Corpus oraux glosés : outils logiciels d'aide à l'analyse

Michel Jacobson

LACITO - CNRS – 7, rue Guy Môquet, Bât 23 – 94801 Villejuif Cedex – France
jacobson@idf.ext.jussieu.fr

Abstract

Field linguistics works on written data, usually created through the analysis of oral recordings. The structure and the type of searches one may desire can be specific to aspects of sound or temporality. Apart from these few characteristics, the methods used to study this type of data have many points in common with those used in textual linguistics.
To illustrate these common points, we will examine several cases of recurrent tasks in field linguistics: morphemic analysis and glossing. We will see how recourse to the study of segments contexts (words, morphemes) given by concordances, or recourse to the calculation of the frequency of use of these units can help to carry out these tasks, or even partially automate them.
To finish, we will present a program designed to help gloss oral data by compiling, alongside the data, a lexicon of all the glosses already used. We will also examine how to use this tool to implement the methods cited above to optimize its usefulness for linguists.

Résumé

La linguistique de terrain travaille sur des corpus écrits, en général issus d'analyses d'enregistrements oraux. La structure et le type de requêtes que l'on peut formuler sur ces corpus peuvent être spécifiques de l'aspect sonore ou temporel. Mis à part ces caractéristiques, les méthodes utilisées pour l'étude de ces corpus ont de nombreux points en commun avec celles utilisées en linguistique textuelle.
Pour illustrer ces points communs, nous examinerons quelques cas de tâches récurrentes en linguistique de terrain : la segmentation et la détermination de gloses. Nous verrons notamment comment le recours à l'examen des contextes d'apparition de segments (mots, morphèmes) donnés par des concordances, ou bien comment le calcul des fréquences d'apparition de ces unités peuvent nous aider à effectuer ces tâches ou bien même à les automatiser en partie.
Nous présenterons enfin un logiciel créé pour aider à gloser des corpus oraux en entretenant parallèlement au corpus, un lexique de toutes les gloses déjà utilisées. Nous examinerons également comment implémenter dans cet outil les méthodes citées plus haut pour optimiser l'aide apportée au linguiste.

Mots-clés : linguistique de terrain, corpus oraux, langage de balisage de texte.

1. Présentation

Beaucoup de laboratoires de recherche en linguistique, dont le LACITO[1] auquel nous appartenons, détiennent et gèrent des matériaux récoltés lors d'enquêtes de terrain, notamment des enregistrements de parole (récits, dialogues, questionnaires, etc.). Ces enregistrements ont fait, pour partie, l'objet d'analyses comportant la plupart du temps une transcription « phonétique », mais aussi parfois des traductions, des gloses, des indications scénographiques, etc. L'ensemble de ces données (enregistrements et analyses) pour une langue donnée constitue un

[1] Laboratoire « LAngues et CIvilisations à Tradition Orale » du CNRS. Le LACITO est un laboratoire dont les chercheurs (linguistes, anthropologues et ethno-musicologues) travaillent depuis plus de 30 ans à la constitution d'une documentation des traditions orales un peu partout dans le monde.

corpus sur lequel les chercheurs vont pouvoir continuer à travailler après leur mission sur le terrain.

Les enregistrements ont été stockés suivant les époques, sur des supports (mécaniques, magnétiques, optiques) analogiques puis plus récemment numériques. Les analyses ont, elles, été stockées parfois sous forme manuscrite, d'autres fois sous forme imprimée et pour les plus récentes d'entre elles, il peut exister une forme électronique informatique dans des formats divers : traitement de texte, base de données ou texte structuré de type Shoebox (Buseman. et Buseman, 1998), Lexware (Hsu, 1989), etc.

1.1. *Les spécificités de la linguistique de terrain*

Ces données (enregistrements et analyses) représentent souvent la rare documentation dont disposent les linguistes sur leurs langues. Ces langues étant souvent minoritaires, peu étudiées et parfois en danger de disparition ou même déjà disparues, les quelques rares données collectées, constituent à la fois un patrimoine culturel irremplaçable et du matériel précieux pour la recherche.

Le volume et la nature même des données disponibles ne nous permettent pas d'utiliser tous les outils que l'on pourrait s'attendre à trouver pour des langues bien documentées comme l'anglais ou le français. Au mieux, nous disposons pour une langue donnée : d'un corpus de quelques heures d'enregistrements accompagnés d'une description au niveau morpho-phonologique, quelquefois d'un lexique ou d'un dictionnaire de quelques milliers de mots, et éventuellement d'une ébauche de grammaire. Les enregistrements dont nous disposons sont assez éloignés des conditions de laboratoire en chambre sourde, d'une part parce qu'ils ont été pratiqués sur le terrain en situations naturelles (récits avec un conteur et un public, cérémonies, dialogues dans la rue), d'autre part parce que le matériel d'enregistrement utilisé devait être aussi léger et autonome en énergie que possible. Il n'est pas possible non plus, étant donné le peu de personnes travaillant sur ces langues de trouver sur le marché des outils de type analyseur syntaxique, ou lemmatiseur… La lourdeur de mise en œuvre de certains de ces outils constitue souvent un obstacle pour des utilisations ponctuelles ou pour des trop petits volumes de données, et rend les interventions manuelles compétitives en termes d'effort et d'investissement en temps.

1.2. *Les méthodes utilisées*

L'examen des données d'un corpus emprunte en partie ses méthodes aux disciplines du traitement du signal. En particulier, l'examen d'un enregistrement oral peut aujourd'hui se faire facilement avec des outils logiciels appliquant des techniques DSP[2] pour présenter le signal de parole sous forme de spectrogrammes, pour faire des analyses formantiques, des extractions de fondamentale, etc. Toutes ces techniques concourent pour le linguiste à se faire une meilleure opinion des contraintes phonético-phonologiques et donc à améliorer son système de transcription.

Une fois établie une première transcription, considérée comme une analyse et non comme une donnée, la plupart des méthodes utilisées par la suite pour questionner, structurer ou enrichir cette annotation sont inspirées des traditions d'études textuelles. En particulier, les linguistes font un usage intensif de concordances. Celles-ci permettent de se faire une meilleure idée des conditionnements qui peuvent exister entre les segments au niveau phonologique. Elles sont aussi utilisées par les linguistes pour distinguer les différents usages d'une forme lexicale afin

[2] Digital Signal Processing.

d'en extraire les différents sens pour la constitution d'un dictionnaire. Un autre outil simple est l'utilisation des fréquences d'apparition de certaines unités, ceci afin de repérer les formes rares, déviantes, voir même les erreurs de transcription.

Dans ce qui suit, nous allons présenter deux types d'activités habituelles aux linguistes de terrain (la segmentation et la détermination de gloses) dans lesquelles les méthodes de travail sur corpus que nous venons de citer sont utilisées en tant qu'aide.

2. Les aides à la segmentation

La segmentation est une opération récurrente que le linguiste utilise à différentes étapes de son travail d'analyse. La transcription « phonétique » telle que la pratique les linguistes, en utilisant des systèmes de notation alphabétiques de type API (Alphabet Phonétique International), présuppose déjà une segmentation implicite de la parole en unités discrètes. D'autres étapes de segmentation en unités plus larges telles que les morphèmes, les mots ou les syntagmes, s'aident d'analyses pratiquées sur des extraits de corpus. Par exemple, on peut facilement concevoir un logiciel d'aide au découpage morphologique, lequel partant d'une transcription d'un mot, proposerait en priorité les segmentations déjà rencontrées, voire les plus fréquentes ou les plus probables.

La segmentation par rétroconversion

Les transcriptions sur lesquelles nous travaillons d'habitude, marquent les unités lexicales et morphologiques au moyen de conventions typographiques. Il est beaucoup plus rare de rencontrer des transcriptions « au kilomètre ». Parfois, le découpage morphologique n'est pas indiqué ou ne l'est pas systématiquement. Cela correspond en général au choix de l'analyste de ne pas pousser son analyse plus loin, par exemple en raison de la difficulté de la tâche (présence de nombreux amalgames, etc.). La segmentation revient donc, la plupart du temps, à utiliser ces indices typographiques pour isoler et étiqueter de manière explicite les segments. C'est ce que l'on appelle la rétroconversion.

La « rétroconversion » est une tâche qui consiste à retrouver la structure logique sous-jacente ayant donné lieu à une structure physique particulière. Suivant les moyens utilisés pour le codage de l'analyse, cette étape de rétroconversion sera ou non complexe et nécessaire. Par exemple, dans les cas que nous traitons dans notre laboratoire, une grosse partie des analyses ayant été faite depuis assez longtemps, parfois avant la popularisation de la micro-informatique, la structure physique est une structure typographique destinée à l'impression sur un support papier. Heureusement, ces documents utilisent la plupart du temps une partie relativement standardisée de conventions de notation, comme par exemple l'utilisation des espaces blancs pour séparer les mots ou bien celle des tirets ou des points pour indiquer les jointures de morphèmes. La rétroconversion peut dans ce cas être effectuée par une simple segmentation basée sur des séparateurs typographiques, et être réalisée par des jeux d'expressions régulières.

Beaucoup de nos documents d'analyse utilisent ces conventions dans un style de présentation « interlinéaire » assez habituel pour les linguistes. Dans un texte interlinéaire, les unités d'analyses mots se suivent traînant avec eux sur plusieurs lignes toute leur analyse: transcription(s), glose(s), etc. La ligne typographique est donc utilisée à la fois pour noter la succession des unités mais aussi les différentes analyses linguistiques de ces unités.

La gestion de l'alignement interlinéaire, suivant le système utilisé pour sa visualisation, peut être effectuée soit par des calculs sur la taille en caractères des unités (Buseman et Buseman,

1998), soit par des structures de données plus abstraites comme les tableaux HTML, ou des boites LaTex, soit enfin par des calculs sur de simples séparateurs de champs comme les caractères de tabulation. La rétroconversion, c'est-à-dire la redécouverte de la segmentation sous-jacente sera, suivant les cas, plus ou moins aisée. Elle peut être effectuée en général, par des jeux d'expressions régulières mais demande parfois quelques interventions manuelles.

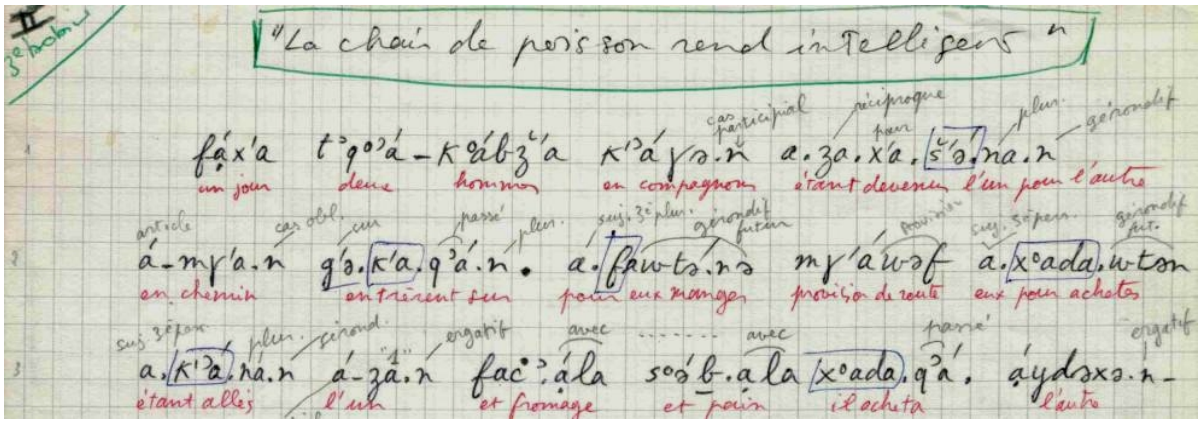

Figure 1. Exemple d'un texte[3] *manuscrit dans une présentation « interlinéaire » [extrait d'un cahier manuscrit attribué à Georges Dumézil].*

En règle générale, les solutions utilisées restent *ad hoc*. Les techniques employées en Traitement du Langage Naturel, à base d'automates à états finis (Kaplan et Martin, 1995), de modèles Markoviens (Rabiner, 1989), de règles sur des connaissances, ou bien à base de systèmes mixtes (Spoart *et al.*, 1996), sont souvent plus lourdes à mettre en œuvre qu'une correction manuelle ou qu'une re-saisie complète des données, à moins d'avoir à traiter de gros volumes de données ce qui est rarement le cas pour les langues que nous étudions.

3. Les aides à la détermination des gloses d'un texte

Une autre tâche assez récurrente dans la constitution de corpus oraux est la détermination des gloses des unités. Par cette opération, nous entendons l'action d'associer un commentaire à une unité d'analyse dans le but d'apporter une explication ou une définition de cette unité. Pour des niveaux d'analyse assez larges, comme le texte ou la phrase, on parlera plus facilement de traduction. Pour des niveaux plus bas d'analyse, comme le mot ou le morphème, on parlera plus volontiers de glose. Cette glose peut être plus ou moins technique, normalisée et en inventaire fini. Cette opération est délicate dans la mesure où elle nécessite une bonne connaissance de la langue en question ainsi qu'une rigueur et un systématisme. Elle est aussi une tâche extrêmement répétitive et fastidieuse car en règle générale il existe dans toute les langues un petit ensemble d'unités très souvent utilisées, et qu'il faut donc gloser de nombreuses fois.

Il s'agit donc d'une situation idéale pour proposer une aide de type logiciel. Un logiciel pourrait permettre à un utilisateur de choisir ses gloses dans un ensemble fini de gloses, ce qui lui permetrait d'être plus facilement systématique et d'éviter de gloser une même unité par des variantes de type : Passé, PASSE. Past, … Un logiciel pourrait aussi permettre de proposer en fonction de critères comme la forme de l'unité, la glose qui lui correspond s'il n'y en a

[3] Texte oubikh (langue du Caucase, aujourd'hui disparue). Il est possible de consulter ce conte (enregistrement, transcription et traduction) sur le site web du LACITO.

qu'une, ou les gloses possibles s'il y a de l'homophonie. Il pourrait aussi pour les cas d'homophonie, trier les propositions afin de faire apparaître en premier les plus probables.

C'est principalement pour des raisons économiques (gain de temps) que nous avons envisagé dans notre laboratoire de construire un outil logiciel d'aide au linguiste pour cette tâche.

3.1. *Description du logiciel*

Le logiciel créé (Interlinear Text Editor[4]) permet de présenter, pour un même document XML (Bary *et al.*, 1998), quatre niveaux d'analyse distincts : le texte, les phrases, les mots et les morphèmes. Il offre pour chacun de ces niveaux une représentation adéquate. Le mode de représentation d'un texte ou d'une phrase permet la visualisation et l'édition de sa transcription et de sa traduction. Le mode de représentation des mots et des morphèmes se fait phrase par phrase sous forme interlinéaire (cf. Figure 2). La glose des mots et des morphèmes est affichée dans un menu déroulant qui permet de changer sa valeur en la choisissant dans l'ensemble de toutes les autres gloses du même document ou des autres documents ouverts.

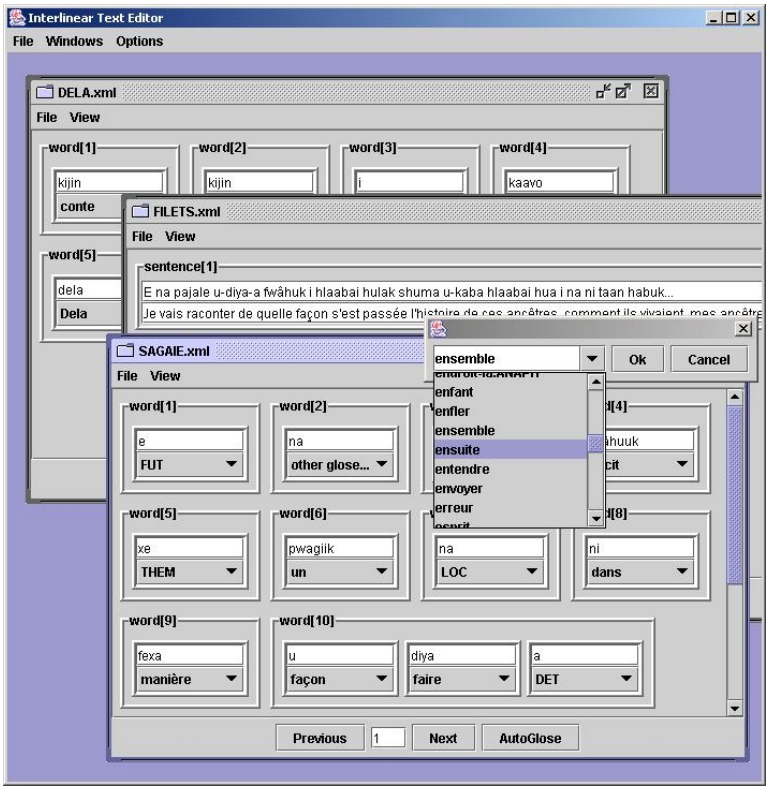

Figure 2. Logiciel « Interlinear Text Editor »

Le logiciel entretient une relation permanente et étroite entre les documents ouverts et un lexique construit de toutes les gloses utilisées dans ces derniers. Chaque transcription de mot ou de morphème est liée à une liste de gloses déjà utilisées pour les gloser. Cette liste comporte pour l'instant uniquement le contenu textuel des gloses et le nombre de fois où elles ont été utilisées. Tout ajout ou changement d'une glose dans un document entraîne un change-

[4] Le logiciel 'Interlinear Text Editor' écrit par l'auteur (Michel Jacobson) est librement disponible sur le site web (http://michel.jacobson.free.fr). Il s'agit d'un programme écrit en Java utilisable sur n'importe quelle plate-forme disposant d'une machine virtuelle JAVA1.3.

ment dans ce lexique pour ajouter ou supprimer un lien ou bien augmenter ou diminuer le nombre d'occurrences de ce lien.

Il est possible aussi d'importer en bloc tout un lexique comportant l'ensemble des liens transcription/glose. Ce lexique sera simplement ajouté à celui calculé à partir des fichiers ouverts.

3.2. *Utilisation du logiciel*

À titre d'exemple, nous avons testé ce logiciel sur un corpus en langue nêlêmwa de Nouvelle-Calédonie. Nous disposons pour cette langue de quelques ressources :

- Un corpus de cinq textes narratifs (Lacito, 2003), représentant environ 1 heure 40 minutes d'enregistrement, accompagné d'une analyse explicitée en XML en phrases, en mots et en morphèmes, avec pour les phrases une transcription phonétique, une traduction en français ou en anglais, et des gloses en français pour chacun des morphèmes ;
- Un corpus d'une trentaine de textes formatés avec Shoebox, représentant 5 heures d'enregistrement. Chaque texte est analysé en phrases comportant une transcription phonétique et une traduction libre en français. Les mots et morphèmes sont indiqués dans la transcription par des marqueurs typographiques. Les morphèmes ne sont pas glosés ;
- Un dictionnaire formaté avec Shoebox comportant près de 3500 entrées (Bril, 2000);

Ceci est un exemple assez représentatif des ressources informatiques dont nous pouvons disposer dans notre laboratoire (et plus généralement en linguistique de terrain) pour une langue donnée.

Dans un premier temps nous avons construit un lexique à partir de 4 textes déjà glosés (environ 700 items), afin de disposer d'une source unique et synthétique d'informations.

Puis nous avons utilisé le logiciel pour gloser un cinquième texte. Nous avons compté :

1) le nombre de fois où le logiciel ne trouve qu'une manière de gloser un morphème. Ce dernier peut donc être glosé sans intervention de l'utilisateur. Nous avons distingué : 1a) les cas où le morphème est correctement glosé ; 2b) des cas où il ne l'est pas ;
2) le nombre de fois où le logiciel propose à l'utilisateur de choisir la glose parmi une liste de gloses. Nous avons distingué : 2a) les cas où la glose correcte faisait partie de la liste en première position, 2b) les cas où elle se trouvait dans les positions suivantes et 2c) les cas où elle n'en faisait pas partie.
3) le nombre d'échecs, c'est-à-dire qu'aucune glose n'est proposée par le logiciel ;

L'hypothèse que nous formulons est que le logiciel sera d'autant plus utile que le nombre de gloses correctement trouvées sera important et le nombre de gloses erronées faible. D'autre part le gain de temps sera d'autant plus grand que la glose correcte proposée en cas d'homophonie sera placée souvent en première position dans la liste. Enfin le logiciel devrait être capable de s'améliorer sur ces deux points par apprentissage. Cet apprentissage consiste à enrichir le lexique au fur et à mesure que le logiciel glose les morphèmes et à corriger à la volée ces choix si ceux-ci sont erronés. Pour simuler cet apprentissage nous avons utilisé le logiciel pour gloser un texte préalablement glosé manuellement. Le logiciel peut alors simuler les corrections d'un utilisateur en contrôlant et rectifiant automatiquement ses choix par comparaison avec les gloses préexistantes.

3.3. *Analyse des résultats*

	Correctement glosé (1a)	Incorrectement glosé (1b)	Total
Sans apprentissage	130	33	163
Avec apprentissage	158	19	177

Tableau 1.
Nombre de morphèmes correctement et incorrectement glosés en fonction de l'apprentissage.

Ce premier tableau (Tableau 1) ne comptabilise que les transcriptions de morphèmes présentes dans le lexique et ne présentant pas d'homophonie. Une seule glose est donc proposée par le logiciel. Elle est jugée correcte ou non en fonction de la glose effectivement présente dans le document d'origine. La différence dans le total des deux situations (avec et sans apprentissage) vient du fait que certains cas d'homophonie ont pu être détecté par l'apprentissage. On remarquera dans le tableau que globalement on obtient plus de bonnes propositions que de mauvaises et que ce taux s'améliore avec l'apprentissage (on passe de 81,6% à 89,2% de gloses correctes).

En examinant de manière qualitative les résultats on peut voir qu'un certain nombre de mauvaises propositions est du à l'utilisation de synonymes pour les gloses. Par exemple avec le corpus testé, la transcription [hulak] laquelle est glosée « ancien » dans le dictionnaire a été glosée « vieux » dans le corpus, alors que la transcription [hulaxa] laquelle, inversement, a été glosée « vieux » dans le dictionnaire à été cette fois glosée « ancien » dans le corpus. Un autre morphème précédemment glosé « semblable » a été cette fois glosé par « être comme ». [kûûlâ] qui avait été jusqu'alors glosé soit par « accomplir » soit par « finir », l'est cette fois dans le corpus par « terminer ». Nous pourrions citer de nombreux autres exemples.

	Position dans la liste			
	Première (2a)	Suivantes (2b)	Absente (2c)	Total
Sans apprentissage	162	194	73	429
Avec apprentissage	164	269	28	461

Tableau 2. Nombre de fois où la glose correcte apparaît dans la liste de propositions, en fonction de sa position dans la liste et de l'apprentissage.

Le tableau 2 comptabilise les transcriptions de morphèmes présentes dans le lexique et qui présentent de l'homophonie. Le logiciel calcule alors la liste de toutes les gloses candidates qu'il trie par fréquence d'apparition. On observe que le nombre de cas d'homophonie non reconnue diminue assez fortement (col. 2c : 28 cas contre 73 cas sans apprentissage), ce qui est logique pour un système avec apprentissage. L'apprentissage semble donc profiter en grande partie aux homophones. Mais le critère de fréquence d'utilisation utilisé ici ne permet pas d'améliorer le taux de bon classement de ces derniers (col. 2a et 2b). Ces résultats semblent nous encourager dans la voie de l'utilisation d'un autre critère tel que celui de la co-occurrence mise en valeur dans les concordances. Ce dernier critère devrait améliorer le classement de gloses des homophones en faisant apparaître en premier les plus probables en termes d'environnement contextuel et non plus en termes de simple fréquence d'utilisation.

	Aucune glose proposée (3)
Sans apprentissage	96
Avec apprentissage	50

Tableau 3. Nombre d'échecs en fonction de l'apprentissage

Le dernier tableau montre simplement que l'apprentissage a permis d'ajouter une cinquantaine d'entrées au lexique. L'examen du tableau 2 nous a déjà montré que ces nouvelles entrées ajoutent principalement de nouveaux cas d'homophonie.

4. Perspectives

Segmenter et gloser la transcription d'un texte oral sont deux tâches à la fois répétitives, fastidieuses et délicates. Un outil logiciel ayant pour objectif d'apporter une aide à ces deux tâches doit utiliser au maximum le peu de renseignements qu'il détient sur la langue pour que ses propositions d'analyses soient les plus fiables et pertinentes possibles. Etant donné le faible niveau de renseignements que l'on possède en général sur ces langues, le système d'aide doit pouvoir s'enrichir au fur et à mesure de l'analyse. C'est ce que nous avons tenté de faire avec le logiciel présenté plus haut. Il reste tout de même de nombreuses fonctions à ajouter à celui-ci pour qu'il puisse constituer une véritable aide au linguiste. De nombreuses interfaces restent aussi à définir pour qu'il soit plus convivial dans son utilisation et plus paramétrable afin de l'adapter à des besoins spécifiques divers.

Il convient aussi d'optimiser certains algorithmes pour que l'aide apportée soit plus pertinente. Le choix et l'ordre de présentation des gloses proposées pourrait venir par exemple de l'utilisation des concordances, ou bien de critères supplémentaire à la transcription, comme par exemple de connaissances syntaxiques (partie du discours, etc.).

Références

Buseman A et Buseman K. (1998). *The Linguists SHOEBOX*. Summer Institute of Linguistics. (http://www.sil.org/computing/catalog/shoebox.html).

Bray T., Paoli J. et Sperberg-McQueen C.M. (Eds) (1998). *Extensible Markup Language (XML) 1.0.* W3C Recommendation, 10 February 1998 (http://www.w3.org/TR/REC-xml).

Bril I. (2000). *Dictionnaire nêlêmwa-nixumwak-français-anglais*. LCP 14. Peeters.

Hsu R. (1989). *Lexware Manual*. Second Edition. Linguistics Dept., University of Hawaii.

Kaplan M.R. et Martin K. (1994). *Regular Models of Phonological Rule Systems*. Computational Linguistics, vol. (20/3) : 331-378.

Lacito. (2003). Site web du programme "Archivage" (http://lacito.vjf.cnrs.fr/archives).

Rabiner R.L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, vol (77/2) : 257-286.

Sproat R. *et al.* (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. In *Proceedings of the Meetng of the Association for Computational Linguistics* : 377-404.