon la méthode usuelle [1], on lui associe canoniquement un automate fini qu'on désigne par $AUT = (V, Q, \varepsilon_g, T'', \varepsilon_d)$. On fait remarquer que, par construction, l'automate $(R^1, \underline{F}, \varepsilon_g, T, \varepsilon_d)$ est déterministe – il est aussi minimal – alors que l'automate AUT, qui en dérive, ne l'est généralement pas.

LEMME 3.A – Pour tout dictionnaire W sur un alphabet V,

(3.2a) $K \subseteq A$;

(3.2b) l'automate AUT reconnaît le langage K ;

(3.2c) K est un langage rationnel. □

Compte tenu de la propriété (3.2b), on notera AUT_K l'automate AUT.

[1] La méthode consiste à appliquer récursivement la règle : pour chaque triplet $(x, c\,p, y)$ de T', $c \in V$, $p \in V^*$, créer un nouvel état q et les deux triplets (x, c, q) et (q, p, y).

LEMME 3.B – Pour tout dictionnaire W sur un alphabet V, $A = W^* K W^*$.

dem. [$\boldsymbol{A \subseteq W^* K W^*}$] Si $a = (a_1, a_2, \dots, a_m)$, $m \geq 1$, est une suite finie non vide sur W, on désigne par $\varphi_a(k)$ l'indice dans $\gamma(a)$ du premier caractère de a_k, $1 \leq k \leq m$, par $\lambda_a(k)$ l'indice de son dernier caractère et par $a[i:j]$, $1 \leq i \leq j \leq m$, la sous-suite $a_i, a_{i+1}, \dots, a_j$ de a. Enfin, si x est une chaîne sur V, $x[i:j]$ désigne le facteur de x qui commence à l'indice i et se termine à l'indice j, $1 \leq i \leq j \leq |x|$.

Soit $p \in A$. D'après la propriété (1.1b), p se factorise en $p = xp'y$, $x \in W^*$, $p' \in AI$ et $y \in W^*$. Il suffit donc de montrer que $p' \in K$ pour établir l'inclusion $A \subseteq W^*K W^*$ visée. Puisque $p' \in AI$, il existe deux suites différentes sur W, $a = (a_1, a_2, \dots, a_m)$, $m \geq 1$, et $b = (b_1, b_2, \dots, b_n)$, $n \geq 1$, telles que $p' = \gamma(a) = \gamma(b)$ et $\forall i, 1 < i \leq m$, $\forall j, 1 < j \leq n$, $\varphi_a(i) \neq \varphi_b(j)$. On suppose que $\lambda_a(1) > \lambda_b(1)$ – voir figure 4. À partir des suites a et b, on considère l'ensemble S des couples (i,j), $1 \leq i \leq m$, $1 \leq j \leq n$, tels que $\varphi_a(i) \leq \varphi_b(j) \leq \lambda_a(i) \leq \lambda_b(j)$. Soit s le cardinal de S. Dans la figure 4, $S = \{ (1,2), (3,4), (5,5) \}$ et ses éléments sont les segments obliques en trait double. On a $\forall (i,j) \in S$, $\forall (i',j') \in S$, $(i < i' \Leftrightarrow j < j')$, autrement dit les segments (i,j) ne « se croisent » jamais. On les numérote à partir de 1 en partant de la gauche et on note (i_k, j_k), $1 \leq k \leq s$, le k-ième. On a :

(1) $\forall k, 1 \leq k < s$, $\varphi_a(i_{k+1}) \leq \lambda_b(j_k)$.

En effet, dans le cas contraire, pour que la suite a « couvre » le caractère d'indice $\lambda_b(j_k)$, il doit exister i, $i_k < i < i_{k+1}$, tel que $\varphi_a(i) < \lambda_b(j_k) < \lambda_a(i)$. Dès lors, pour tous j, $j_k < j < j_{k+1}$, soit $\lambda_b(j) < \lambda_a(i)$, soit $\varphi_b(j) > \lambda_a(i)$, sinon le couple (i_{k+1}, j_{k+1}) ne serait pas le $(k+1)$-ième élément de S. Il en résulte que le caractère d'indice $\lambda_a(i)$ n'est pas « couvert » par la suite b. Or, par hypothèse, la suite b « couvre » la chaîne p'. D'où (1). Par ailleurs,

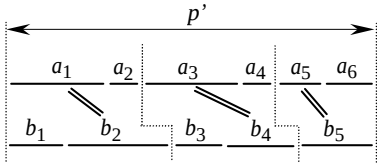


Fig. 4 : lemme 3.B

(2) $i_1 = 1$ et $j_s = n$.

On considère alors, d'une part, les sous-suites α_k, $1 \leq k \leq s$, de la suite a telles que $\alpha_k = a[i_k : i_{k+1} - 1]$ lorsque $k < s$ et $\alpha_s = a[i_s : m]$, d'autre part, part les sous-suites β_k, $1 \leq k \leq s$, de la suite b telles que $\beta_1 = b[1 : j_1]$ et $\beta_k = b[j_{k-1} + 1 : j_k]$ lorsque $1 < k \leq s$. Par ailleurs, l'inégalité (1) permet de considérer d'une part, pour tous k, $1 < k \leq s$, les facteurs $x_k = p'[\varphi_a(i_k) : \lambda_b(j_{k-1})]$ de p', d'autre part, pour tous k, $1 \leq k < s$, les facteurs $y_k = p'[\varphi_a(i_{k+1}) : \lambda_b(j_k)]$ de p'. De plus, on pose $x_1 = y_s = \varepsilon$. On vérifie alors aisément que, par construction et à partir des propriétés (1) et (2), les quadruplets $(\alpha_k, \beta_k, x_k, y_k)$, $1 \leq k \leq s$, sont des 1-recouvrements – indiqués figure 4 par les lignes verticales en pointillé –, qu'ils sont emboîtables dans l'ordre croissant de leur indice et que leur emboîtement est égal au recouvrement $(a, b, \varepsilon, \varepsilon)$. Donc $p' \in K$.

[$\boldsymbol{W^* K W^* \subseteq A}$] D'après les propriétés (1.1a) et (3.2a).

À partir du lemme 3.B, le théorème suivant fournit une expression explicite du langage AI.

THÉORÈME 3.C – Pour tout dictionnaire W sur un alphabet V,

(3.3a) $AI = K - (W^+ K \cup K W^+ \cup W^+ K W^+)$;

(3.3b) AI est un langage rationnel.

dem. (3.3a) D'après la propriété (1.1c) et le lemme 3.B, on a $AI = W^* K W^* - (W^+ W^* K W^* \cup W^* K W^* W^+ \cup W W^* K W^* W^+)$. On établit (3.3a) en remplaçant W^* par $W^+ \cup \{\varepsilon\}$. (3.3b) D'après les propriétés de fermeture des langages rationnels pour les opérations ensemblistes.

4. LIENS AVEC L'ALGORITHMIQUE DES CODES.

On présente brièvement dans ce chapitre comment ce qui précède conduit à deux algorithmes pour décider si un dictionnaire est ou n'est pas un code et comment le premier s'apparente à l'algorithme de Sardinas et Patterson (SP) et le second à celui de Spehner. Le chapitre est limité aux dictionnaires, donc aux codes finis.

4.1 *Caractérisation des codes à partir des recouvrements*. Un dictionnaire W est un code si, pour toutes suites a et b non vides sur W, l'égalité $\gamma(a) = \gamma(b)$ implique l'égalité $a = b$, i.e. γ est injective. Soit la suite $(R_n , n \geq 0)$ d'ensembles de recouvrements définie par :

(4.1) $R_0 = \{ (a, b, \varepsilon, y) \in R^1 \}$ et, pour tout $n \geq 0$, $R_{n+1} = R_n \circ R^1$.

On note FD_n, pour tout $n \geq 0$, l'ensemble des feuillures droites des éléments de R_n.

LEMME 4.A – Soit un dictionnaire W. Les propositions ci-dessous sont équivalentes :

(4.2a) W est un code ;

(4.2b) l'ensemble A des éléments ambigus de W^+ est vide ;

(4.2c) le langage L sur l'alphabet des 1-recouvrements de W est vide ;

(4.2d) pour tout $n \geq 0$, $\in \notin FD_n$;

(4.2e) l'ensemble des états utiles [1] de l'automate $(R^1, \underline{F} , \varepsilon_g, T, \varepsilon_d)$ est vide. □

Chacune des propriétés (4.2d) et (4.2e) conduit naturellement à un algorithme de décision pour les codes. Le premier, basé sur la suite des ensembles FD_n, est une visite en largeur des états de l'automate du langage L à partir de l'état ε_g. Il s'apparente à l'algorithme SP. Le second consiste à calculer l'ensemble des états utiles de l'automate du langage L et s'apparente à l'algorithme de Spehner.

4.2 *Eléments de comparaison avec l'algorithme* SP. Si X et Y sont des langages sur V, on note $X^{-1} Y$ le langage $\{z \in V^* \mid \exists x \in X , \exists y \in Y , y = x z \}$. Dès lors, soit la suite de langages définie par :

(4.3) $U_0 = W^{-1} W - \{ \varepsilon \}$ et, pour tout $n \geq 0$, $U_{n+1} = W^{-1} U_n \cup U_n{}^{-1} W$.

---

[1] Un état q d'un automate fini est *utile* s'il est *accessible* (il existe un chemin d'un état initial à q) et *coaccessible* (il existe un chemin de q à un état final).

L'algorithme SP est basé sur la propriété selon laquelle W est un code ssi, pour tout $n \geq 0$, $\varepsilon \notin U_n$ (Berstel & Perrin, 1985). Son itération est celle des ensembles U_n tandis que celle issue de la propriété (4.2d) est celle des ensembles FD_n. Le lemme ci-après fournit une définition directe et récursive des ensembles FD_n, qui constitue un élément pour comparer les deux suites.

LEMME 4.B – Soit un dictionnaire W. On a :

(4.4a) $FD_0 = (W^{-1} W - \{\varepsilon\}) \cup (W W^+)^{-1} W \cup W^{*-1}((W^{+-1} W)^{-1} W)$;

(4.4b) $FD_{n+1} = W^{*-1}((W^{*-1}(FD_n^{-1} W))^{-1} W)$, pour tout $n \geq 0$. □

indications. Dans l'expression (4.4a), le premier terme de l'union est l'ensemble des feuillures à droite f des 1-recouvrements de la forme $((a_1),(b_1),\varepsilon, f)$, $a_1 \in W$, $b_1 \in W$, le second celui de ceux de la forme $((a_1),(b_1, \ldots, b_n), \varepsilon, f)$, $n \geq 2$, et le dernier celui de ceux de la forme $((a_1, \ldots, a_m),(b_1, \ldots, b_n), \varepsilon, f)$, $m \geq 2$, $n \geq 2$.

4.3 *Eléments de comparaison avec l'algorithme de Spehner*. Un quadruplet (w, s, u, v), dans lequel $w \in W$, s est une suite non vide sur W, $u \in V^*$ et $v \in V^*$, est un *S-quadruplet* si $w = u\,\gamma(s)\,v$. Soient la relation binaire $B = \{ (u, v) \mid (w, s, u, v)$ est un S-quadruplet $\}$ sur V^* et l'ensemble C des éléments $c \neq \varepsilon$ de V^* tels qu'il existe, dans B, un chemin de ε à c et un chemin de c à ε. L'algorithme de Spehner est basé sur la propriété selon laquelle W est un code ssi l'ensemble C est vide (Lallement, 1979). Moyennant les transformations entre S-quadruplets et 1-recouvrements que décrit le lemme ci-après et que montre la figure 5, il est aisé d'établir que la vacuité de l'ensemble C est équivalente à la propriété (4.2e).

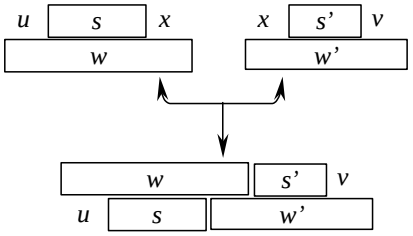


Fig. 5 : S-quadruplets et 1-recouvrements

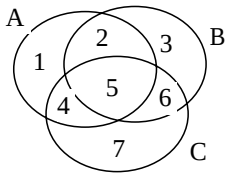


Fig. 6 *:* partition Π

LEMME 4.C – Etant donné un dictionnaire W,

(A) si deux S-quadruplets sur W sont de la forme (w, s, u, x) et (w', s', x, v), alors le quadruplet $(w+s', s+w', u, v)$, dans lequel + désigne l'adjonction à gauche ou à droite d'un élément à une suite, est un 1-recouvrement sur W ;

(B) si (a, b, u, v) est un 1-recouvrement avec $a = (a_i ; 1 \leq i \leq m)$ et $b = (b_j ; 1 \leq j \leq n)$, et si x désigne la chaîne sur laquelle a_1 et b_n se chevauchent, alors les quadruplets $(a_1, (b_j ; 1 \leq j \leq n-1), u, x)$ et $(b_n, (a_i ; 2 \leq i \leq m), x, v)$ sont des S-quadruplets. □

5. CALCUL D'UNE EXPRESSION RÉGULIÈRE DU LANGAGE ***AI***

La méthode de calcul présentée est en plusieurs étapes. Certaines utilisent des techniques avérées, d'autres sont plus spécifiques. À partir de l'expérience déjà menée sur la langue laotienne, ce chapitre expose la méthode puis ses aspects spécifiques.

5.1 *La méthode*. Les règles (3.1) et la propriété (3.3a) conduisent à la succession des étapes ci-dessous pour calculer une expression régulière du langage *AI* :

(5.1a) calcul de l'ensemble R^1 des 1-recouvrements du dictionnaire *W* ;

(5.1b) calcul de l'automate AUT_L du langage *L* à partir des règles (3.1) ;

(5.1c) calcul de l'automate AUT_K du langage *K* à partir de l'automate AUT_L ;

(5.1d) calcul de l'automate AUT_{AI} de *AI* à partir de l'automate AUT_K et de l'expression (3.3a) de *AI* ;

(5.1e) calcul d'une expression régulière de *AI* à partir de l'automate AUT_{AI}.

Avant d'être implantée en vraie grandeur, la chaîne de calculs (5.1) fut testée « à la main » sur des exemples réduits grâce au logiciel intéractif (*Automate*, 2000) dédié aux automates finis. Elle fut alors entièrement réalisée sur la langue laotienne dont le dictionnaire des syllabes compte environ 56 000 entrées. En fonction des conclusions de cette expérience, la chaîne est en cours de reprogrammation. La première leçon est que les temps d'exécution et l'espace mémoire qu'elle requiert peuvent devenir « redoutables » avec la taille du dictionnaire *W*. Or le dictionnaire du thaï compte plus d'un million de syllabes, celui du khmer plus d'un milliard ! Ces dictionnaires dont décrits par des grammaires de forme hors contexte sans symbole non terminal récursif de façon qu'ils restent finis. Pour chacune des langues étudiées, sa grammaire recense non seulement les syllabes attestées dans les mots mais aussi — et cela explique la taille volumineuse des dictionnaires — les syllabes « théoriques », autrement dit celles qui, bien qu'en accord avec leur langue, sont peu ou pas attestées mais dont l'occurence dans un texte ne peut être exclue. Ainsi en français, de la syllabe inhabituelle *xieng* pour l'écriture de la ville laotienne *Xieng-Khouang*. Compte tenu de la taille des dictionnaires, il est essentiel de pouvoir la réduire. Aussi commence-t-on par présenter un prétraitement qui substitue à tout dictionnaire un dictionnaire de cardinalité plus petite sur lequel opérera, sans perte d'information, la chaîne (5.1).

Dans la suite, on suppose que chaque dictionnaire est stocké sous la forme de l'automate d'états finis minimal qui le reconnaît. Le récent travail de Yeu (2003) a montré qu'en pratique cette hypothèse était réaliste. Le calcul des automates minimaux s'est en effet révélé rapide — moins d'une seconde pour les syllabes du laotien, moins de quatre pour celles du khmer(1). Partant de la grammaire du dictionnaire, l'algorithme en calcule une expression régulière, puis un automate non déterministe, finalement l'automate minimal. On souligne que l'algorithme « profite » de la forme grammaire originelle des dictionnaires : il ne réclame pas de stocker *in extenso* les dictionnaires ou de les énumérer syllabe après syllabe. Dans l'exemple du khmer, ces solutions seraient difficiles, voire impossibles, à mettre en œuvre compte tenu du grand nombre de syllabes, même à partir des techniques proposées par Daciuk, Watson & Watson (1998) pour la construction des automates minimaux acycliques. On précise enfin que l'algorithme qui réalise toutes les déterminisations d'automates mentionnées dans l'article a été implanté à partir des techniques décrites par Leslie (1995) et que l'algorithme de Brzozowski (Watson, 1993) est celui utilisé pour toutes les minimisations.

(1) Les temps d'exécution cités sont tous relatifs à une machine cadencée à 1,5 GHz.

5.2 *Réduction de la taille des dictionnaires*. Etant donné un dictionnaire W, on considère la relation d'équivalence ≡ sur l'alphabet V définie par :

$\forall a \in V, \forall b \in V, a \equiv b$ ssi $\forall m \in V^*, \forall m' \in V^*, m\ a\ m' \in W \Leftrightarrow m\ b\ m' \in W$.

Intuitivement, $a \equiv b$ si a et b « jouent le même rôle » dans le dictionnaire W. Soient $V/_{\equiv}$ l'ensemble quotient et π la surjection canonique qui associe à tout $v \in W$ sa classe d'équivalence πv. Considérant alors $V/_{\equiv}$ comme un nouvel alphabet, la surjection π s'étend canoniquement en un morphisme de monoïdes libres de V^* vers $(V/_{\equiv})^*$: $\pi v_1 v_2 \ldots v_n = \pi v_1 \pi v_2 \ldots \pi v_n$ et $\pi\, \varepsilon = \varepsilon$. Par ailleurs, on note πW le dictionnaire $\{ \pi w \mid w \in W \}$ sur $V/_{\equiv}$. Prenant $W = \{ aa, ab, ba, bb, c \}$ et posant $C^1 = \{ a, b \}$ et $C^2 = \{ c \}$, alors $V/_{\equiv} = \{ C^1, C^2 \}$, $\pi a = \pi b = C^1$, $\pi c = C^2$ et $\pi W = \{ C^1\ C^1, C^2 \}$.

LEMME 5.A – Pour tout dictionnaire W,

(5.2a) $W = \pi^{-1}\pi W$; (5.2b) $AI(W) = \pi^{-1}AI(\pi W)$. □

Eu égard à la propriété (5.2a), l'évaluation du langage $AI(W)$ sur V se ramène à celle du langage $AI(\pi W)$ sur $V/_{\equiv}$. Or le cardinal de $V/_{\equiv}$ est, par construction, au plus égal à celui de V et, par conséquent, celui du dictionnaire πW au plus égal à celui du dictionnaire W. Selon les dictionnaires, le gain de cardinalité peut se révéler important. Il l'est pour le dictionnaire laotien dont la taille est divisée par 5 (56 000 → 11 000), bien plus pour celui du khmer dont la taille est divisée par 10^4 (1 milliard → 100 000).

On montre maintenant comment évaluer la partition $V/_{\equiv}$ à partir de l'automate minimal AUT_W du dictionnaire W. Pour cette circonstance, on considérera une version complète de l'automate[(1)]. Soit T_W l'ensemble de ses transitions. Pour tout couple (q_i,q_j) d'états, on note $V_{i,j}$ le sous-ensemble $\{ c \in V \mid (q_i,c,q_j) \in T_W \}$ de V. L'union des sous-ensembles $V_{i,j}$ est égale à V. On note Π la moins fine [(2)] des partitions de V compatibles avec la famille des sous-ensembles $V_{i,j}$ [(3)]. Sur la figure 6, les ovales étiquetés par des lettres représentent les ensembles $V_{i,j}$ et la partition Π est constituée des « tommettes » numérotées.

LEMME 5.B – Pour tout dictionnaire W, les partitions Π et $V/_{\equiv}$ de V sont égales.

dem. [Π plus fine que $V/_{\equiv}$] Soient c^1 et c^2 dans la même classe C de Π. Puisque l'automate est déterministe et complet, pour tout état q_i, il existe un unique état q_j tel que $(q_i, c^1, q_j) \in T_W$, autrement dit, $c^1 \in V_{i,j}$. Or, de par la définition de la relation Π, $C \subseteq V_{i,j}$, donc $(q_i, c^2, q_j) \in T_W$. Il en résulte que $c^1 \equiv c^2$. [$V/_{\equiv}$ plus fine que Π] Soient c^1 et c^2 et soit $(q_i, c^1, q_j) \in T_W$. Puisque l'automate est déterministe et complet, il existe un unique état q_k tel que $(q_i, c^2, q_k) \in T_W$. Soit $m \in V^*$ tel que q_i soit l'état d'arrivée de l'automate opérant sur m à partir de l'état initial. Puisque $c^1 \equiv c^2$, pour tout $m' \in V^*$, $m\ c^1 m' \in W \Leftrightarrow m\ c^2 m' \in W$. Le langage reconnu par l'automate à partir de l'état q_j et celui à partir de l'état q_k sont donc égaux. Or l'automate est minimal. Il en résulte que $q_j = q_k$ d'après le théorème de Myhill-Nérode (Aho & Ullman, 1972). Il s'ensuit que c^1 et c^2 appartiennent à la même classe de Π.

(1) Un automate est *complet* si sa relation de transition est le graphe d'une application du produit $Q \times V$ dans Q.

(2) Si Π et Π' sont des partitions d'un même ensemble, Π est plus *fine* que Π' si chaque élément de Π est inclus dans un élément de Π'.

(3) Une partition Π d'un ensemble E est *compatible* avec une famille F de sous-ensembles non vides de E si pour chaque élément p de Π et chaque élément f de F, soit $p \subseteq f$, soit p et f sont disjoints.

5.3 *Calcul* (5.1a) *de l'ensemble des* 1-*recouvrements*. Dans ce paragraphe, on note, pour tout $v = c_1c_2...c_l$, dans V^*, $v^\sim = c_l...c_2c_1$ son image miroir et on désigne par $W^\sim$ le dictionnaire $\{ w^\sim \mid w \in W \}$. Par ailleurs, si $v \in V^+$ est un chemin dans un automate AUT, on dit que v *se prolonge* dans AUT s'il existe c dans V tel que la concaténation vc est un chemin dans AUT.

Un 1-recouvrement est *minimal* s'il est de la forme $((a_1), (b_1), x, y)$ dans laquelle les suites a et b sont de longueur 1. Le calcul des 1-recouvrements minimaux est la base pour celui de tous les 1-recouvrements puisque ceux-ci s'obtiennent en « comblant » de toutes les manières possibles, avec des éléments de W, les feuillures de ceux-là. L'algorithme « spontané » pour calculer l'ensemble des 1-recouvrements minimaux consiste à examiner tous les couples (w_1, w_2) d'éléments du dictionnaire W et, pour chacun d'eux, à calculer toutes les manières qu'ont w_1 et w_2 de se chevaucher, sans se confondre lorsqu'ils sont égaux. Cet examen est une itération de $|W|^2$ tours si $|W|$ est le cardinal de W. Sa complexité peut être significativement réduite en considérant les automates minimaux AUT_W de W et $AUT_{W^\sim}$ de $W^\sim$, qu'on supposera, pour cette circonstance, tous deux émondés[1]. Tout 1-recouvrement minimal est de la forme $((x\ \mu), (\mu\ y), x, y)$, $\mu \in V^+$, $x,y \in V^*$ et $xy \neq \varepsilon$. On dit de μ qu'il est le *milieu* du 1-recouvrement.

LEMME 5.C – Pour tout dictionnaire W, un élément μ de V^+ est un milieu ssi la conjonction des conditions (5.3) est vérifiée :

(5.3a) μ est un chemin dans l'automate AUT_W en partant de son état initial ;

(5.3b) $\mu^\sim$ est un chemin dans l'automate $AUT_{W^\sim}$ en partant de son état initial ;

(5.3c) ($\mu \notin W$) <u>ou</u> (μ se prolonge dans AUT_W) <u>ou</u> (($\mu^\sim$ se prolonge dans $AUT_{W^\sim}$).

dem. $\mu \in V^+$ est un milieu ssi $F(\mu) \equiv (\exists x \in V^*, \exists y \in V^*, \mu^\sim x^\sim \in W^\sim, \mu y \in W, x y \neq \varepsilon)$ vaut *vrai*. D'une part, on a $F(\mu) \Rightarrow$ (5.3a) <u>et</u> (5.3b). D'autre part, si la conjonction (5.3a) <u>et</u> (5.3b) est vérifiée, alors lorsque $\mu \notin W$, $F(\mu)$ vaut *vrai* puisque les automates AUT_W et $AUT_{W^\sim}$ sont émondés et, lorsque $\mu \in W$, $F(\mu)$ est équivalente à la disjonction (μ se prolonge dans AUT_W) <u>ou</u> (($\mu^\sim$ se prolonge dans $AUT_{W^\sim}$). Donc, $F(\mu) \Leftrightarrow$ (5.3a) <u>et</u> (5.3b) <u>et</u> (5.3c) [2].

Ce lemme conduit à un algorithme basé sur l'énumération de tous les chemins non vides μ de AUT_W. Pour chacun d'eux, il vérifie que $\mu^\sim$ est un chemin dans $AUT_{W^\sim}$, s'assure que la condition (5.3c) est satisfaite, calcule alors tous les 1-recouvrements qui ont μ comme milieu. Il existe autant de chemins non vides dans AUT_W à partir de son état initial que de préfixes non vide dans W. L'algorithme réalise donc au plus $lg(W) \times |W|$ tours, si $lg(W)$ désigne la longueur moyenne moyenne des éléments de W. Pour le laotien, dont la version réduite du dictionnaire compte environ 11 000 syllabes de longueur moyenne 3,76 caractères, l'algorithme combinatoire réaliserait $11\,000^2$ itérations tandis que celui du lemme (5.3c) en réalisera au plus $3{,}76 \times 11\,000$.

(1) Un automate est *émondé* (*trim automata*) si tous ses états sont utiles (Bellot & Sakarovitch, 1998).

(2) En appliquant la règle : si $p \Rightarrow q$ et $q \Rightarrow ((\neg r$ <u>et</u> $p)$ <u>ou</u> $(r$ <u>et</u> $(p \Leftrightarrow s)))$, alors $p \Leftrightarrow (q$ <u>et</u> $(\neg r$ <u>ou</u> $s))$.

5.4 *Calculs* (5.1b) *et* (5.1c) ; *réduction du nombre de transitions de* AUT_K. Les Calculs (5.1b) et (5.1c) se résument à appliquer les règles du chapitre 3, puis à émonder les automates visés. Ces étapes introduisent une nouvelle réduction des données, moins importante que celle du paragraphe 5.2, mais non négligeable. En effet, le nombre de transitions de l'automate AUT_L, partant celui de l'automate AUT_K, peut être réduit. À partir de l'automate AUT_L du langage L, on considère l'automate obtenu en supprimant toutes les transitions qui « partent » de l'état final ε_d, i.e. les transitions de la forme $(\varepsilon_d,\ (a, b, \varepsilon, y),\ y)$. Cet automate reconnaît le langage $L^{\sim}$ de l'ensemble R^1 des 1-recouvrements. Le même cheminement que celui qui, à partir du langage L, aboutit au langage K (chapitre 3) conduit, à partir de $L^{\sim}$, au langage $K^{\sim}$ et à son automate $AUT_{K^{\sim}}$.

LEMME 5.D – Pour tout dictionnaire W,

(5.4) $AI = K^{\sim} - (W^+ K^{\sim} \cup K^{\sim} W^+ \cup W^+ K^{\sim} W^+)$.

dem. Par construction, $K = K^{\sim} \cup K\ K^{\sim}$. Donc, à partir de la propriété (3.3a), $AI = (K^{\sim} \cup K\ K^{\sim}) - (W^+ K \cup K\ W^+ \cup W^+ K\ W^+)$. Or, $K\ K^{\sim} \subseteq W^+ K$, donc, $AI = K^{\sim} - (W^+ K \cup K\ W^+ \cup W^+ K\ W^+)$. Par ailleurs, $W^+ K = W^+ K^{\sim} \cup W^+ K\ K^{\sim} = W^+ K^{\sim}$ puisque $W^+ K \subseteq W^+$. De même, $W^+ K = W^+ K^{\sim}$ et $W^+ K\ W^+ = W^+ K^{\sim} W^+$.

Ce lemme permet d'optimiser l'étape (5.1d) en y remplaçant le langage K par le langage $K^{\sim}$ dont l'automate a moins de transitions et l'égalité (3.3a) par l'égalité (5.4).

5.5 *Calcul* (5.1d) *de l'automate* AUT_{AI}. Ce Calcul est probablement le plus délicat. En effet, la présence dans la formule (5.4) d'une différence ensembliste réclame que les automates de ses opérandes soient déterministes, à savoir celui de $K^{\sim}$ et celui de l'union $(W^+ K^{\sim} \cup K^{\sim} W^+ \cup W^+ K^{\sim} W^+)$. Or l'algorithme pour rendre déterministe un automate fini est, dans le pire des cas, de l'ordre de 2^q, q étant le nombre d'états. Le tableau de la figure 7 est lié au laotien. Il indique, pour certains des langages intermédiaires intervenant dans la formule (5.4), le nombre d'états (en gras) et le nombre de transitions (en italique) des versions non déterministe et minimale de leur automate. Il montre le fort « peuplement » des automates non déterministes opposé à celui de leur correspondant minimal. Ce constat a conduit à systématiquement minimiser chacun des automates intermédiaires. Le temps de calcul de l'étape (5.1d), appliquée au laotien, fut de sept minutes.

	non déterministe	minimal
$K^{\sim}$	13829 ; *88881*	41 ; *262*
$W^+ K^{\sim}$	14403 ; *95574*	85 ; *1486*
$K^{\sim} W^+$	15550 ; *96147*	75 ; *913*
$W^+ K^{\sim} W^+$	15550 ; *102840*	126 ; *2133*
AI		38 ; *253*

Fig. 7 – *taille des automates.*

5.6 *Calcul* (5.1e) *d'une expression régulière du langage AI.* Une première expression régulière est obtenue, à partir de l'automate AUT_{AI}, par la méthode basée sur l'algorithme de résolution d'équations linéaires par élimination gaussienne (Aho & Ullman, 1972), (Bellot & Sakarovitch, 1998). En développant cette expression, on aboutit à la liste des motifs irréductibles. Le laotien en compte 240 000. Certains contiennent l'itérateur * de Kleene. Ainsi de

ແກ(ອ (ອ ອ ນ ອ ອ)* ງ ອ ຍ ງ et ເ ກ ວ ອ (ອ)* ນ ອ ອ (ອ ອ ນ ອ ອ)* ງ ອ ຍ ງ.

Le langage des ambiguïtés irréductibles issu de la grammaire des syllabes laotiennes utilisée est donc infini. Cependant, on fait remarquer qu'aucun des motifs infinis ne correspond à des suites de mots de la langue laotienne. Ceci ne doit pas surprendre puisque toutes les concaténations de syllabes, i.e. tous les éléments de W^+, ne font pas partie de la langue laotienne.

BIBLIOGRAPHIE,

Automate, logiciel. Caylux, B., (2000).
Téléchargeable à partir de brassens.upmf-grenoble.fr/IMSS/logiciels/

Aho A. V. & Ullman J. D., (1972). *The theory of parsing, translation and compiling.* Prentice Hall.

Bellot P. & Sakarovitch J., (1998). *Logique et automates.* Ellipses.

Berment V., (1997). *Traitement automatique du laotien. Quelques aspects morphologiques.* Mémoire de DREA, Institut National des Langues et Civilisations Orientales (INALCO), 96 p.

Berment V., (1998). *Prolégomènes graphotaxiques du laotien.* Mémoire de DEA, Institut National des Langues et Civilisations Orientales (INALCO), 160 p.

Berment V., (2002). *Several Technical Issues for Building New Lexical Bases.* Papillon Seminar, July 2002. 5 p.

Berstel J. & Perrin D., (1985). Theory of codes. Academic Press, Orlando.

Daciuk J., Watson B. W. & Watson R. E., (1998). *Incremental construction of minimal acyclic finite state automata and transducers*. Actes de Finite State Methods in Natural Language Processing, Université de Bilkent, Ankara, Turquie, juin-juillet 1998.

Head T. & Weber A., (1995). *Deciding multiset decipherability*. IEEE transactions on information theory, vol. 41, pp 291-297.

Lallement G., (1979). *Semigroups and combinatorial applications*. John Wiley & sons, New York.

Lao Word, logiciel. Berment V., (2000).
Téléchargeable à partir de www.LaoSoftware.com

Leslie T., (1995). *Efficient Approaches to Subset Construction*. Masters thesis, Computer Science, University of Waterloo. Téléchargeable à partir de
www.csd.uwo.ca/research/grail/.papers/subset.ps

Lothaire M., (2002). *Algebraic combinatorics on words*. Cambridge University Press.

Watson B. W., (1993). *A taxonomy of finite automata minimization algorithms*. Université de Technologie d'Eindhoven, Pays-Bas. Téléchargeable à partir de
http://www.cs.up.ac.za/cs/bwatson/publications.html#1993

Weber A. & Head T., (1994). *The finest homophonic partition and related code concepts.* Actes de Mathematical Foundations of Computer Science, Lecture Notes in Computer Science, n° 841, pp 618-628, Springer-Verlag, ed. Privara, Rovan & Ruzicka.

Yeu T., (2003). *Découpage syllabique des langues du Sud-Est Asiatique*. Rapport de stage d'IUT, Université René Descartes, Paris 5, 28 pages.

A.13 LICENCE GPL TYPE

Une version française de la licence GPL, non officielle, est disponible sur le site http://www.linux-france.org/article/these/gpl.html.

Pour une analyse juridique de cette licence, voir [Clément-Fontaine 1999].

GNU GENERAL PUBLIC LICENSE
Version 2, June 1991

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.

b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients'

exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED

WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

<one line to give the program's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

Gnomovision version 69, Copyright (C) year name of author
Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type 'show w'.
This is free software, and you are welcome to redistribute it under certain conditions; type 'show c' for details.

The hypothetical commands 'show w' and 'show c' should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than 'show w' and 'show c'; they could even be mouse-clicks or menu items--whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the program ‘Gnomovision’ (which makes passes at compilers) written by James Hacker.

<signature of Ty Coon>, 1 April 1989
Ty Coon, President of Vice

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.

A.14 GÉNÉRATION DE SYLLABES À PARTIR D'UNE GRAMMAIRE EN PROLOG

Le programme Prolog ci-dessous a été réalisé pour vérifier la génération de l'arbre lexicographique des syllabes (voir III.2.4 et III.2.6). Il est basé sur la grammaire des syllabes laotiennes que nous utilisions au moment de cette vérification.

```
imprimer_syllabes:-
  open('syllabes.imp',write,_,[alias(imp)]),
  imp_syl,
  close(imp).

imp_syl:-
  syl(S),
  ecrire_codes(imp,S),
  nl(imp),
  fail.

imp_syl.
ecrire_codes(_,[]).

ecrire_codes(Fimp,[C|Cs]):-
  ((integer(C),put(Fimp,C));write(Fimp,C)),!,
  ecrire_codes(Fimp,Cs).

syl(S) :- syllabe(S,[]).
syllabe(X1,X2) :- gcin(X1,X),suitegcin(X,X2).
syllabe(X1,X2) :- e(X1,X),suitee(X,X2).
syllabe(X1,X2) :- ee(X1,X),suiteee(X,X2).
syllabe(X1,X2) :- o(X1,X),suiteo(X,X2).
syllabe(X1,X2) :- ai(X1,X),suiteai(X,X2).
syllabe(X1,X2) :- ae(X1,X),suiteai(X,X2).
syllabe(X1,X2) :- cin(X1,X),r(X,X2).
syllabe("3s[",[]).

e([99|R],R).
ee([103|R],R).
o([51|R],R).
ai([119|R],R).
ae([46|R],R).

suitegcin([113,59,116|R],R).
suitegcin([116|R],R).
suitegcin(X1,X2) :-
  acc(X1,X),suitegcin1(X,X2).
suitegcin(X1,X2) :- suitegcin1(X1,X2).
suitegcin1([101|R],R).
suitegcin1([80|R],R).
suitegcin1([113,59|R],R).
suitegcin1([61|R],R).
suitegcin1(X1,X2) :-
  avantcfobl(X1,X),cfobl(X,X2).
suitegcin1(X1,X2) :-
  avantcfopt(X1,X),cfopt(X,X2).
cfopt([R],R).
cfopt([],[]).
cfopt([X|R],R) :- cfobl([X|R],R).
avantcfobl([97|R],R).
avantcfobl([113|R],R).
avantcfobl([80,97|R],R).
avantcfobl([80|R],R).
avantcfobl([59,97|R],R).
avantcfobl([59|R],R).
avantcfobl([118,97|R],R).
avantcfobl([118|R],R).
avantcfopt([107|R],R).
avantcfopt([121|R],R).
avantcfopt([117|R],R).
avantcfopt([98|R],R).
avantcfopt([110|R],R).
avantcfopt([53|R],R).
avantcfopt([54|R],R).

suitee(X1,X2) :- gcin(X1,X),suitee1(X,X2).
suitee1([116|R],R).
suitee1(X1,X2) :- acc(X1,X),suitee2(X,X2).
suitee1(X1,X2) :- suitee2(X1,X2).
suitee2(X1,X2) :- a(X1,X),cfobl(X,X2).
suitee2(X1,X2) :- cfobl(X1,X2).
suitee2([R],R).
a([97|R],R).

suiteee(X1,X2) :- gcin(X1,X),suiteee1(X,X2).
suiteee1(X1,X2) :- suitee1(X1,X2).
suiteee1([97,80,116|R],R).
suiteee1([97,112|R],R).
suiteee1([107,116|R],R).
suiteee1(X1,X2) :- acc(X1,X),suiteee2(X,X2).
suiteee1(X1,X2) :- suiteee2(X1,X2).
suiteee2(X1,X2) :- b(X1,X),cfopt(X,X2).
suiteee2([113,107|R],R).
suiteee2([112|R],R).
suiteee2([97,80|R],R).

b([121|R],R).
b([117|R],R).
b([98,118|R],R).
b([110,118|R],R).

suiteo(X1,X2) :- gcin(X1,X),suiteo1(X,X2).
suiteo1(X1,X2) :- acc(X1,X),suiteo2(X,X2).
suiteo1(X1,X2) :- suiteo2(X1,X2).
suiteo1([116|R],R).
```

```
suiteo2(X1,X2) :- cfobl(X1,X2).
suiteo2([R],R).

suiteai(X1,X2) :- gcin(X1,X),suiteai1(X,X2).
suiteai1(X1,X2) :- acc(X1,X2).
suiteai1([R],R).

gcin(X1,X2) :- cin(X1,X2).
gcin(X1,X2) :- gci(X1,X2).
gci(X1,X2) :- gca(X1,X2).
gci(X1,X2) :- gce(X1,X2).
gci(X1,X2) :- gcg(X1,X2).
gce(X1,X2) :- cer(X1,X),suitecer(X,X2).
gce(X1,X2) :- cel(X1,X),suitecel(X,X2).
gce(X1,X2) :- cels(X1,X),suitecels(X,X2).
gce(X1,X2) :- cev(X1,X),suitecev(X,X2).
gcg(X1,X2) :- avantcgsh(X1,X),cgsh(X,X2).
gcg(X1,X2) :-
   avantcgkhh(X1,X),cgkhh(X,X2).
gcg(X1,X2) :- avantcgthh(X1,X),cgthh(X,X2).
gcg(X1,X2) :-
   avantcgphh(X1,X),cgphh(X,X2).
gcg(X1,X2) :- avantcgfh(X1,X),cgfh(X,X2).
gcg(X1,X2) :- avantcgch(X1,X),cgch(X,X2).
gcg(X1,X2) :- avantcgt(X1,X),cgt(X,X2).
gcg(X1,X2) :- avantcgp(X1,X),cgp(X,X2).
gcg(X1,X2) :-
   avantcgphb(X1,X),cgphb(X,X2).
cfobl(X1,X2) :- cfo(X1,X2).
cfobl(X1,X2) :- cfs(X1,X2).

gca([115,39|R],R).
gca([115,111|R],R).
gca([115,112|R],R).
gca([115,80|R],R).
gca([115,105|R],R).
gca([115,94|R],R).
gca([115,59|R],R).
gca([115,93|R],R).

suitecer([105|R],R).
suitecel([93|R],R).
suitecels([94|R],R).
suitecev([59|R],R).

cer([100|R],R).
cer([55|R],R).
cer([56|R],R).
cer([120|R],R).
cer([114|R],R).

cel([100|R],R).
cel([55|R],R).
cel([56|R],R).
cel([120|R],R).

cels([100|R],R).
cels([55|R],R).
cels([56|R],R).
cels([120|R],R).

cev([100|R],R).
cev([48|R],R).
cev([55|R],R).
cev([39|R],R).
cev([57|R],R).
cev([108|R],R).
cev([45|R],R).
cev([52|R],R).
cev([109|R],R).
cev([93|R],R).

avantcgsh([108|R],R).
avantcgkhh([48|R],R).
avantcgthh([52|R],R).
avantcgphh([122|R],R).
avantcgfh([47|R],R).
avantcgch([57|R],R).
avantcgt([56|R],R).
avantcgp([120|R],R).
avantcgphb([114|R],R).

cgsh([39|R],R).
cgsh([112|R],R).
cgsh([111|R],R).
cgsh([44|R],R).
cgsh([105|R],R).
cgsh([106|R],R).
cgsh([94|R],R).
cgsh([59|R],R).

cgkhh([112|R],R).
cgkhh([111|R],R).
cgkhh([44|R],R).
cgkhh([93|R],R).
cgkhh([94|R],R).

cgthh([112|R],R).
cgthh([111|R],R).
cgthh([93|R],R).
cgthh([94|R],R).
cgthh([59|R],R).

cgphh([112|R],R).
cgphh([111|R],R).
cgphh([93|R],R).
cgphh([94|R],R).

cgfh([105|R],R).
cgch([105|R],R).
```

```
cgt([93|R],R).
cgt([94|R],R).

cgp([112|R],R).
cgp([44|R],R).
cgp([93|R],R).
cgp([94|R],R).
cgp([59|R],R).

cgphb([112|R],R).
cgphb([93|R],R).
cgphb([94|R],R).

cin([100|R],R).
cin([48|R],R).
cin([55|R],R).
cin([39|R],R).
cin([57|R],R).
cin([108|R],R).
cin([45|R],R).
cin([112|R],R).
cin([102|R],R).
cin([56|R],R).
cin([52|R],R).
cin([109|R],R).
cin([111|R],R).
cin([91|R],R).
cin([120|R],R).
cin([122|R],R).
cin([47|R],R).
cin([114|R],R).
cin([50|R],R).
cin([44|R],R).
cin([49|R],R).
cin([105|R],R).
cin([59|R],R).
cin([115|R],R).
cin([118|R],R).
cin([73|R],R).
cin([93|R],R).

cfo([100|R],R).
cfo([102|R],R).
cfo([91|R],R).
cfo([39|R],R).
cfo([111|R],R).
cfo([44|R],R).
cfo([112|R],R).
cfo([80|R],R).
cfo([59|R],R).

cfs([48|R],R).
cfs([55|R],R).
cfs([56|R],R).
cfs([52|R],R).
cfs([109|R],R).
cfs([57|R],R).
cfs([108|R],R).
cfs([45|R],R).
cfs([120|R],R).
cfs([114|R],R).
cfs([50|R],R).
cfs([105|R],R).
cfs([93|R],R).

acc([106|R],R).
acc([249|R],R).
acc([104|R],R).
acc([232|R],R).
acc([72|R],R).
acc([233|R],R).
acc([74|R],R).
acc([224|R],R).

r([105|R],R).
```

GLOSSAIRE ET BIBLIOGRAPHIE

GLOSSAIRE DES TERMES LINGUISTIQUES UTILISÉS

Voir aussi le glossaire en ligne : http://alis.isoc.org/glossaire/index.html.

Affinité entre langues : « On parle d'affinité entre deux ou plusieurs langues, qui n'ont entre elles aucune parenté génétique, quand elles présentent certaines ressemblances structurelles (organisation de la phrase, vocabulaire général, déclinaison, etc.). Par exemple, les similitudes existant entre la déclinaison latine et la déclinaison russe sont dues à une parenté génétique puisque la grammaire comparée attribue aux deux langues une origine commune : l'indo-européen ; en revanche, les ressemblances entre le takelma et l'indo-européen sont dues, elles, à une certaine affinité. » ([Dubois et al. 1994]). *Voir aussi Parenté linguistique.*

Contact de langues : « Le *contact de langues* est la situation humaine dans laquelle un individu ou un groupe sont conduits à utiliser deux ou plusieurs langues. Le contact de langues est donc l'événement concret qui provoque le bilinguisme ou en pose les problèmes. Le contact de langues peut avoir des raisons géographiques : aux limites de deux communautés linguistiques, les individus peuvent être amenés à circuler et à employer ainsi leur langue maternelle, tantôt celle de la communauté voisine. C'est là, notamment, le contact de langues des pays frontaliers… Mais il y a aussi contact de langues quand un individu, se déplaçant, par exemple, pour des raisons professionnelles, est amené à utiliser à certains moments une autre langue que la sienne. D'une manière générale, les difficultés nées de la coexistence dans une région donnée (ou chez un individu) de deux ou plusieurs langues se résolvent par la commutation ou usage alterné, la substitution ou utilisation exclusive de l'une des langues après élimination de l'autre ou par amalgame, c'est-à-dire l'introduction dans des langues de traits appartenant à l'autre… » ([Dubois et al. 1994]). *Voir aussi Affinité entre langues.*

Dialecte : « Employé couramment pour *dialecte régional* par opposition à 'langue', le *dialecte* est un ensemble de signes et de règles combinatoires de même origine qu'un autre système considéré comme une langue, mais n'ayant pas acquis le statut culturel et social de cette langue indépendamment de laquelle il s'est développé : quand on dit que le picard est un dialecte français, cela ne signifie pas que le picard est né de l'évolution (ou à plus forte raison de la 'déformation') du français… » ([Dubois et al. 1994]). *Voir aussi Langue et Idiome.*

Famille linguistique ou famille de langue : « On dit que deux langues appartiennent à la même *famille* quand elles sont apparentées génétiquement, c'est-à-dire quand tout laisse à penser qu'elles se sont développées à partir d'une origine commune. Généralement, on réserve la dénomination de *famille de langues* à un ensemble formé par toutes les langues connues de même origine ; dans cet ensemble, les sous-ensembles constitués par certaines langues apparentées plus étroitement entre elles qu'avec les autres sont des *branches* ou *sous familles*. Le terme de groupe s'applique indifféremment à un ensemble de familles, à une famille, à un ensemble de branches d'une même famille, à un ensemble de langues d'une même branche : il implique que le classement n'est pas encore établi… » ([Dubois et al. 1994]). *Voir aussi Groupe de langues.*

Forme canonique : Forme d'une représentation textuelle définie par convention comme représentation unique. Par exemple, si 'e et é peuvent être équivalents dans certains cas, on peut considérer que é est la forme canonique. Nous utilisons cette définition de la locution pour des cas de représentations multiples dues à l'existence :

- ⇨ de plusieurs saisies possibles d'un même terme donnant le même résultat visuel,
- ⇨ de plusieurs encodages possibles pour un système d'écriture.

Nous n'utilisons pas cette définition de la locution pour les multiples formes possibles d'un même terme dues à l'absence d'une orthographe figée.

Groupe linguistique ou groupe de langues : « Le terme de groupe de langues désigne un ensemble de langues réunies pour une raison génétique, typologique ou géographique. » ([Dubois et al. 1994]). *Voir aussi Famille de langues.*

Homographes : On dit que deux formes sont *homographes* quand elles ont la même graphie mais des sens différents ([Dubois et al 1994]). Par exemple, le *fils* de Marc et les *fils* de coton.

Idiome : « Les linguistes considèrent que tout parler ou idiome est :
- ⇨ soit une langue, utilisée par une population entière dans un ou plusieurs pays,
- ⇨ soit une variante de celle-ci, propre à une partie de cette population, un 'lecte'.
  Si elle est propre à une région, elle est appelée alors dialecte, ou, s'il s'agit du parler populaire, de celui des classes moyenne ou de la forme châtiée, officielle, académique).
- ⇨ soit, enfin, la pratique particulière caractérisant chaque individu, dite 'idiolecte'.

Mais l'usage général tend à réserver le terme langue aux parlers institutionnalisés des États-nations et à traiter tous les autres de dialectes ou à les affubler de termes péjoratifs comme celui de 'patois'… L'univers linguistique est particulièrement instable et ses unités de base qui sont les langues et les dialectes, en perpétuelle évolution, peuvent amener les linguistes eux-mêmes à des analyses assez divergentes, ne serait-ce que sur la qualification et le dénombrement des unes et des autres. » ([Breton 2003], page 15).

Langue : « On reconnaît l'existence d'une plurarité de *langues* dès qu'on parle de langue française, anglaise, etc. Ce terme entre en concurrence avec les autres mots (dialectes, parlers, patois) qui désignent aussi des systèmes de communication linguistiques. La notion de langue est une notion pratique, mais complexe, introduite avant que la linguistique ne se constitue.

Langue écrite et institutions

Quand on applique le mot aux pays modernes, les institutions et les habitudes donnent par énumération la liste des langues. Il s'agit alors de réduire les langues aux formes standard dont les utilisateurs, généralement pour des raisons extralinguistiques, considèrent que ce sont des langues. Les caractères définitoires de la langue peuvent être alors l'existence d'une tradition d'écriture et même de littérature, mais aussi le statut institutionnel. Selon que l'on fait intervenir celui-ci ou non, le nombre de langues est plus ou moins grand. Ce statut institutionnel peut exclure tout enseignement au moins officiel (c'est le cas du sort réservé à certains parlers) ou leur confère un rôle d'appoint (c'est le cas des langues qu'on peut présenter à certains examens en épreuves facultative : occitan, breton). En France, on ne reconnaît le statut de langue maternelle, à apprendre à l'école primaire, qu'au français standard…

Langue à formes écrites non enseignées

On parle aussi de langues là où il n'y a pas d'enseignement ou, en tous cas, pas d'enseignement de certains systèmes linguistiques que l'on appelle langues (ainsi, au Sénégal, où l'enseignement est donné en français, le ouolof est une langue). On n'a pas toujours dans ce cas-là le critère de l'écriture pour dire qu'un ensemble de parlers locaux est une langue, par opposition à un autre ensemble voisin ou occupant la même zone qui est considérée comme une autre langue. Le critère qui semble le plus évident dans ce cas est celui de l'intelligibilité mutuelle, ou intercompréhension. On poserait comme principe que si deux personnes ayant des dialectes différents se comprennent en parlant chacun dans son dialecte, elles parlent la même langue ; sinon elles parlent des langues différentes. En réalité, l'intercompréhension est quelque chose de relatif : on ne se comprend jamais entièrement, on se comprend toujours un peu : un Bonifacien (de dialecte génois) comprend bien un Porto-Vecchiais (de dialecte corso-gallurais), mais l'inverse n'est pas vrai ; et entre un Porto-Vecchiais et un Cap-Corsin (ayant tous deux conscience de parler la même langue), l'intercompréhension sera possible par l'acceptation de la polynomie.

Un autre critère peut être l'énumération des éléments communs. On peut établir une liste du vocabulaire fondamental de 100 mots et établir la concordance de 0 à 100 p. 100. On pourrait sans doute procéder de même pour la morphologie ou la syntaxe, mais le problème est de savoir à partir de quel pourcentage d'écart on dira qu'il y a deux langues. Le problème est que le parler d'un village B sera proche de celui d'un village voisin A, celui de C proche de celui de B, et ainsi de suite jusqu'à Z, mais qu'il y aura un énorme écart entre les dialectes A et Z. Il y a très souvent continuité linguistique dans toute la zone des langues romanes, alors qu'on parle de langues différentes. De même, les isoglosses ne coïncident jamais entièrement, et il faut alors choisir entre les traits négligeables et importants…

En dehors des formes écrites, la définition des langues est donc compliquée, dans la mesure où la continuité linguistique est chose fréquente. » ([Dubois et al. 1994]). *Voir aussi Idiome.*

Langue grégaire : Terme sans connotation péjorative défini par Louis-Jean Calvet dans [Calvet 1987], page 80 et signifiant : « avec connivence ». Ainsi, une langue grégaire est « une langue de petit groupe, qui limite donc la communication à quelques-uns et dont la forme est marquée par cette volonté de limitation », par exemple les argots à clefs (verlan…), les registres sociaux, les formes linguistiques de classes d'âge ou les langues familiales. Calvet donne encore l'exemple de Français qui, travaillant aux Etats-Unis, utiliseraient le français entre eux (fonction grégaire).

Langue nationale : Langue dont le statut est variable d'un pays à l'autre : langue unique de l'école, de l'administration (Burundi=kirundi, République Centrafricaine=sango), langues régionales en nombre limité (Zaïre=4, Guinée=8), toutes les langues du pays (Burkina Faso=70), aucune des langues du pays (Tchad=0) ([Calvet 1999], page 54 et 55).

Langue officielle : Langue de fonctionnement de l'État, langue de l'école, des média, etc. ([Calvet 1999], page 54). « Sur 200 États souverains dans le monde, 160 sont officiellement unilingues au niveau national. Une trentaine sont bilingues, 7 trilingues (Belgique, Luxembourg, Bosnie, Érythrée, Rwanda, Seychelles, Vanuatu) et deux quadrilingues (Suisse et Singapour)… Certains États reconnaissent des langues officielles ou co-officielles au niveau régional. C'est le cas des deux pays officiellement multinationaux, la Russie avec ses 130 nationalités et leurs 55 entités autonomes, et la Chine pour ses 55 nationalités minoritaires et leurs 150 entités autonomes. L'Inde, résolument uninationale et authentiquement fédérale, liste, après ses deux langues officielles au niveau national, 18 langues 'constitutionnelles', dont la plupart sont celles des 35 États et Territoires de l'Union. » ([Breton 2003], page 31).

Langues parentes : *Voir Parenté linguistique.*

Latin : Systèmes d'écriture incluant les lettres latines et les différentes lettres, symboles et signes diacritiques ajoutés et utilisés conjointement.

Parenté génétique : *Voir Parenté linguistique.*

Parenté linguistique : « La linguistique historique définit deux sortes de *parentés*, l'une historique ou génétique, l'autre typologique. Deux langues sont apparentées génétiquement quand elles proviennent de l'évolution d'une langue unique. L'histoire permet parfois de fonder la parenté historique ; c'est le cas, par exemple, des langues romanes issues du latin. Plus souvent, la parenté est prouvée par comparaison ; c'est le cas pour le groupe de langues relevant de la famille indo-européenne. On peut établir aussi des parentés typologiques ; on constate ainsi que, dans certaines régions, des langues, différentes au départ, tendent à converger, à se rapprocher (contact de langues). Il se produit aussi des convergences fortuites, comme on en a constaté entre le tswana d'Afrique du Sud et le germanique (consonantismes ressemblants) ; de même, le takelma et l'indo-européen ont six importants traits typologiques en commun. On réserve le nom d'affinité aux convergences fortuites et celui de parenté dans l'hypothèse d'une origine commune. » ([Dubois et al. 1994]). *Voir aussi Affinité entre langues et Contact de langues.*

Parenté typologique : *Voir Parenté linguistique.*

Parler : *Voir Idiome.*

Syllabe : Dans ce document, nous employons le terme *syllabe* pour représenter la chaîne de caractères correspondant à la syllabe phonologique définie ci-dessous.
« On appelle *syllabe* la structure fondamentale qui est à la base de tout regroupement de phonèmes dans la chaîne parlée. Cette structure se fonde sur le contraste de phonèmes appelés traditionnellement *voyelles* et *consonnes*. La structure phonématique de la syllabe est déterminée par un ensemble de règles qui varient de langue à langue… » ([Dubois et al. 1994]).

Système d'écriture : Nous utilisons la locution *système d'écriture* dans le sens donné par le Grand Robert pour *écriture* : « Système de représentation de la parole et de la pensée par des signes conventionnels tracés et destinés à durer. » ([Rey 2001]).

RÉFÉRENCES BIBLIOGRAPHIQUES

Agirre E., Aldezabal I., Alegria I., Arregi X, Arriola J.M., Artola X., Diaz de Ilarraza A., Ezeiza N., Gojenola K, Sarasola K., Soroa A., (2001). *Developing language technology for a minority language: progress and strategy*, in Elsnews 10.1 (printemps 2001), ISSN 1350-990X.

Arendt Hannah, (1995). *Qu'est-ce que la politique ?*. Éditions du Seuil, ISBN 2-02-021769-4.

Armentier Louis, (1980). *Orientalisme et linguistique*. Montréal : Univers (voir plus particulièrement p. 122-134 et p. 162-181).

Aroonmanakun Wirote, (2002). *Collocation and Thai word segmentation*. Proceeding of the Fifth Symposium on Natural Language Processing / Oriental COCOSDA, p. 68-75, Hua Hin, Thailand, ISBN 974-572-947-7.

Bauhahn Maurice, (2002). *Rendering the world's complex scripts : a case study in Khmer*. 21e conférence internationale Unicode, Dublin (Irlande), mai 2002.

Berment Vincent, (1997). *Traitement automatique du laotien. Quelques aspects morphologiques*. Mémoire de DREA, INALCO, Paris, 1997.

Berment Vincent, (1998). *Prolégomènes graphotaxiques du laotien*. – Mémoire de DEA, INALCO, Paris.

Berment Vincent, (2002a). *Several technical issues for building new lexical bases*. Papillon Seminar, Tokyo, Juillet 2002.

Berment Vincent, (2002b). *UNL annotation of minority languages web pages using the Papillon lexical base*. Pre-COLING Conference, Penang, Août 2002.

Berment Vincent, (2002c). *Several directions for minority languages computerization*. COLING Conference, Taipei, Août 2002.

Berment Vincent, Thongvilu Houmphanh, (2003). *Cooperative Lao ICT framework. Case study: construction of Lao lexical resources*. Regional Conference on Digital GMS, Bangkok, Février 2003. 2003.

Berment Vincent, Jacqmin Thakkhinh, Dechanet Blandine, (2003). *Parts of Speech for the LaoLex Dictionary*. STEA PAN-Laos National Seminar, Vientiane, Mars 2003.

Berment Vincent, (2003a). *LaoWord's Word Processing Functions*. STEA PAN-Laos National Seminar, Vientiane, Mars 2003.

Berment Vincent, (2003b). *Current status of the Papillon-Lao database: Tools (LaoLex), dictionary (LaoDict), XML schema and export towards Papillon*. Papillon Seminar, Sapporo, Juillet 2003.

Berment Vincent, (2003c). *XML schema for Lao*. Papillon Seminar, Sapporo, Juillet 2003.

Bernot Denise, Cardinaud Marie-Hélène, Yin Yin Myint Marie, (1990). *Manuel de birman, volume 1*. Langues de l'Asie – INALCO, L'Asiathèque, ISBN 2-901795-36-6.

Berstel J., Perrin D., (1985). *Theory of codes*. Academic Press, Orlando.

Boite René, Bourlard Hervé, Dutoit Thierry, Hancq Joël, Leich Henri, (2000). Traitement de la parole, Presses polytechniques et universitaires romandes, ISBN 2-88074-388-5

Boitet Christian, (1990). *La TAO à Grenoble en 1990, 1980-90 : TAO du réviseur et TAO du traducteur*, partie des supports de l'école d'été de Lannion organisée en 1990 par le LATL et le CNET.

Boitet Christian, (1993). *La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue*, in : « La Traductique », pages 109 à 148, AUPELF-UREF, édité par André Clas & Pierrette Bouillon, Les Presses de l'Université de Montréal et AUPELF-UREF, 1993, ISBN 2-7606-1616-9.

Boitet Christian, (1996). *La synergie entre THAM, réseau et TA comme facteur de progrès théoriques et pratiques en TAO*, communication au colloque TAL+AI/NLP/IA, Moncton, actes édités par Ch. Moghrabi.

Boitet Christian, (1999). *A research perspective on how to democratize machine translation and translation aids aiming at high quality final output*. MT Summit, 1999.

Boitet Christian, (2000). *Traduction assistée par ordinateur*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Boitet Christian, (2001a). *Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes*, TALN 2001, Tours, 2-5 juillet 2001.

Boitet Christian, (2001b). *Four technical and organizational keys for handling more languages and improving quality (on demand) in MT*, MTS 2001, Workshop on "MT2010 – Towards a Road Map for MT", 18 septembre 2001, Santiago de Compostela.

Bonhomme Patrice, (2000). *Codage et normalisation de ressources textuelles*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Bouillon Pierrette, (1998). *Traitement automatique des langues naturelles*, Ouvrage collectif, Duculot, ISBN 2-8011-1181-3.

Breton Roland, (1995). *Géographie des langues*, 3e édition, PUF (collection *Que sais-je ?*), ISBN 2-13-046663-x.

Breton Roland, (2003). *Atlas des langues du monde*, Autrement (collection *Atlas/Monde*), ISBN 2-7467-0400-5.

Brown Marie-Hélène, (1991). *Lire et écrire le thaï*, Duangkamol, ISBN 9-7421-0495-6.

Bür Jürgen, Bauder Irene, (1992). *Windows 3.1*, Micro Application – ISBN: 2-86899-701-5.

Calvet Louis-Jean, (1987). *La guerre des langues et les politiques linguistiques*, 2e édition, Hachette Littératures (collection *Pluriel*), ISBN 2-01-278985-4.

Calvet Louis-Jean, (2002). *Le marché aux langues*, Plon, ISBN 2-259-19660-8.

Campbell George, (1991). *Compendium of the World's Languages*, 2 volumes, Routledge, ISBN 0-4150-6978-5.

Canals-Marote R., Esteve-Guillén A., Garrido-Alenda A., Guardiola-Savall M.I., Itturaspe-Bellver A., Montserrat-Buendia S., Ortiz-Rojas S., Pastor-Pina H., Pérez-Antón P.M., Forcada M.L., (2001). *The Spanish-Catalan machine translation system interNOSTRUM*, MT Summit, 2001.

Charnyapornpong Surin, (1983). *A Thai syllable separation algorithm*, Master thesis, Asian Institute of Technology, Thailand.

Charoenporn Thatsanee, Chotimongkol Ananlada and Sornlertlamvanich Virach, (1999). *Automatic romanization for Thai*, in Proceeding of the 2nd International Workshop on East-Asian Language Resources and Evaluation, Taipei, Taiwan, 1999.

Clément-Fontaine, Mélanie, (1999). *La licence publique générale GNU*, Mémoire de DEA, Université de Montpellier I, Faculté de Droit.

CMDL, Institutions et organisations non gouvernementales diverses, (1996). *Déclaration Universelle des Droits Linguistiques*, Conférence mondiale sur les droits linguistiques, Barcelone, 1996.

Cœdes Georges, (1964). *Les Etats hindouisés d'Indochine et d'Indonésie*, Paris, Boccard.

Cole Ron, Mariani Joseph, Uszkoreit Hans, Varile Giovanni-Batista, Zaenen Annie, Zampolli Antonio, Zue Victor, (1997). Survey of the State of the Art in Human Language Technology, Cambridge University Press et Giardini, 1997 ISBN 0-521-59277-1.

Comité Littéraire (ກົມວັນນະຄະດີ), (1962). *Lao grammar (ໄວຍາກອນລາວ)*, Vientiane 1962.

Comrie Bernard, (1991). *The World's Major Languages*, Oxford University Press, ISBN 0-415-04516-9.

Coyaud Maurice, (1995). *Les langues dans le monde actuel, graphies et phonies, tome 1*, P.A.F.

Coyaud Maurice, (1997). *Les langues dans le monde actuel, graphies et phonies, tome 2*, P.A.F., ISBN 2-902684-35-5.

Dalby Andrew, (1998). *Dictionary of languages*, Bloomsbury Publishing, ISBN 0-7475-3118-8.

Dalby David, (1999). *The Linguasphere register of the world's languages and speech communities*, Linguasphere press, 2 volumes.

Dale Robert, Moisl Hermann, Somers Harold, (2000). *Handbook of natural language processing*, Marcel Dekker, New York & Basel, ISBN 0-8247-9000-6.

Daniels Peters & Bright William, (1996). *The world's writing systems*. Oxford University Press, ISBN 0-19-507993-0.

Del Vigna Claude & Berment Vincent, (2004). *Ambiguïtés irréductibles dans les monoïdes de mots*. Journal de la Société Royale Belge de Mathématique, parution prévue en 2004.

Del Vigna Claude, (2003). *Web-powered databases : the low level in C++*. Chapitre 7 du livre Web-powered databases, Ed. D. Taniar et J. W. Rahayu, Idea Group Publishing, Hershey, Pennsylvania, USA, ISBN 1-59140-035-X.

Delouche Gilles, (1988). *Méthode de thaï, volume 1*. Langues de l'Asie – INALCO, L'Asiathèque, ISBN 2-901795-32-8.

Di Filippo, Éric, (1999). *Les logiciels libres*, Mémoire de DEA, Université de Nice-Sophia Antipolis.

Diz Gamallo Inés, (2001). *The importance of MT for the survival of minority languages: Spanish-Galician MT system*, MT Summit, 2001.

Doan-Nguyen Hai, (1998). *Techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes*, Mémoire de doctorat, UJF Grenoble 1998.

Dubois Jean, Giacomo Mathée, Guespin Louis, Marcellesi Christiane, Marcellesi Jean-Baptiste, Mével Jean-Pierre, (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse, ISBN 2-03-340334-3.

Ferlus Michel, (1988). *Langues et Ecritures en Asie du Sud-Est*, The 21st International Conference on sino-tibétaine Languages and Linguistics, University of Lund, Sweden.

Février James, (1959). *Histoire de l'écriture*, Grande Bibliothèque Payot, ISBN 2-228-88976-8.

Finot Louis, (1917). *Recherches sur la littérature laotienne*, Bulletin de l'ÉFEO, tome XVII n°5, Imprimerie d'Extrême-Orient, Hanoï 1917.

Frémy Dominique, Frémy Michèle, (1995). *Statistiques langues parlées*, in : QUID 96 p. 154-155, Paris : Laffont.

Gauthier François, Leclerc Jacques, Maurais Jacques, (1993). *Langues et Constitutions-Recueil des clauses linguistiques des Constitutions du monde*. Larousse, Office de la langue française (Québec) / Conseil international de la langue française (CILF), Québec/Paris, ISBN 2-551-15620-3.

Gelb I.J., (1973). *Pour une théorie de l'écriture*, Flammarion - Idées et Recherches (traduction de l'édition américaine de 1952 : A stydy of
writing, The university of Chicago Press).

Goodman Danny, (1998). *JavaScript, le guide du développeur*. ISBN: 2-7464-0347-1.

Grimes Barbara F., Grimes Joseph E. , editors, (2000). *Ethnologue. Volume 1: languages of the world; volume 2: maps and indexes. 2 vols; xi, 855, vi, 729 p.* SIL International, Dallas.

Hagège Claude, (2000). *Halte à la mort des Langues*, Odile Jacob.

Hospitalier J.J., (1937). *Grammaire laotienne*, Paul Geuthner, Paris.

Huffman Franklin E., (1970). *Cambodian system of writing and beginning reader*, Yale University Press.

Inthamone Lamvieng, (1987). *Je lis et j'écris lao (ອ້ອຍອ່ານອ້ອຍຂຽນພາສາລາວ)*, INALCO

Inthamone Lamvieng, (à paraître). *Je parle lao*.

Jurafsky Daniel, Martin James, (2000). *Speech and language processing*, Prentice Hall. ISBN: 0-13-095069-6.

Kanlayanawat Witoon, Prasitjutrakul Somchai, (1997). *Automatic Indexing for Thai Text with Unknown Words using Trie Structure*. In Proceedings of the Natural Language Processing Pacific Rim Symposium, p. 115-120, Phuket, Thailand, 2-4 décembre 1997 (NLPRS 97), ISBN 974-89570-9-8.

Karoonboonyanan Theppitak, Sornlertlamvanich Virach & Meknavin Surapant, (1997). *A Thai Soundex system for spelling correction*. Proceeding of the National Language Processing Pacific Rim Symposium, pp. 633-636, Phuket, Thailand, ISBN 974-89570-9-8.

Katzner Kenneth, (1995). *The Languages of the World*, Routledge, 3e édition, ISBN 0-415-11809-3.

Kawtrakul Asanee, Thumkanon Chalathip, Seriburi Sapon, (1995). *A Statistical Approach to Thai Word Filtering*. Proceedings of the 2nd Symposium on NLP.

Kawtrakul Asanee, Kumtanode Supapas, Jamjanya Thitima, Jewriyavech Chanvit, (1995). *A Lexibase model for writing production assistant system*. In Proceedings of the Symposium on Natural Language Processing in Thailand.

Kawtrakul Asanee, Thumkanon Chalatip, Poovorawan Yuen, Varasrai Patcharee, Suktarachan Mukda, (1997). *Automatic Thai unknown word recognition*. In Proceedings of the Natural Language Processing Pacific Rim Symposium, p. 341-346, Phuket, Thailand, 2-4 décembre 1997 (NLPRS 97), ISBN 974-89570-9-8.

Kloss Heinz, McConnel Grant, (1974). *Composition linguistique des nations du monde, Volume 1. L'Asie du Sud, secteurs central et occidental*., Presses de l'Université Laval, ISBN 2-7637-6710-9.

Kloss Heinz, McConnel Grant, (1978). *Composition linguistique des nations du monde, Volume 2. L'Amérique du Nord*., Presses de l'Université Laval, ISBN 2-7637-6869-5.

Kloss Heinz, McConnel Grant, (1979). *Composition linguistique des nations du monde, Volume 3. L'Amérique centrale et l'Amérique du Sud*., Presses de l'Université Laval, ISBN 2-7637-6884-4.

Kloss Heinz, McConnel Grant, (1981). *Composition linguistique des nations du monde, Volume 4. L'Océanie*., Presses de l'Université Laval, ISBN 2-7637-6967-5.

Kloss Heinz, McConnel Grant, (1984). *Composition linguistique des nations du monde, Volume 5. L'Europe et l'URSS*., Presses de l'Université Laval, ISBN 2-7637-7044-4.

Labrie Normand, (1999). *The Historical Development of Language Policy in Europe*, in "A Language Strategy for Europe, Retrospect and Prospect, Pádraig Ó Riagáin and Síle Harrington, editors, Dublin.

Lafourcade Mathieu, (1994). *Génie logiciel pour le génie linguiciel*, Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble 1.

Lepage Yves, (2000). *De l'analogie*, Exposé du 26 février devant le groupe Koin, Kyoto, Japon.

Leslie T., (1995). *Efficient approaches to subset construction*, Master thesis, Computer Science, University of Waterloo.

Mahsut Muhtar, Ogawa Yasuhiro, Sugino Kazue, Inagaki Yasuyoshi, (2001). *Utilizing Agglutinative Features in japonaise-Uighur Machine Translation*, MT Summit, 2001.

Malherbe Michel, (1983). *Les langages de l'humanité*, Paris : Seghers, ISBN 2-221-01243-7.

Mangeot Mathieu, (2002). *How to import an existing XML dictionary into the Papillon platform*, Séminaire Papillon 2002, 16-18 juillet 2002, NII, Tokyo, Japon.

Mason Marilyn, (2000). *Spelling issues for Haitian Creole Authoring and Translation Workflow*, International journal for language and documentation, numéro 4, avril 2000, ISBN 1468-5728.

Mason Marilyn, (2002). *Closing the digital divide – Issues in expanding localization efforts to minority languages*, The LISA Newsletter, volume X, numéro 2, 2002 (réimpression d'un article publié en avril 2001), ISBN 1420-3693.

Maspero Georges, (1929). *« Langues »* , in : Un Empire colonial français l'Indochine vol. 1, p. 63-80, Paris-Bruxelles : G. van Oest, 1929.

Meillet Antoine, Cohen Marcel (ouvrage collectif), (1952). *Les langues du monde,* 2 volumes, réimpression Slatkine 1981 – ISBN: 2-05-100254-1.

Meknavin Surapant, Charoenpornsawat Paisarn, Kijsirikul Boonserm, (1997). *Featured-based Thai word segmentation*, Natural Language Processing Pacific Rim Symposium (NLPRS 97), ISBN 974-89570-9-8.

Mel'čuk Igor, Clas André, Polguère Alain, (1995). *Introduction à la lexicographie explicative et combinatoire*, Editions Duculot, ISBN 2-8011-1106-6.

Microsoft Corporation, (1995). *Microsoft Word Developer's Kit Third Edition*, Microsoft Press, ISBN 1-55615-880-7.

Mittrapiyanuruk Pradit, Hansakunbuntheung Chatchawarn, Tesprasit Virongrong & Sornlertlamvanich Virach, (2000). *Issues in Thai text-to-speech synthesis: the NECTEC approach*, Proceedings of NECTEC Annual Conference 2000, Bangkok Thailand.

Mittrapiyanuruk Pradit, Sornlertlamvanich Virach, (2000). *The automatic Thai sentence extraction*, Thailand.

Moseley Christopher, Asher Ronald, (1994). *Atlas of the World's Languages*, Routledge Reference, ISBN 0-415-01925-7.

Palmer David D., (2000). *Tokenisation and sentence segmentation*, Handbook of Natural Language Processing, Robert Dale et al. Editors.

Paroubek Patrick, Rajman Martin, (2000). *Étiquetage morpho-syntaxique*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Paul Michael, (2001). *Translation Knowledge Recycling for Related Languages*, MT Summit, 2001.

Perret Michèle, (1998). *Introduction à l'histoire de la langue française*, Campus Linguistique, SEDES, ISBN 2-7181-9032-9.

Peterson James Lyle, (1980). Computer Programs for Spelling Correction, Lecture Notes in Computer Science, Vol 96, Springer-Verlag, New York, ISBN 0387102590.

Pierrel Jean-Marie, (2000). *Ingénierie des langues*, Ouvrage collectif, Hermes Science, ISBN 2-7462-0113-5.

Poowarawan Yuen, Imarom Wiwat, (1986). *Thai Syllable Separater by dictionary*, Proceedings of the Ninth Electrical Engineering Conference, Khonkhaen, Thailand.

Rarunrom Sampan, (1991). *Dictionary-based Thai word separation,* Senior project report, Thailand.

Rastier François, Cavazza Marc, Abeillé Anne, (1994). *Sémantique pour l'analyse*, Masson, ISBN 2-225-84537-9.

Reinhorn Marc, (1970). *Dictionnaire laotien-français*, CNRS.

Reinhorn Marc, (1975). *Grammaire de la langue lao*, INALCO.

Rey Alain, (2001). *Le Grand Robert de la langue française, 2e édition*, (Ouvrage collectif dirigé par Alain Rey), ISBN 2-85036-673-0.

Romary Laurent, (2000). *Outils d'accès à des ressources linguistiques*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Roop D . Haigh, (1972). *An introduction to the Burmese writing system*, Yale University Press.

Ruhlen Merritt, (1987). *A Guide To The World's Languages, volume 1*, Stanford University 1987, ISBN 0-8047-1250-6.

Sabah Gérard, (1990). *L'intelligence artificielle et le langage, volume 1, 2e édition, Représentation des connaissances*, Hermes, ISBN 2-86601-134-1.

Sabah Gérard, (1989). *L'intelligence artificielle et le langage, volume 2, Processus de compréhension*, Hermes, ISBN 2-86601-187-2.

Sala Marius, Vintila-Radulescu Ioana, (1984). *Les langues du monde, petite encyclopédie (traduit du roumain)*. Paris : Société d'édition « les belles lettres », ISBN 2-251-37401-9.

Sampson Geoffrey, (2001). *What is a minority language*, in Elsnews 10.1 (printemps 2001), ISSN 1350-990X.

Sérasset Gilles, (1994). *SUBLIM: un Système Universel de Bases Lexicales Multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble 1.

Shimohata Sayori, Kitamura Mihoko, Sukehiro Tatsuya, Murata Toshiki, (2001). *Collaborative Translation Environment on the Web*, MT Summit, 2001.

Somers Harold, (1997). *Machine Translation and Minority Languages*, Translating and the Computer 19, Papers from the ASLIB Conference 13/14 November 1997.

Sornlertlamvanich Virach, (1993). *Word segmentation for Thai in a machine translation system*, Thailand *(en thaï)*.

Sornlertlamvanich Virach, Potipiti Tanapong, Wutiwiwatchai Chai, Mittrapiyanuruk (2000). *The state of the art in Thai language processing*, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000), Hong Kong, pp 597-598, October 2000.

Teston Eugène, Percheron Maurice, (1931). *L'ethnographie, la linguistique*, in : Indochine moderne p. 289-310. Paris : Librairie de France.

Thairatananond Yupin, (1981). *Towards the design of a Thai text syllable analyser*, Master thesis, Asian Institute of Technology, Thailand.

Theeramunkong Tharanuk, Usanavasin Sasiporn, Machomsomboon Tanin and Opasanont Borisuth, (2000). *Thai word segmentation without a dictionary by using decision trees*, Thailand.

Thierry S., (1980). *Thaï*, in : Encyclopædia Universalis : vol. 15 p. 1031-1033. Paris, Encyclopædia Universalis.

The Unicode Consortium, (2000). *The Unicode Standard Version 3.0*, Addison-Wesley, ISBN 0-201-61633-5.

Véronis Jean, (2000a). *Alignement de corpus multilingues*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Véronis Jean, (2000b). *Annotation automatique de corpus : panorama et état de la technique*, in : Ingénierie des langues, sous la direction de Jean-Marie Pierrel, Hermes Science, ISBN 2-7462-0113-5.

Voegelin Charles Frederick, Voegelin Florence Marie, (1977). *Classification and index of the world's languages*, Elsevier, New York, ISBN 0-444-00155-7.

Watson Bruce W., (1993). *A taxonomy of finite automata minimization algorithms*, Université d'Eindhoven, Pays-Bas.

Watson Bruce W., (1994). *An introduction to the FIRE engine: A C++ toolkit for FInite automata and Regular Expressions*, Université d'Eindhoven, Pays-Bas.

Wehrli Éric, (1997). *L'analyse syntaxique des langues naturelles*, Masson, ISBN 2-225-85432-7.

Wright Marcus, Briggs Will, Van Buskirk Robert, Harmer Craig, (2001). *The use of international encoding standards for local language computer software, documents and data*. Independent Eritrea: Lessons and Prospects, The International Conference Commemorating The 10th Anniversary Of The Independence Of Eritrea, July 2001, Asmara, Eritrea.

Wurm Stephen, Heywards Ian, (2001). *Atlas of the world's languages in danger of disappearing, second edition*, UNESCO publishing, ISBN 92-3-103255-0.

Yacoub Joseph, (1998). *Les minorités dans le monde, faits et analyses*, Desclée de Brouwer, ISBN 2-220-04171-9.

Yeu Tri, (2003). *Découpage syllabique des langues du Sud-Est Asiatique*, Université René Descartes (Paris 5).