Langues peu dotées

Définition et problématiques pour le TALN et le TALP

Vincent Berment

C&S - LIG/GETALP

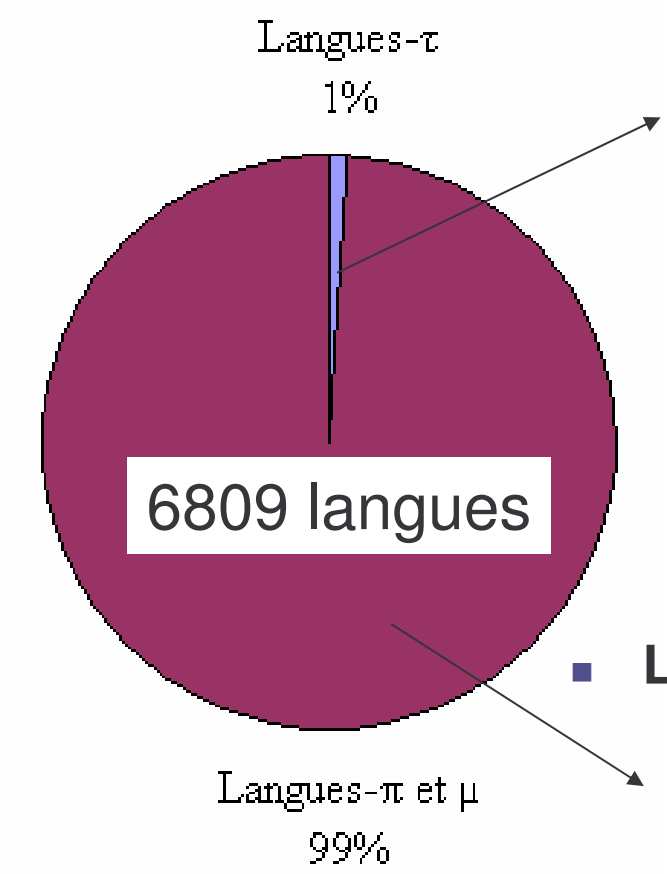
Une mesure du niveau d'informatisation : l'indice-σ

	Services / ressources	Criticité (de 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**

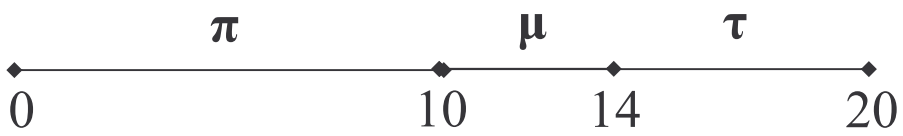
Langues-π et Nations-Unies (Unesco)

- Déclaration universelle sur la diversité culturelle
  - Novembre 2001
  - Extrait : « Toute personne doit ainsi pouvoir **s'exprimer, créer et diffuser ses œuvres** dans la langue de son choix et **en particulier dans sa langue maternelle** »
- Recommandation sur la promotion et l'usage du multilinguisme
  - Octobre 2003
  - Extrait : « Les États membres, les organisations internationales et les entreprises spécialisées dans les technologies de l'information et de la communication 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. »

Langues bien et mal dotées informatiquement



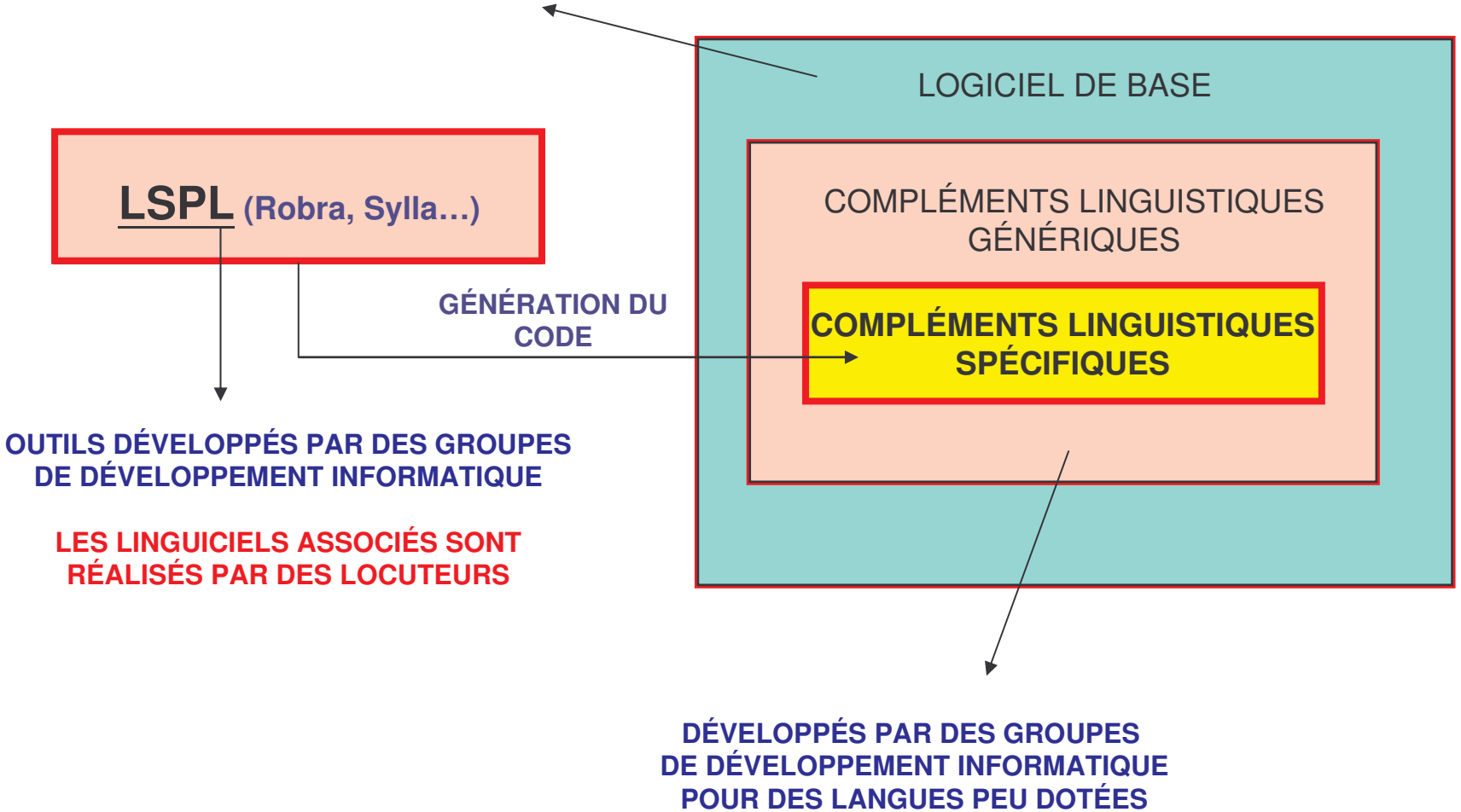
- **Langues Très bien dotées informatiquement**
  - **Quelques dizaines de langues-τ**
    - Allemand, anglais, français, japonais, russe…
  - **Informatisation rentable => éditeurs de logiciels**
    - Apple, IBM, Microsoft, Xerox…



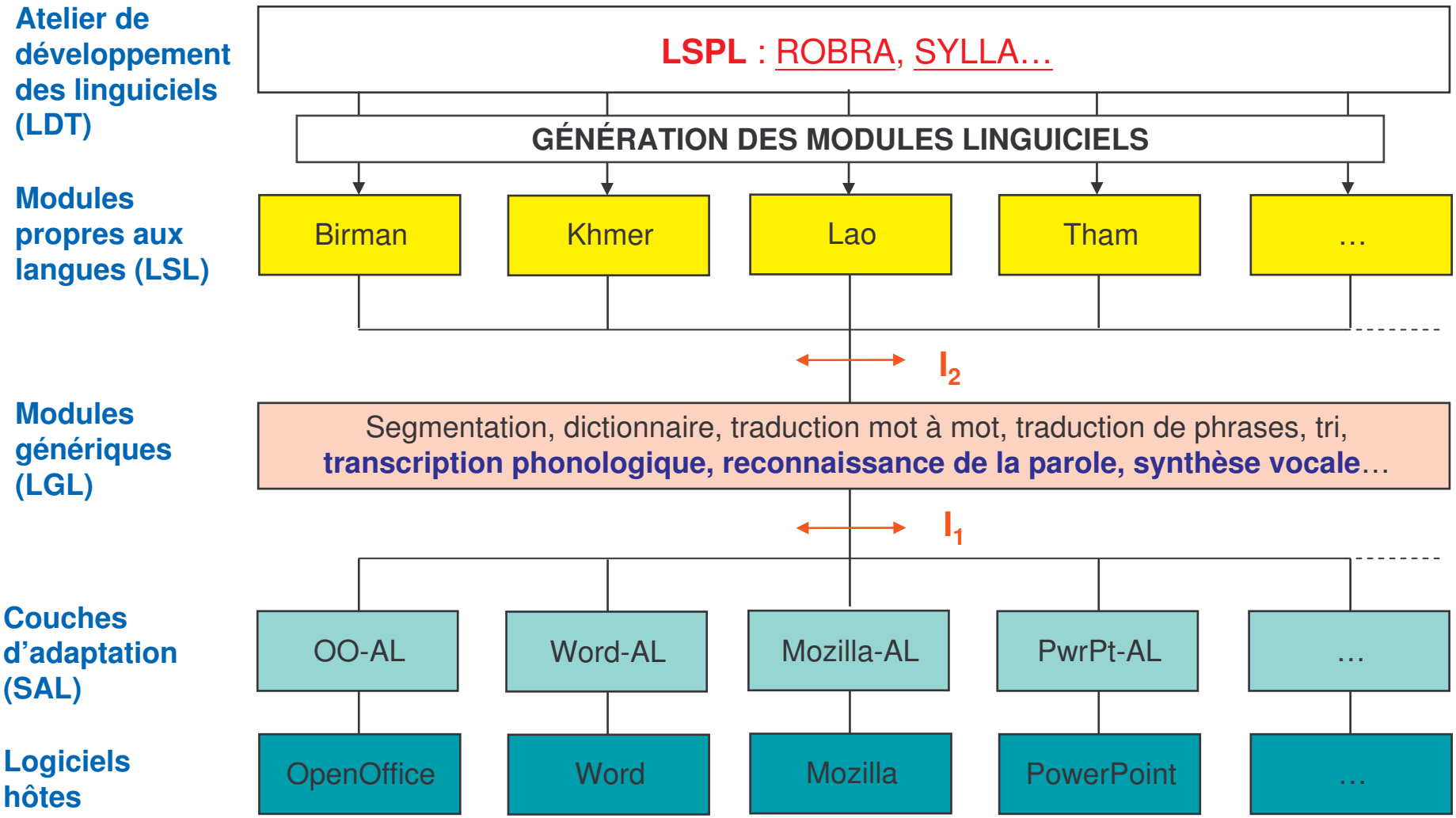
- **Langues Peu ou Moyennement dotées**
  - **Plus de 6000 langues-π et μ**
  - **Informatisation pas ou peu rentable => autres**
    - Groupes de locuteurs créés spontanément
    - Projets de développement…

Une méthodologie adaptée

DÉVELOPPÉ PAR LES GRANDS ÉDITEURS DE LOGICIELS
POUR DES LANGUES BIEN DOTÉES
DOIT PRÉSENTER DES INTERFACES STANDARDISÉES



Des groupes locaux s'investissent pour leur langue



Gain obtenus pour la saisie grâce à la généricité

Coût la 1ère fois (laotien : LaoWord)	Coût les fois suivantes (bengali : BanglaWord)	**Gain**
250 h	18 h	**92,8 %**
5 millions de locuteurs au Laos	300 millions de locuteurs au Bengladesh	

Ici, on a substitué le module spécifique au lao par un module spécifique au bengali.

Formule du gain de réutilisation

Économie = (nL*tL*rL*(nE-1)+nE*tE*rE*(nL-1)) / (nL*nE*(tL+tE))

nL / nE = nombre de langues / d'environnements
tL / tE = temps de développement du code linguistique / générique
rL / rE = taux de réutilisation du code linguistique / générique

- **Si nE=4, tL=1000 heures, tE=500 heures, rL=rE=95 %**

Nb langues	Économie	Économie (en heures)
2	63,33 %	7 600 (4 400 au lieu de 12 000)
5	72,83 %	21 850 (8 150 au lieu de 30 000)
10	76,00 %	45 600 (14 400 au lieu de 60 000)
100	78,85 %	473 100 (126 900 au lieu de 600 000)
1000	79,14 %	4 748 100 (1 251 900 au lieu de 6 000 000)

Un projet d'informatisation de plusieurs langues

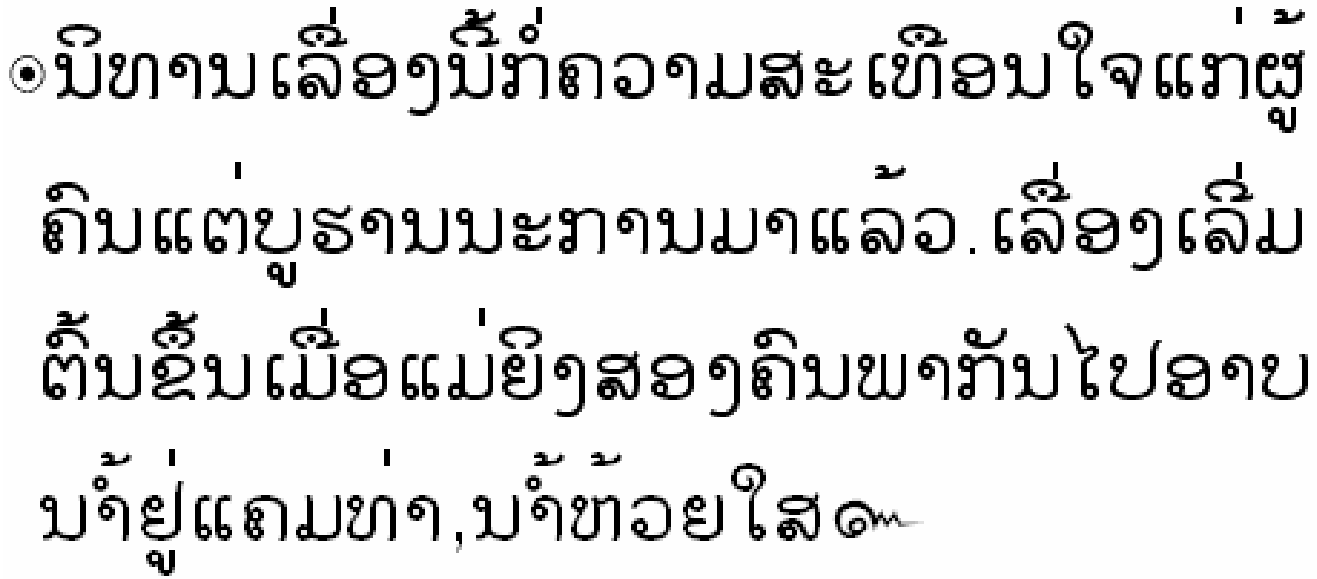
- Projet PAN Localization
  - Lancé par le Centre de Recherche pour le Développement International (Canada)
  - Mené par le Centre de traitement de l'Ourdou (Pakistan)
- Budget : 1,3 M€
- Informatisation de six langues : bengali, dzongkha (Bhoutan), khmer, lao, népali, cingalais

Résultats du projet PAN Localization

- Quelques factorisations :
  - Saisie (bengali, lao)
  - Conversion Unicode (khmer, cingalais)
  - Correcteur d'orthographe (bengali, khmer)
  - Tri (bengali , khmer)
  - Reconnaissance de caractères (bengali, cingalais)
- Réussite ?
  - + Sensibilisation des équipes locales
  - - Non-réutilisation de l'existant venant d'Occident
  - - Refus de faire participer les spécialistes occidentaux
  - - Résultat de niveau inférieur à l'état de l'art

Autre projet (mené au LIG / GÉTALP / GÉTA)

Segmentation de textes en mots



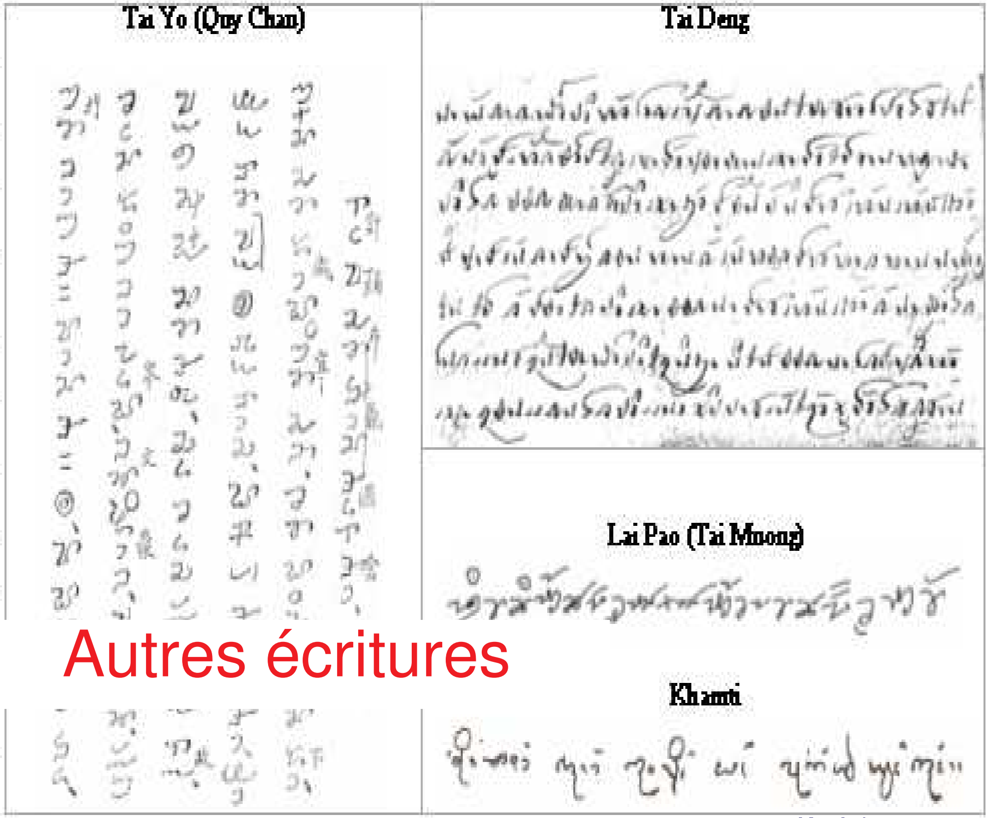
Ce texte en lao ne contient pas d'espaces (non segmenté)

Environ 30 systèmes d'écriture en Asie du Sud-Est

Texte en lao

◉ນິທານເລື່ອງນີ້ກໍ່ຄວາມສະເທືອນໃຈແກ່ຜູ້ຄົນແຕ່ບູຮານນະການມາແລ້ວ. ເລື່ອງເລີ່ມຕົ້ນຂຶ້ນເມື່ອແມ່ຍິງສອງຄົນພາກັນໄປອາບນ້ຳຢູ່ແຄມທ່າ,ນ້ຳຫ້ວຍໃສ໑

Famille	Sous-Famille	Système d'écriture	Exemple
Khmer	Khmer	Khmer	មានរឿងមួយដំណាលថា មានភិក្ខុ ១ អង្គ
	Lao-thai	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]



Autres écritures

Algorithmes de segmentation en vue de la traduction

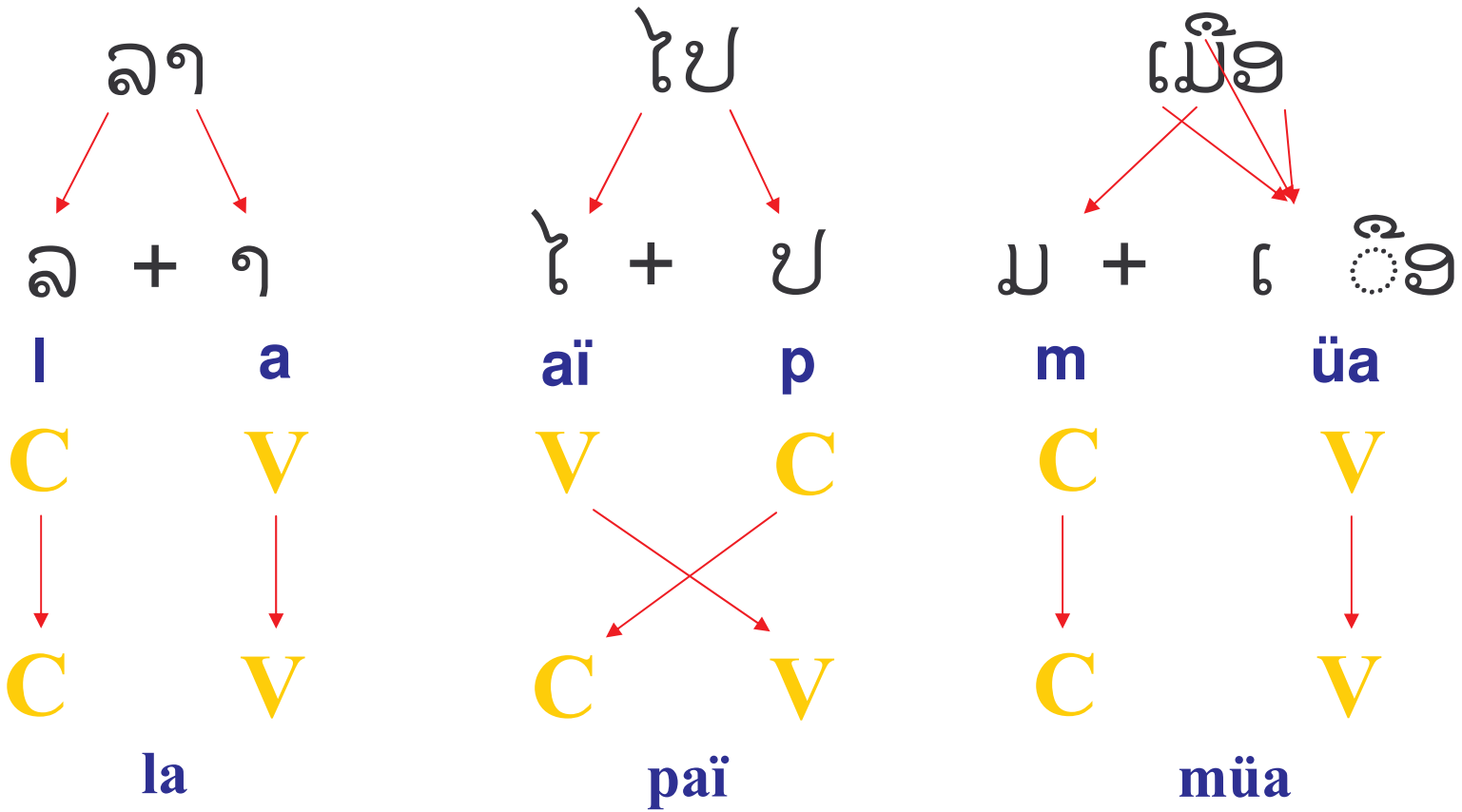
- On segmente le texte en syllabes
  - ສະບາຍດີທຸກໆທ່ານ → ສະ-ບາຍ-ດີ-ທຸກ-ໆ-ທ່ານ
- On regroupe les syllabes pour former des mots contenus dans le dictionnaire (algorithme de « plus longue chaîne d'abord »)
  - ສະ-ບາຍ-ດີ-ທຸກ-ໆ-ທ່ານ → ສະບາຍດີ-ທຸກ-ໆ-ທ່ານ
- On présente le résultat
  - ສະບາຍດີ (bonjour) ທຸກ (tout, tous) ໆ (?) ທ່ານ (personne)

La reconnaissance des syllabes est complexe

Exemple sur le lao

Forme générale des syllabes = **C [C] [A] V [C]**

Formes et positions problématiques des voyelles



Le nombre des syllabes peut être contraignant

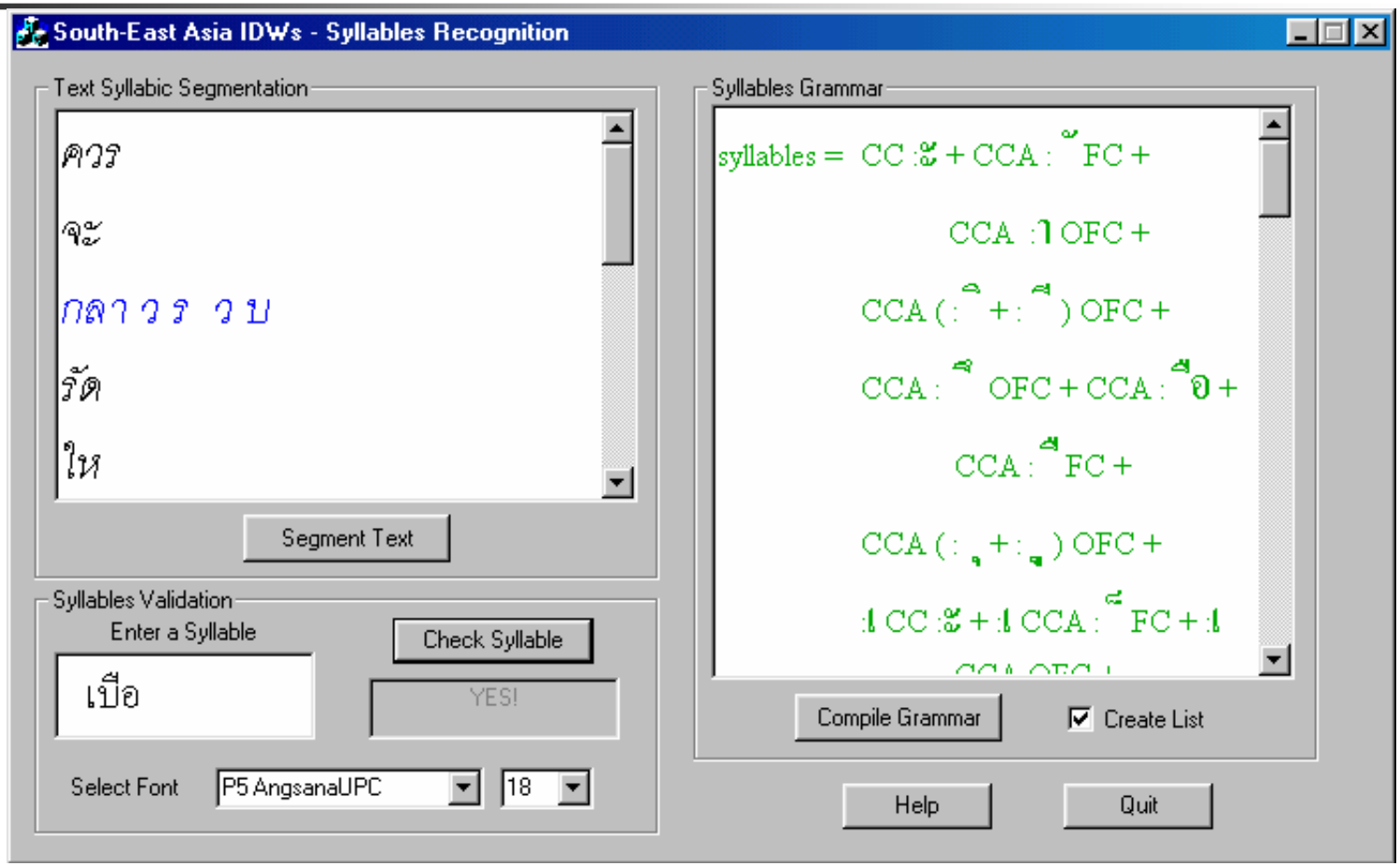
- **Birman** :
  - Forme générale : (C ou CS) [L] V [C ou ◌̊ ou ◌̊ɛ [◌̊ᴄ] [◌̥ ou ◌ː]]
  - Cardinaux des constituants : |C| = 33, |CS| = 20, |L| = 15, |V| = 35
  - Majorant : (33+20)x16x35x(34+2)x2x3 = **6 410 880 syllabes**
- **Khmer** :
  - Forme générale : (C [CS [CS]] [D1] ou CS [CS]) V [C [CS] [D2]] ou VI ou L
  - 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**
- **Laotien** :
  - Forme générale : (C ou GC) [A] V [CF]
  - Cardinaux des constituants : |C| = 27, |A| = 4, |GC| = 36, |V| = 38 et |CF| = 8
  - Majorant : (27+36)x5x38x9 = **95 760 syllabes**
- **Siamois** (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**

Syllabes : expression hors-contexte

Syllabes =	CC :ะ + CCA :ั CF + CCA :า CFO + CCA (:ิ + :ี + :ึ + :ี + :ุ + :ู) CFO + …
CCA =	CC + CC Acc ;
CC =	GC + CI ;
GC =	:ห (:ง + :ย + :ม + :ม + :ล + :ั + :อ) + CI :อ ;
CI =	:ก + :ข + :ค + :ง + :จ + :ส + :ช + :ย + :ด + :ต + :ถ + :ท + :น + :บ + :ป + :ผ + :ฝ + :พ + :ฟ + :ม + :ย + :ร + :ล + :อ + :ห + :อ + :ร + :ฆ + :ฆ ;
CFO =	CF + {} ;
CF =	:ก + :ด + :บ + :ง + :น + :ม + :ย + :อ ;
Acc =	: + :่ + :๊ + :๋ ;

- Appropriable par un linguiste
- Testé sur le laotien et sur le khmer

Le LSPL SYLLA



- Mise au point des automates de reconnaissance des syllabes
  - Réduit le temps de développement d'environ 80 % (60 h au lieu de 300 en moyenne)
  - Facilement appropriable
  - Utilisé pour : birman, khmer, laotien et thaï

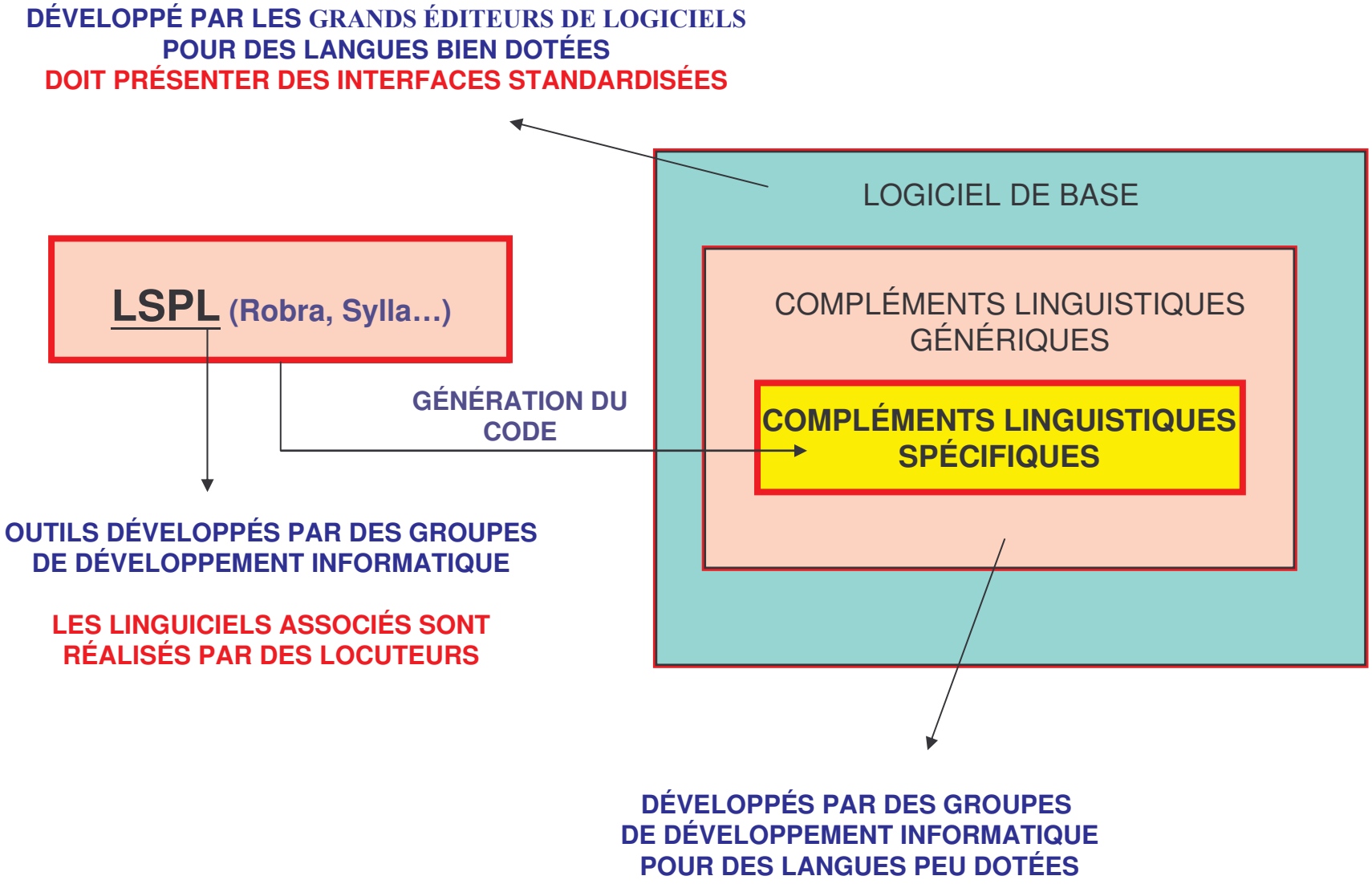
Applications au traitement de la parole

- Transcription phonétique
  - Travaux des thaïs (Thai Soundex, Thai TTS)
  - Transcriptions sur LaoSoftware.com
- Synthèse vocale et reconnaissance de la parole
  - Travaux du LIG
    - Laurent Besacier
    - Le Viet Bac
    - Sethserey Sam
  - Travaux des thaïs (Thai Soundex, Thai TTS)

Problème résiduel mais essentiel des ambiguïtés

- Exemple d'ambiguïté (thaï)
  - ผม-ตาก-ลม → ผมตากลม → ผม-ตา-กลม
  - (« je prends l'air » / « j'ai les yeux globuleux »)
- Exemples de motifs irréductibles (lao)
  - ແກງອວ(ວອວນອວ)*ງອັງ
  - ເກືອວ(ວ)*ບອວ(ວອວນອວ)*ງອັງ

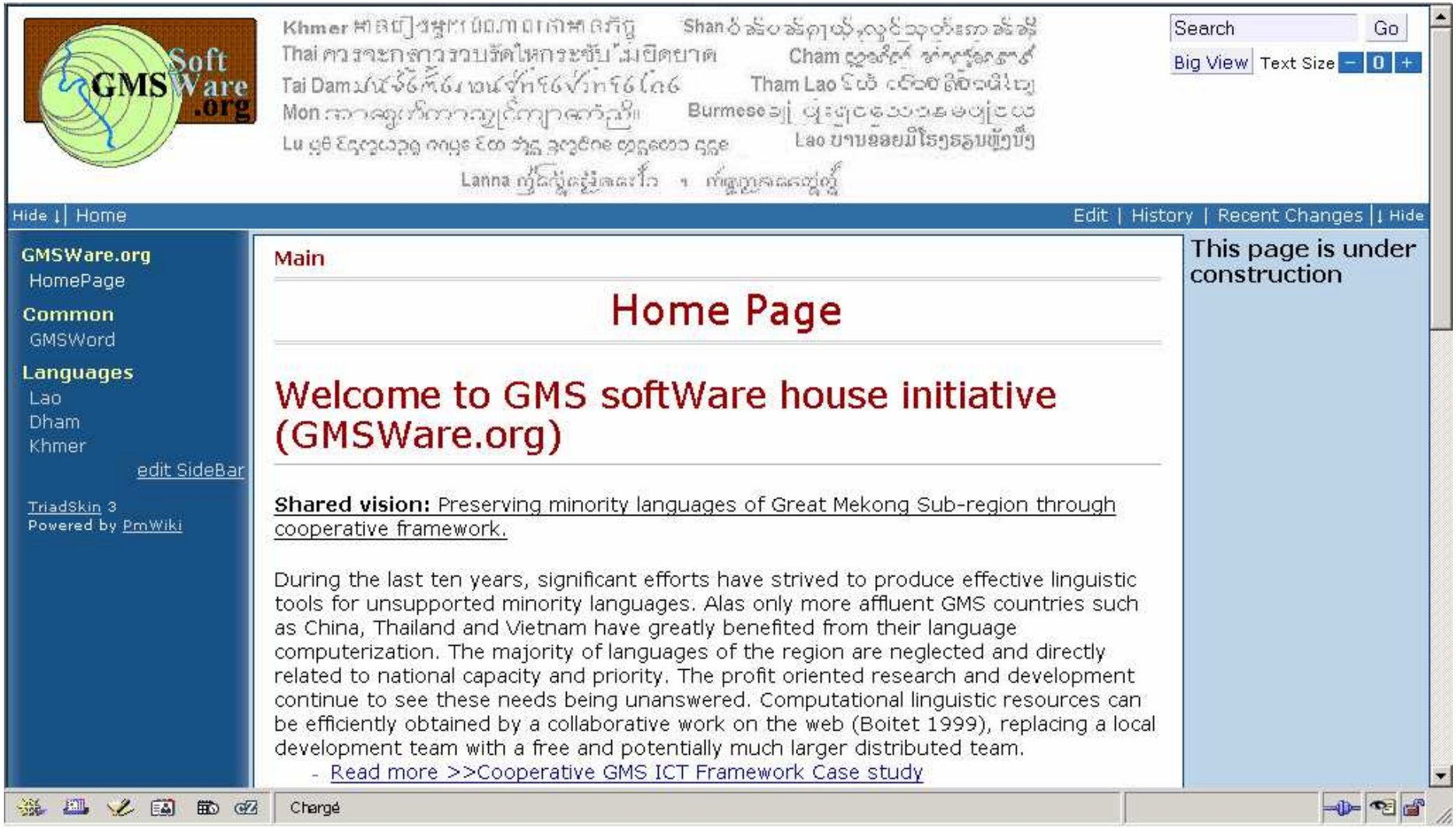
En conclusion : une répartition des tâches…



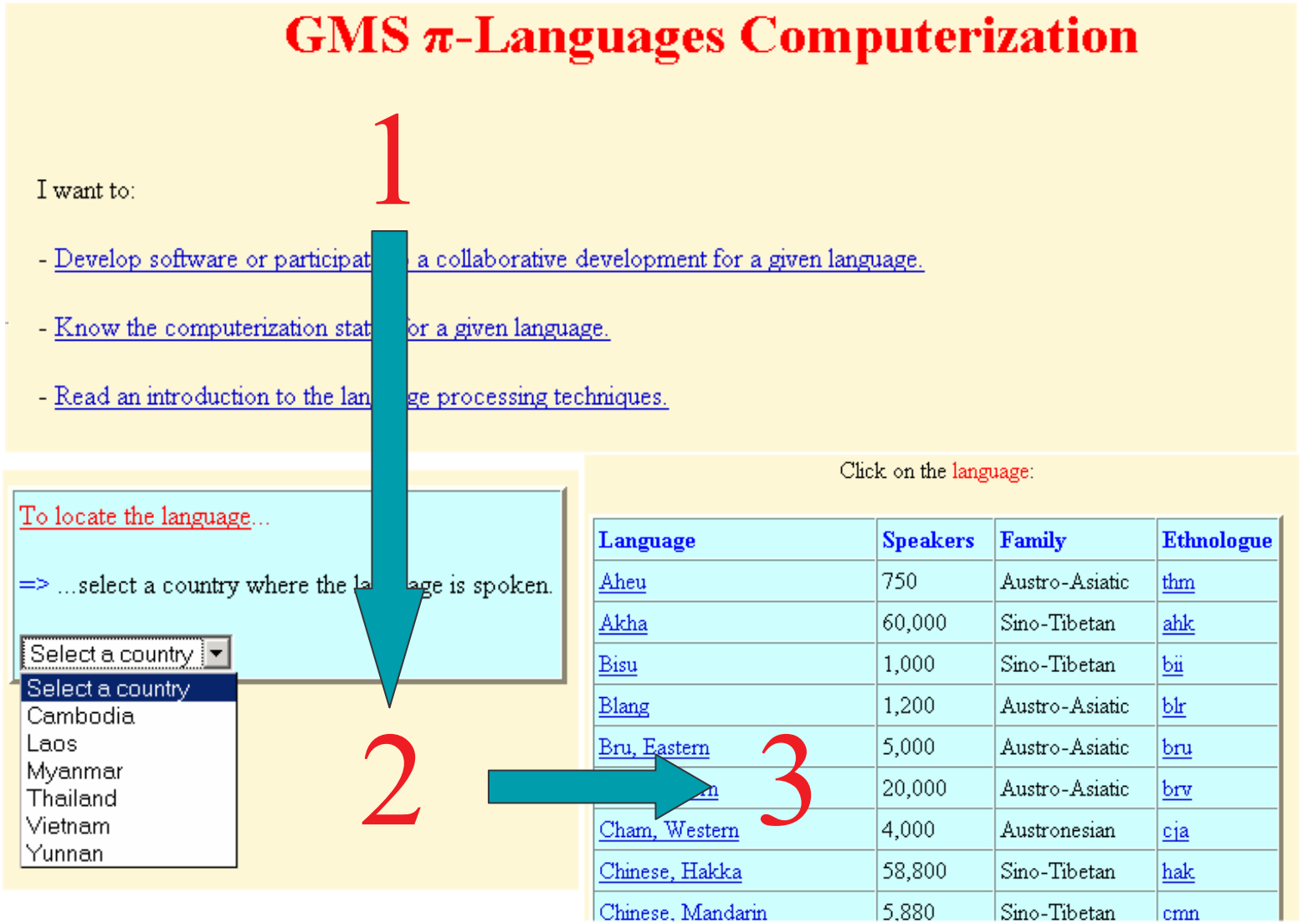
En conclusion : … associée à une méthodologie

- Normaliser les interfaces entre modules
- Définir et réaliser les parties génériques
- Définir les LSPL et les outils les compilant
- Développer des linguiciels pour quelques langues, qui serviront de modèles
- Permettre la prise en charge des langues par leurs locuteurs (outils en ligne)

Site collaboratif pour un groupe de langues



Guider les candidats à la contribution



Merci de votre attention

Questions ?