Les langues africaines sur la Toile
Étude des cas haoussa, somali, lingala et isixhosa

Les langues africaines sont présentes sur la Toile. Cette présence facilite la création de corpus linguistiques. Pour illustrer notre propos et notre méthodologie, nous nous concentrons sur quelques langues des quatre coins de l'Afrique: le haoussa, le somali, le lingala et l'isixhosa. La taille et la composition des corpus respectifs donnent une idée de ce qui est disponible sur la Toile pour ces langues et permettent aussi de comparer ces langues entre elles. Afin d'illustrer le potentiel de ces corpus linguistiques nous fabriquerons, à titre d'exemple, des logiciels de grande utilité: à savoir, des correcteurs d'orthographe pour chacune de ces langues africaines.

Termes-clés:
la Toile; correcteurs d'orthographe; langues africaines; haoussa; somali; lingala; isixhosa.

1 Introduction

On entend souvent dire que les langues africaines ne sont pas présentes sur la Toile. Cette vision doit être sérieusement révisée. Plusieurs centaines de langues africaines ont déjà réussi à conquérir une place dans le cyberespace. Le nombre de sites et donc de textes par contre, dépend de la langue en question et n'est pas toujours en correspondance avec le nombre de locuteurs ni avec l'importance supposée de la langue.

Chanard et Popescu-Belis (2001) ont déjà détaillé les trois phases nécessaires pour qu'on puisse parler de l'informatisation d'une langue. Dans leur article ils ont traité la toute première étape, à savoir le stockage des documents «*sous forme alphabétique, syllabique ou idéographique* sur un support informatique, c'est-à-dire sous la forme d'une suite de caractères qui seule permet l'édition, la recherche, bref la manipulation *analytique* du contenu linguistique» (Chanard et Popescu-Belis 2001: 33).
La deuxième et la troisième étape sont d'après eux respectivement les outils informatiques adaptés à la langue et la diffusion de la langue sur la Toile.

Dans la présente contribution par contre, nous renversons leur séquence en choisissant la Toile comme point de départ. En effet, nous nous basons sur des documents disponibles dans le cyberespace, pour créer des corpus linguistiques, et ensuite des logiciels comme par exemple des correcteurs d'orthographe. Il va de soi que l'existence de documents en langues africaines sur la Toile présuppose le stockage de ces textes sur un support informatique – la première phase de Chanard et Popescu-Belis.

2 La Toile et les langues africaines

Quoique l'anglais reste de loin la langue dominante sur la Toile, Grefenstette (2002) montre que beaucoup d'autres langues de par le monde gagnent du terrain. Ceci est aussi le cas pour les langues africaines, puisque la proportion de celles-ci ne cesse de croître. Pour une esquisse traitant des langues africaines dans le cyberespace dans ce contexte, voir De Schryver (2002). Néanmoins, une question valable reste la suivante: pourquoi est-ce que la présence des langues africaines sur la Toile reste aussi précaire bien que celles-ci constituent 30% de toutes les langues du monde?

2.1 La situation précaire des langues africaines dans le cyberespace

Est-ce un problème d'existence de caractères? Non, de nos jours toute écriture peut être utilisée sur la Toile, aussi bien les systèmes alphabétiques (comme par exemple le kiswahili[1] ou l'arabe[2]), que les systèmes syllabiques (comme par exemple l'amharique[3]), que les systèmes idéographiques (comme pour par exemple le *han* chinois ou le *kanji* japonais). En ce qui concerne les langues africaines, il est bien connu que la plus grande partie (de celles qui sont déjà écrites) est basée sur l'alphabet latin. De la même façon que pour plusieurs langues indo-européennes qui emploient l'alphabet latin comme base, des signes diacritiques peuvent avoir été ajoutés ou des caractères latins de base peuvent avoir été adaptés en ajoutant des glyphes (pseudo-)phonétiques. Pour pouvoir «voir» l'orthographe correcte, ou du moins l'orthographe du

1 *p. ex. ippmedia.com/alasiri.htm*
2 *p. ex. news.bbc.co.uk/hi/arabic/news*
3 *p. ex. www2.dw-world.de/amharic/*

document, il suffit d'employer un navigateur plutôt récent (par exemple *Explorer 4+* ou *Netscape 4+*), et/ou de télécharger le jeu de caractères qui a été employé pour la création du document.

Est-ce un problème de standard orthographique? Pour le haoussa par exemple, il existe une orthographe officielle. Par contre sur la Toile, comme on le verra plus loin, on trouve au moins six types d'orthographe. Cette situation ne facilite pas nécessairement l'échange de documents et peut même être un obstacle dans la création et le téléchargement depuis les serveurs de documents.

Est-ce un problème d'accès? Les difficultés d'accès aux technologies de l'information et de la communication (TIC) ne sont en effet pas à sous-estimer en Afrique. Les cybercafés ne se trouvent que dans les grandes villes et ils ne sont pas accessibles à tout le monde. En plus, l'analphabétisme constitue un obstacle à la fréquentation de textes en l'occurrence sur ordinateur.

Est-ce un problème de motivation? Le manque de motivation parmi les Africains à écrire dans leur propre langue est une des raisons que l'on peut citer pour expliquer le relatif insuccès des langues africaines sur la Toile. Le cybernaute qui s'exprime sur la Toile veut être lu et compris, il va donc écrire dans une langue connue par le plus grand nombre de gens. En effet, une grande partie des textes en langues africaines trouvés sur la Toile n'a pas été écrit par des Africains, comme nombre de documents religieux ou de textes destinés à l'enseignement. Des forums où des Africains communiquent avec d'autres Africains, en langues africaines, sont l'exception et non la règle.

2.2 Estimer le nombre de mots en langues africaines dans le cyberespace

Grefenstette et Nioche (2000) ont proposé une méthode pour estimer le nombre de mots, pour une langue donnée, dans le cyberespace. Même si le nombre de mots en langues africaines est, jusqu'à ce jour, en fait toujours trop faible pour ce genre de calcul, on peut déjà employer la méthode de Grefenstette et Nioche avec succès. L'idée de base est toute simple: il suffit de diviser le nombre de fois qu'une sélection de mots uniques d'une langue donnée a été indexée par un moteur de recherche, par la fréquence relative de ces mêmes mots dans un corpus. La moyenne des résultats offre alors une estimation du nombre de mots sur la Toile dans la langue étudiée.

On peut brièvement illustrer la méthode pour le sepedi. La colonne 1 dans [1] liste les dix mots qui ont été utilisés, avec leurs équivalents français dans la colonne 2. La colonne 3 montre les fréquences relatives dans un corpus du sepedi (pour ce corpus, voir De Schryver & Lepota 2001: 3). Les dix mots de la colonne 1 ont été soumis au moteur de recherche *Google*[4] (mars 2003) et le nombre de «pages vues» pour chacun de ces mots est montré dans l'avant-dernière colonne. Les estimations de la dernière colonne sont basées sur les chiffres des deux colonnes précédentes.

[1] Estimation du nombre de mots sepedi sur la Toile

Mot en sepedi	Équivalent en français	Fréquence dans un corpus (en %)	Résultat *Google*	# mots sepedi sur la Toile
kgopela	demander	0,03961934	46	116 105
latelago	suivre (+relatif)	0,03302768	57	172 582
kgauswi	prêt (de)	0,02695642	31	115 000
bolelago	parler (+relatif)	0,02260245	21	92 910
mafelelong	à la fin	0,01974029	26	131 710
tsebago	savoir (+relatif)	0,01925458	18	93 484
mangwalo	lettres	0,01075481	20	185 963
hlogong	dans la tête	0,01030380	8	77 641
sengwalwa	manuscrit	0,00995687	9	90 390
phapano	différence	0,00988749	19	192 162
				126 795

En considérant la moyenne, l'on peut conclure qu'il y a au moins cent vingt-cinq mille mots sepedi sur la Toile. Nous remarquons aussi que même si l'on n'a pas accès à un corpus d'une certaine langue (pour en extraire les chiffres de la colonne 3), il est évidemment possible de télécharger d'abord quelques textes (dans la langue à examiner) de la Toile. Les fréquences relatives peuvent alors être dérivées de l'ensemble de ces textes.

3 Quatre études de cas : haoussa, somali, lingala et isixhosa

Pour illustrer l'existence des langues africaines sur la Toile, nous avons choisi quatre langues représentant quatre régions géographiques de l'Afrique sub-saharienne : le haoussa pour l'Ouest, le somali pour l'Est, le lingala pour la région centrale et l'isixhosa pour la région australe. Notons en passant que le Nord de l'Afrique est évidemment bien représenté par l'arabe, et que la langue sub-saharienne dominante sur la Toile est le kiswahili. Ces deux dernières langues sont d'ailleurs tellement présentes, avec plusieurs millions de mots chacune, qu'il est possible de rechercher uniquement des pages écrites dans ces langues avec le moteur de recherche *All the Web*[5].

Rechercher des sites et/ou des pages dans une langue africaine donnée peut facilement se faire à l'aide d'un des nombreux moteurs de recherche. Il suffit d'introduire un ou plusieurs mots de la langue recherchée dans le champ de recherche, d'appuyer sur la touche *Entrée* du clavier, et – à condition que les mots n'existent que dans cette langue – tous les liens cliquables sur la page avec les résultats mèneront à des sites et/ou des pages dans la langue recherchée.

C'est exactement ce procédé-là que nous avons suivi pour le haoussa, le somali, le lingala et l'isixhosa. Pour chacune de ces langues, nous avons navigué sur la Toile pendant environ quatre jours (mars 2003). À partir des textes trouvés, nous avons alors construit un corpus pour chaque langue. La taille et le contenu de ces différents corpus dépendent bien sûr de ce qui est disponible sur la Toile. Ensuite nous avons utilisé ces corpus pour créer plusieurs correcteurs d'orthographe dans les quatre langues.

Dans ce qui suit, nous présenterons les difficultés que nous avons rencontrées, les résultats, puis nous expliquerons comment on peut fabriquer un correcteur d'orthographe. Nous mettrons aussi les correcteurs d'orthographe à l'épreuve à l'aide d'un texte qui existe dans les quatre langues concernées, c'est-à-dire la *Déclaration universelle des droits de l'homme*[6].

4 *www.google.com*
5 *www.alltheweb.com*
6 *www.unhchr.ch/udhr*
7 *www.ethnologue.com*

3.1 L'Afrique de l'Ouest : haoussa

Pour l'Afrique de l'Ouest, le choix s'est facilement porté sur le haoussa qui est, avec ses trente-neuf millions de locuteurs dans huit pays différents[7], la plus grande *lingua franca* de cette région géographique. Le haoussa est une langue de la famille chadique qui fait partie de la grande unité des langues afro-asiatiques. Ce grand nombre de locuteurs du haoussa comme première ou deuxième langue et sa fonction comme langue véhiculaire ont malheureusement pour conséquence que les textes haoussa sur la Toile se présentent sous des formes très diverses. Nous avons aussi remarqué que la plupart des sites sont d'origine nigériane.

Le grand nombre d'orthographes différentes, utilisées dans des textes haoussa trouvés sur la Toile, pose problème. Historiquement, le haoussa connaît deux types d'écritures, l'ajami et le boko. L'écriture ajami, étant basée sur des caractères arabes, était utilisée dès le XVII^{e} siècle. À présent par contre, l'écriture basée sur l'alphabet latin, le boko, a gagné beaucoup de terrain. Cette écriture boko, utilisée par les Européens depuis le début du XIX^{e} siècle, est la seule trouvée sur la Toile. Dans l'étude présente nous ne nous arrêterons pas sur la question de l'écriture ajami.

Malgré l'imposition d'une orthographe standard pour le haoussa en janvier 1980 pendant une conférence à Niamey (Wolff 1991), au moins six types d'orthographe sont utilisés dans le cyberespace. L'alphabet latin « pur » ne répond pas aux besoins d'une écriture haoussa, puisque des caractères symbolisant des coups de glottes, des implosifs et des éjectifs sont absents. Dans l'orthographe de 1980, le « b implosif » est représenté par le caractère phonétique ɓ et le « d implosif » par le caractère phonétique ɗ. Le « k éjectif » correspond au caractère phonétique ƙ. Le « coup de glotte » est représenté par l'apostrophe ' et la combinaison du « coup de glotte et y » s'écrit y'. On peut considérer les autres types d'orthographe (trouvés sur la Toile et ailleurs) comme des adaptations graduelles de cette orthographe standard. Le y' peut être remplacé par 'y. Une deuxième simplification est le remplacement de ɓ et ɗ par 'b et 'd et le remplacement de ƙ par k'. Ici on voit encore bien la différence entre les consonnes implosives et la consonne éjective. Dans une troisième étape, par contre, cette distinction est omise et on trouve b', d' et k'. Une simplification qui va encore plus loin est l'omission du caractère implosif ou éjectif des

consonnes, on garde seulement l'apostrophe pour le coup de glotte et avant y. Dans la dernière étape, seul le coup de glotte est présent. Ces six types d'écriture haoussa sont non seulement tous présents sur la Toile, mais parfois plusieurs orthographes sont utilisées sur un seul site ou même dans un seul texte. Bien que la simplification ne semble pas poser trop de problèmes pour les locuteurs maternels du haoussa – dans les différents forums, les cybernautes ont tendance à utiliser l'orthographe la plus simplifiée – les orthographes simplifiées ne contribuent pas à la compréhension de la langue pour ceux qui étudient le haoussa.

Bien que le haoussa soit une langue tonale, les tons ne sont pas marqués dans la plupart des textes haoussa disponibles sur la Toile. Nous avons trouvé seulement cinquante-six mots sur lesquels des tons sont marqués; il s'agit des tons hauts (´), des tons bas (`) et des tons descendants (^). Il n'y a que neuf textes sur trois sites différents qui utilisent ces mots avec tons.

Surtout pour le traitement informatique efficace d'une langue, l'utilisation d'un seul type d'orthographe est nécessaire. Le meilleur choix pour le haoussa est sans aucun doute l'orthographe standard. L'utilisation des caractères phonétiques ne pose pas de problèmes, puisque ceux-ci sont présents dans le jeu Unicode. Même pour l'élaboration d'un correcteur d'orthographe, ces caractères phonétiques (Unicode) peuvent être utilisés sans aucun problème.

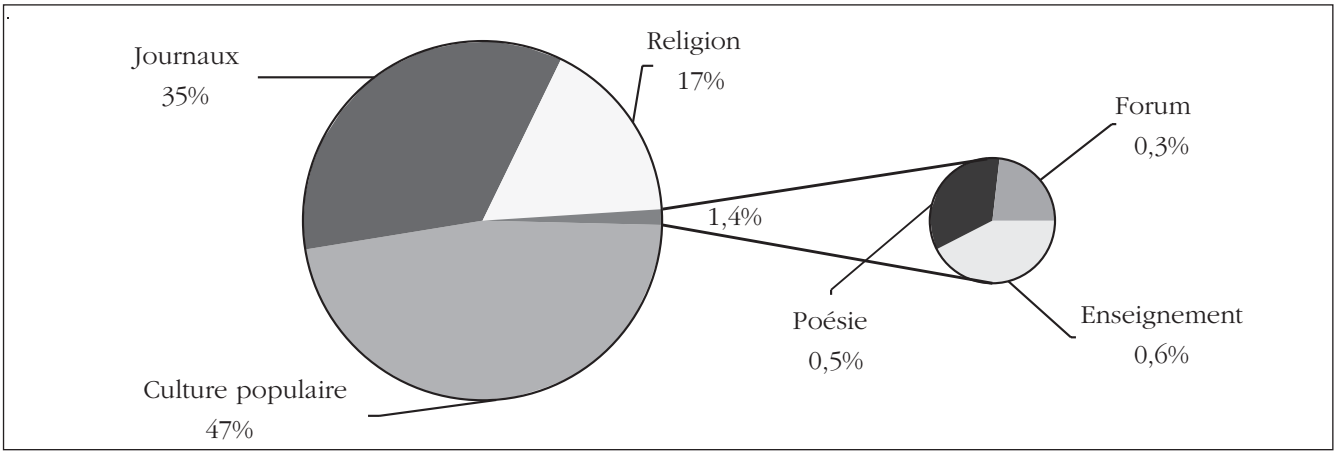
3.1.1 Distribution du haoussa sur la Toile selon le contenu

De la navigation sur la Toile pendant quatre jours à la recherche de documents haoussa résulte un corpus linguistique contenant 858 734 mots au total (les « *tokens* »), dont 30 996 mots différents (les « *types* »). Comme on peut le voir dans [2], les textes haoussa peuvent être divisés en trois grandes catégories: culture populaire (47 %), journaux (35 %) et religion (17 %). Le pourcentage de textes produits sur des forums par des cybernautes est minuscule (0,3 % du total des mots).

3.1.2 Correcteurs d'orthographe haoussa

Il y a plusieurs façons de construire un correcteur d'orthographe. Pour un aperçu avec l'accent sur les langues africaines, voir Prinsloo et De Schryver (2003). Une des pistes est tellement simple qu'on pourrait parler de « correcteur d'orthographe maison ». En effet, tous les logiciels populaires de traitement de textes, comme *Microsoft Word* ou *Corel WordPerfect*, ont une fonction qui

[2] Distribution du haoussa sur la Toile selon le contenu



permet d'ajouter un ou plusieurs « dictionnaires personnels ». Il suffit donc de composer une liste des mots fréquents dans un corpus, ou même une liste de tous les mots différents dans un corpus, et d'en faire un ou plusieurs correcteurs d'orthographe.

À titre d'exemple, [3] montre une partie d'un texte du site *Bisharat*[8] : ce texte ne se trouve *pas* dans le corpus, et il est vérifié avec un correcteur d'orthographe dérivé du corpus contenant 858 734 *tokens* et 30 996 *types*. On voit clairement qu'il n'y a aucun problème à traiter l'orthographe standard du haoussa (avec les caractères phonétiques) dans un outil d'édition, même au niveau du logiciel de vérification.

Dès qu'on dispose d'un correcteur d'orthographe utilisant une certaine orthographe, il est généralement trivial d'employer la fonction « rechercher et remplacer » pour en faire un vérificateur capable de reconnaître une autre orthographe. Puisque nous avons décidé de contrôler l'efficacité de tous nos correcteurs d'orthographe à l'aide du même texte disponible dans les quatre langues – la *Déclaration universelle des droits de l'homme*, un texte qui n'est évidemment pas inclus dans les corpus respectifs – nous avons produit un correcteur d'orthographe haoussa qui utilise l'écriture de ce texte, et dans le cas présent, l'écriture la plus simplifiée. Les résultats des tests sont résumés dans [4].

Comme on peut le voir, un total de cinq correcteurs d'orthographe a été produit, chacun correspondant à une

[3] Texte haoussa dans l'orthographe standard, vérifié à l'aide d'un correcteur d'orthographe dérivé de la Toile (2 mots corrects ne sont pas reconnus)

Ga albarkacin ilimin na'urar zamani, halayen al'amuran duniya na cudayyar jama'a, sai ci gaba da tasowa suke yi da sauri. Amma duk da haka, a fadin Afirka da sauran wurare na kudu a bisa doron duniya, akwai ƙarancin albarkatun na'urar zamani wadanda za'a fuskanci hanyoyin labarai da sadarwa na zamani da su (kuma ga buƙatu iri-iri na sahun gaba sun fuskanto a gaggauce).

Ga halin haka kuwa, hanyoyin dacewa da dama sukan iya ɓacewa in ba'a ɗauki wasu matakai na aikata abubuwan da wuri ba. Bayan dogon nazari bisa ga dukan al 'amuran nan ne wasu ma'aikata da kungiyoyi masu zaman kansu, suke aiki domin su taimaki Afirka ta fannin sababbin hanyoyin labarai da sadarwa na zamani, ta a maido hankali wajen sawa a gane da yadda hanyoyin nan suke, da kuma yadda ake iya cin moriyarsu a aikace.

Les changements technologiques et socio-économiques se déroulent rapidement. Cependant, en Afrique, aussi bien qu'ailleurs dans le Sud, il y a peu de ressources disponibles pour traiter des aspects de TIC (et en même temps, il y a beaucoup d'autres besoins de base exigeant ces ressources).

Des opportunités peuvent être perdues dans la mesure où des actions opportunes ne sont pas prises. En reconnaissant ce fait, des agences et des organismes extérieurs cherchent à aider l'Afrique dans le domaine des TIC, se concentrant d'abord sur les questions essentielles de la connectivité de base et des aspects de « l'accès physique ». Cependant, peu abordent encore les questions également importantes de « l'accès significatif » et du contenu.

8 *www.bisharat.net*

[4] Construction d'un correcteur d'orthographe haoussa

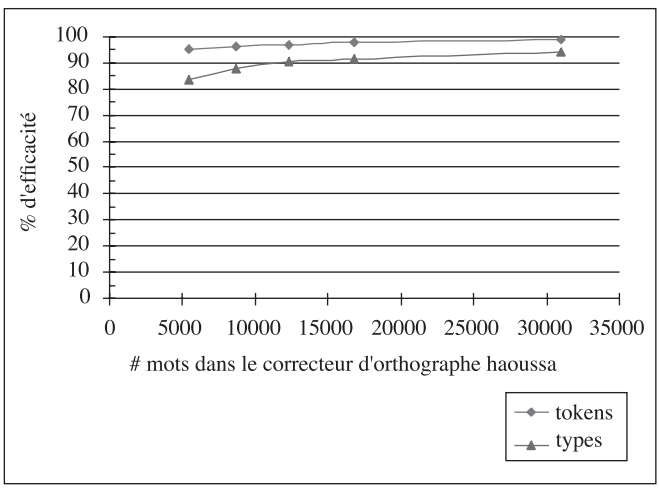
Correcteur d'orthographe haoussa (dérivé de 858 734 *tokens*)			*Déclaration universelle des droits de l'homme* haoussa (2 751 *tokens*; 556 *types*)				
mots dans chaque niveau			pas reconnu		% d'efficacité		
fréquence	nombre	%	*tokens*	*types*	*tokens*	*types*	utilisateurs
10 ou plus	5 418	17,48	130	90	95,27	83,81	96,73
5 jusqu'à 9	3 218	10,38	100	67	96,36	87,95	97,56
3 et 4	3 657	11,80	81	53	97,06	90,47	98,07
2	4 519	14,58	64	46	97,67	91,73	98,33
1 (hapax)	14 184	45,76	35	31	98,73	94,42	98,87
	30 996	100,00					

différente section ou « couche » du corpus : (i) tous les mots avec une fréquence d'au moins 10 ; (ii) tous les mots avec une fréquence de cinq à neuf ; (iii) tous les mots avec une fréquence de trois et quatre ; (iv) tous les mots avec une fréquence de deux ; et (v) tous les mots qui n'apparaissent qu'une fois dans le corpus (les « hapax », bons pour 45,76 % du total). La *Déclaration universelle des droits de l'homme* contient 2 751 *tokens* et 556 *types*. Nous présumons qu'il n'y a pas de fautes dans ce texte, et donc qu'un vérificateur « parfait » devrait pouvoir reconnaître tous les mots. Autrement dit, l'efficacité (ou le « taux de rappel ») devrait être de 100 %. En utilisant seulement le premier niveau, l'efficacité du point de vue des *tokens* est déjà de 95,27 %. En ajoutant le second niveau au premier, l'efficacité grimpe jusqu'à 96,36 % ; en ajoutant le troisième niveau, l'efficacité est de 97,06 % ; en ajoutant ensuite le quatrième niveau, l'on atteint 97,67 % ; finalement avec tous les niveaux ensembles, l'efficacité grimpe jusqu'à 98,73 %. Du point de vue des *types*, l'efficacité est moins bonne, mais du point de vue des utilisateurs, l'efficacité grimpe même jusqu'à 98,87 %. Pour calculer ce dernier, le nombre de *types* non reconnu est comparé au nombre de *tokens* dans le texte. Le raisonnement est que, du point de vue de l'utilisateur, il suffit d'ajouter chaque *type* une seule fois avec la fonction « Ajouter » afin que le correcteur d'orthographe reconnaisse le mot en question. Les graphiques dans [5] résument l'efficacité.

3.2 L'Afrique de l'Est : somali

Le somali – aussi une langue afro-asiatique, mais faisant partie de la famille cushitique – est parlé en Afrique de l'Est, plus précisément au Somali et dans les pays limitrophes : le Djibouti, l'Éthiopie et le Kenya. La population totale de locuteurs du somali peut être chiffrée à

[5] Efficacité d'un correcteur d'orthographe haoussa

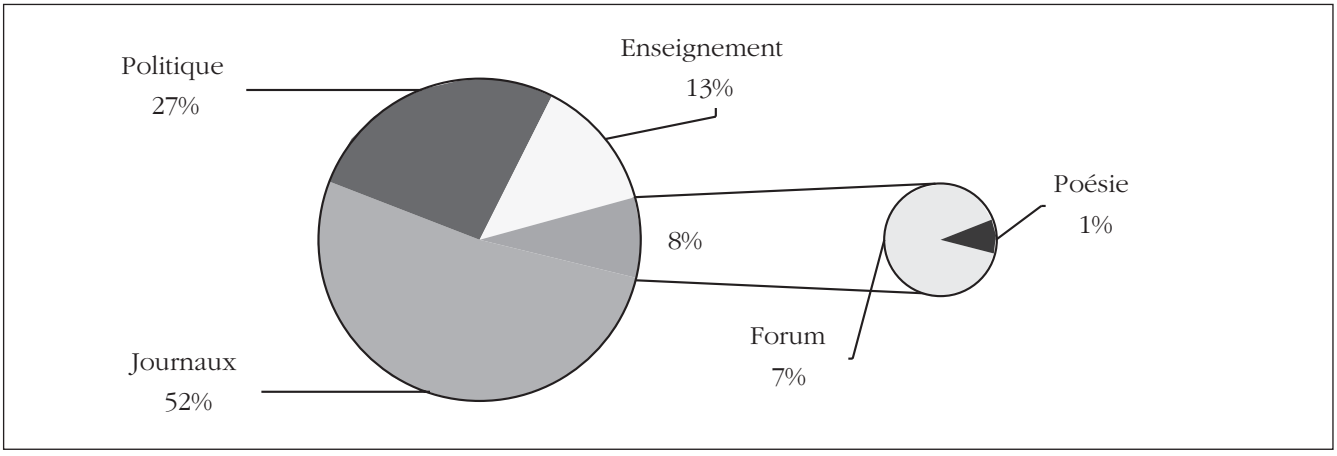


environ dix millions[9]. Ce nombre de locuteurs indique que cette langue joue un rôle important parmi les langues africaines. L'écriture du somali se fait à l'aide de l'alphabet latin.

3.2.1 Distribution du somali sur la Toile selon le contenu

Un corpus de 304 361 *tokens* et 40 251 *types* a été rassemblé, dont la distribution est montrée dans [6].

[6] Distribution du somali sur la Toile selon le contenu



Comparé à la distribution pour le haoussa, on voit l'apparition d'un grand bloc de politique (27 %) et d'enseignement (13 %). Aussi les forums grandissent substantiellement (7 %). Notons toutefois que malgré le fait qu'il n'y a pas de textes religieux dans le corpus, il y en a qui sont disponibles sur la Toile. En ce qui concerne les documents politiques, il est étonnant qu'un grand nombre provienne de sites officiels de gouvernements non-africains (australiens, finlandais, suédois, etc.). La présence de locuteurs du somali dans ces pays peut expliquer cela. Une autre partie des documents politiques est originaire de la « République de Somaliland ». Environ la moitié (52 %) du matériel somali sur la Toile est composé d'articles de journaux.

3.2.2 Correcteurs d'orthographe somali

Pour le somali aussi, cinq correcteurs d'orthographe ont été produits selon la même procédure que pour le haoussa. Malgré le fait que le corpus somali soit trois fois plus petit que celui du haoussa, il contient un quart de *types* de plus, dont 60,09 % d'hapax. Le nombre de *types* en soi ne dit évidemment pas tout, et l'efficacité des cinq niveaux du correcteur d'orthographe est en effet moins bonne que pour le haoussa. Les chiffres et graphiques sont résumés dans [7] et [8].

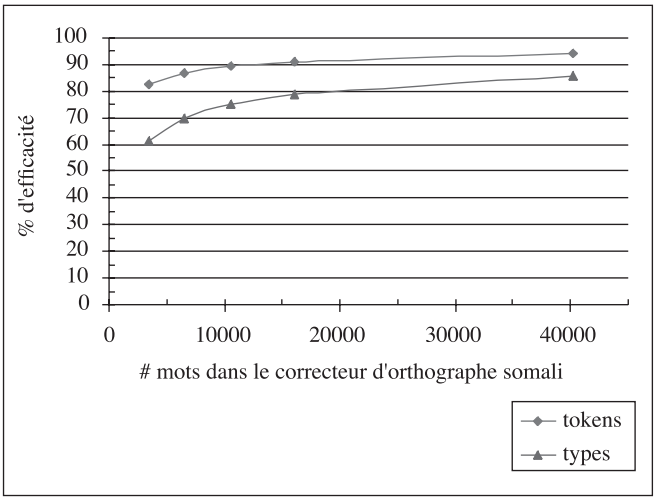
Pour le somali, la compagnie *SomiTek* distribue un correcteur d'orthographe appelé « Hikaadiye » qui peut être téléchargé depuis la Toile[10]. Après une comparaison entre notre correcteur d'orthographe et le correcteur offert par SomiTek, nous avons constaté que beaucoup plus de mots n'étaient pas reconnus par Hikaadiye. Il est possible que SomiTek se soit basée sur un autre dialecte du somali pour l'élaboration de son logiciel de vérification. Dans [9] les deux correcteurs d'orthographe sont appliqués sur le même texte, notamment l'Article 21 de la *Déclaration universelle des droits de l'homme*. Les mots qui ne sont pas reconnu sont en gras.

9 *www.ethnologue.com*
10 *www.somitek.com*

[7] Construction d'un correcteur d'orthographe somali

Correcteur d'orthographe somali (dérivé de 304 361 *tokens*)			*Déclaration universelle des droits de l'homme* somali (1 919 *tokens*; 710 *types*)				
mots dans chaque niveau			pas reconnu		% d'efficacité		
fréquence	nombre	%	*tokens*	*types*	*tokens*	*types*	utilisateurs
10 ou plus	3 433	8,53	334	276	82,60	61,13	85,62
5 jusqu'à 9	3 026	7,52	255	215	86,71	69,72	88,80
3 et 4	4 056	10,08	207	175	89,21	75,35	90,88
2	5 551	13,79	176	151	90,83	78,73	92,13
1 (hapax)	24 185	60,09	107	100	94,42	85,92	94,79
	40 251	100,00					

[8] Efficacité d'un correcteur d'orthographe somali



Pour ce petit extrait, le nombre de mots qui ne sont pas reconnus par le correcteur d'orthographe somali Hikaadiye est, comparé au nôtre, trois fois plus grand.

[9] Comparaison entre le correcteur d'orthographe somali Hikaadiye et le nôtre

Correcteur d'orthographe somali Hikaadiye
Qod **XXI**
1. Qof kastaa wuxuu xaq u leeyahay inuu ka qayb galo maamulka Dalkiisa si toos ah ama isagoo si xornimo ah u **dooranaya** cid wakiil ka ah.
2. Dadka oo dhami way u **simanyihiin** inay ka qayb galaan maamulka sare ee dalka ka jira.
3. Rabitaanka dadweynuhu waa saldhigga awoodda Dawladda; **rabitaankaas** waxaa lagu **muujinyaa dooroshooyin** run ah oo **xilliyo** joogta ah dhacaya, laga wada qayb-**galaya**, **waxayna** u **dhacayaan** si qarsoodi ah ama hab kale oo wax lagu doorto oo xor ah.

Notre correcteur d'orthographe pour le somali
Qod **XXI**
1. Qof kastaa wuxuu xaq u leeyahay inuu ka qayb galo maamulka Dalkiisa si toos ah ama isagoo si xornimo ah u dooranaya cid wakiil ka ah.
2. Dadka oo dhami way u **simanyihiin** inay ka qayb galaan maamulka sare ee dalka ka jira.
3. Rabitaanka dadweynuhu waa saldhigga awoodda Dawladda; **rabitaankaas** waxaa lagu muujinyaa **dooroshooyin** run ah oo xilliyo joogta ah dhacaya, laga wada qayb-galaya, waxayna u dhacayaan si qarsoodi ah ama hab kale oo wax lagu doorto oo xor ah.

3.3 L'Afrique centrale : lingala

Le lingala – une langue bantoue de la zone C de Guthrie (C36d), parlée dans la République démocratique du Congo (Kinshasa), dans la République du Congo (Brazzaville) et en Angola – peut être considérée comme une *lingua franca* de l'Afrique centrale. Le nombre de locuteurs tourne autour de sept millions[11]. Malgré ce nombre de locuteurs et le statut de la langue, il n'y a pas d'orthographe fixe. Seules les lettres de l'alphabet latin sont utilisées, du moins sur la Toile, mais l'indication des tons est plutôt sporadique et capricieuse. Les trois systèmes plus ou moins « systématiques » utilisés sur la Toile sont : (i) aucun ton est indiqué ; (ii) les tons sont seulement indiqués pour les « paires minimales » ; et (iii) tous les tons sont indiqués.

3.3.1 Distribution du lingala sur la Toile selon le contenu

La première chose que l'on remarque pour le lingala sur la Toile est que les sites ne sont pas nombreux. Des quatre langues, le corpus téléchargé pour le lingala est aussi le plus petit avec 193 772 *tokens* et 11 557 *types*. En plus, la distribution selon le contenu est très différente comme on peut le voir dans [10].

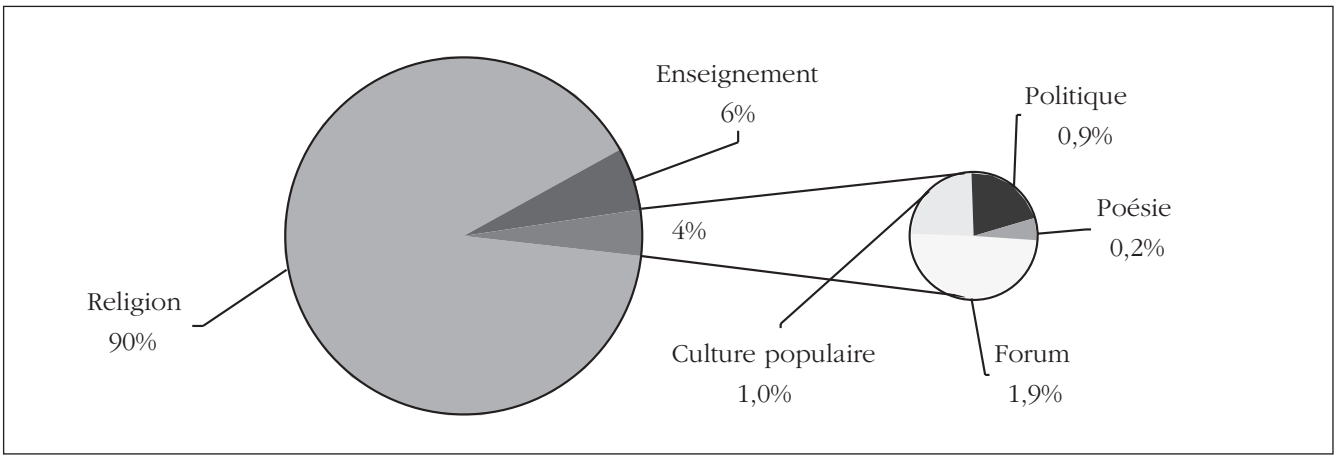
La très grande majorité des textes lingala se trouve dans la catégorie religion (90 %). La deuxième catégorie, l'enseignement (6 %), est beaucoup moins répandue. Le pourcentage des forums (1,9 %) est plus haut que pour le haoussa, mais n'atteint pas du tout le pourcentage du somali.

3.3.2 Correcteurs d'orthographe lingala

Puisque l'orthographe de la *Déclaration universelle des droits de l'homme* est sans tons, nous avons fabriqué un correcteur « lingala sans tons » à partir de tous les textes dans le corpus. Ceci est facile, puisqu'il suffit d'utiliser la fonction « rechercher et remplacer » pour se débarrasser des tons avant de faire la liste des mots pour le correcteur d'orthographe. Les résultats du correcteur « lingala sans tons » sont résumés dans [11] et [12].

Comparée aux correcteurs haoussa et somali, l'efficacité du correcteur lingala se trouve entre les deux, et cela avec beaucoup moins de mots. Notons tout de même qu'en plus de ce correcteur « lingala sans tons », nous avons aussi fabriqué un « lingala paires minimales » et un « lingala avec tons ».

[10] Distribution du lingala sur la Toile selon le contenu

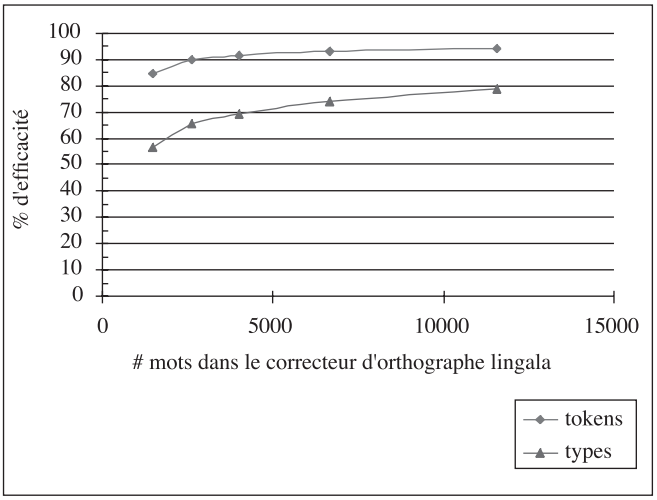


11 *www.ethnologue.com*

[11] Construction d'un correcteur d'orthographe lingala

Correcteur d'orthographe lingala (dérivé de 193 772 *tokens*)			*Déclaration universelle des droits de l'homme* lingala (1 854 *tokens*; 361 *types*)				
mots dans chaque niveau			pas reconnu		% d'efficacité		
fréquence	nombre	%	*tokens*	*types*	*tokens*	*types*	utilisateurs
10 ou plus	1 469	12,71	282	156	84,79	56,79	91,59
5 jusqu'à 9	1 130	9,78	182	125	90,18	65,37	93,26
3 et 4	1 388	12,01	154	110	91,69	69,53	94,07
2	2 668	23,09	132	94	92,88	73,96	94,93
1 (hapax)	4 902	42,42	104	77	94,39	78,67	95,85
	11 557	100,00					

[12] Efficacité d'un correcteur d'orthographe lingala



3.4 L'Afrique australe : isixhosa

Pour l'Afrique australe finalement, nous avons décidé d'étudier l'isixhosa, une langue de l'Afrique du Sud mais aussi parlée au Botswana et au Lesotho. Comme le lingala, l'isixhosa compte environ sept millions de locuteurs[12]. Cette langue bantoue de la zone S de Guthrie (S41) a été influencée par les langues khoisans. L'alphabet latin est utilisé pour l'écriture, et même les « clics » sont représentés par des caractères de l'alphabet latin. Les tons ne sont jamais indiqués.

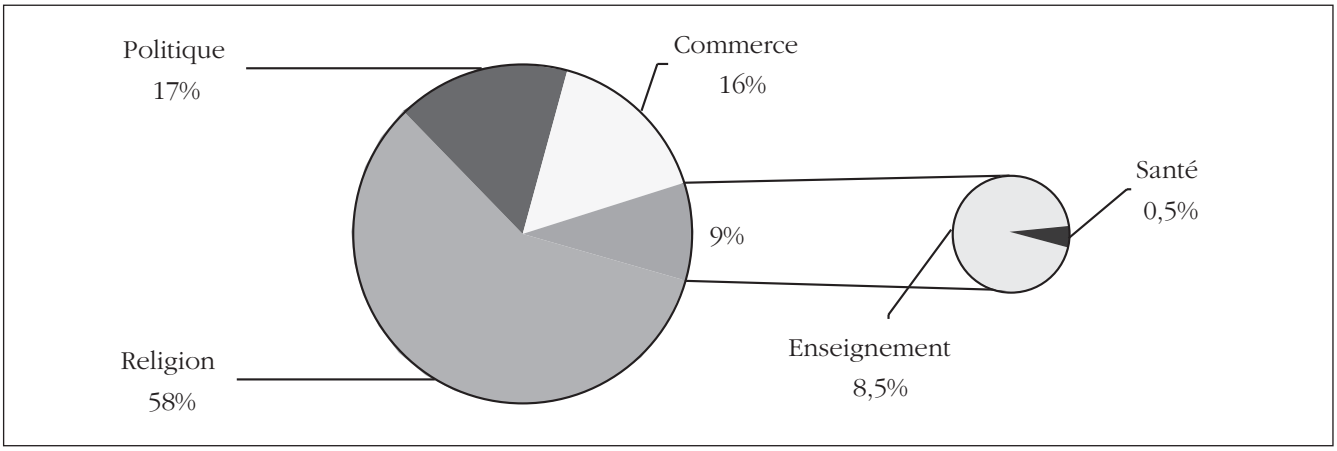
3.4.1 Distribution de l'isixhosa sur la Toile selon le contenu

Des neuf langues bantoues officielles de l'Afrique du Sud, l'isixhosa est certainement celle pour laquelle il y a « beaucoup » de matériel sur la Toile. Un corpus de 943 772 *tokens* et 149 553 *types* a été rassemblé, dont la distribution est montrée dans [13].

Les textes isixhosa peuvent être divisés en trois grandes catégories. Comme pour le lingala, la religion forme le plus grand bloc (58 %). La deuxième catégorie est constituée, comme pour le somali, de documents politiques (17 %). La troisième catégorie, le commerce (16 %), n'existe pas pour les autres langues de notre échantillon. Puisqu'il nous a semblé qu'il n'y a pas de forums isixhosa sérieux ou substantiels sur la Toile, cette catégorie n'est pas représentée dans notre corpus. Un bloc enseignement (8,5 %) par contre, est présent comme pour les autres langues.

12 *www.ethnologue.com*

[13] Distribution de l'isixhosa sur la Toile selon le contenu



3.4.2 Correcteurs d'orthographe isixhosa

L'isixhosa est écrit de manière dite « conjonctive » (voir Prinsloo et De Schryver 2002). En conséquence, le nombre de mots différents, ou donc le nombre de *types*, est très élevé. Ceci ne facilite pas la tâche d'un correcteur d'orthographe, comme on peut le voir dans [14] et [15].

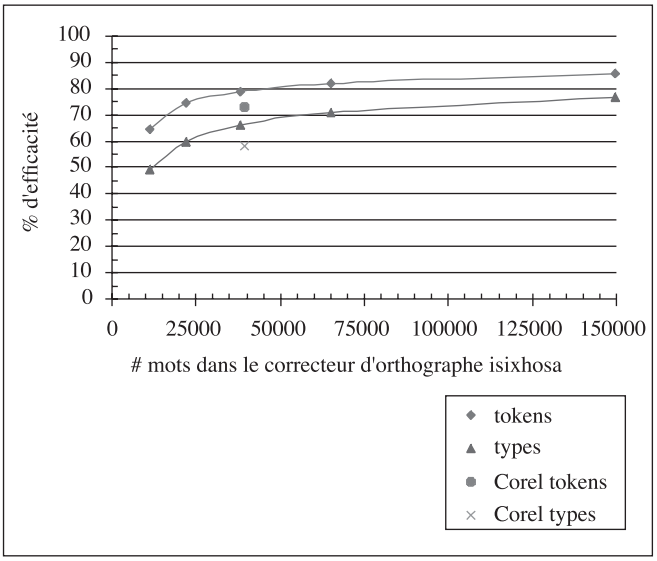
Même avec un total de 150 000 mots, l'efficacité n'atteint qu'un peu plus de 85 %.

Pour l'isixhosa, un correcteur d'orthographe dit « de première génération » (voir Prinsloo et De Schryver 2001 : 129) est inclus dans le logiciel de traitement de textes *Corel WordPerfect 9* distribué en Afrique du Sud depuis l'année 2000. Ce correcteur contient 39 192 mots. L'efficacité de ce correcteur a été calculée sur le même texte, à savoir la *Déclaration universelle des droits de l'homme*, et on a trouvé que la reconnaissance était un peu moins bonne, comme on peut le voir dans [15].

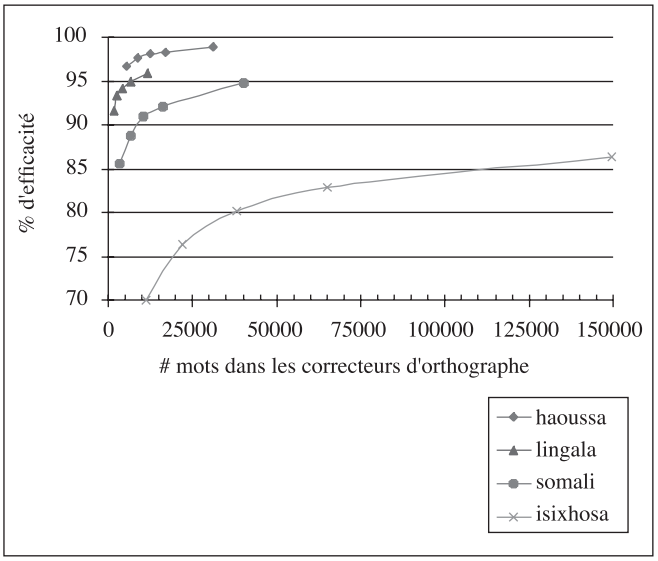
[14] Construction d'un correcteur d'orthographe isixhosa

Correcteur d'orthographe isixhosa (dérivé de 943 772 *tokens*)			*Déclaration universelle des droits de l'homme* isixhosa (1 196 *tokens* ; 705 *types*)				
mots dans chaque niveau			pas reconnu		% d'efficacité		
fréquence	nombre	%	*tokens*	*types*	*tokens*	*types*	utilisateurs
10 ou plus	11 026	7,37	422	358	64,72	49,22	70,07
5 jusqu'à 9	11 123	7,44	302	282	74,75	60,00	76,42
3 et 4	16 069	10,74	252	238	78,93	66,24	80,10
2	26 902	17,99	216	206	81,94	70,78	82,78
1 (hapax)	84 433	56,46	172	164	85,62	76,74	86,29
	149 553	100,00					

[15] Efficacité d'un correcteur d'orthographe isixhosa



[16] Comparaison de l'efficacité des correcteurs d'orthographe pour les quatre langues



4 Conclusion

Dans cette contribution nous avons démontré que les langues africaines sont non seulement présentes sur la Toile – et certaines même en force – mais aussi qu'il n'y a aucune raison technique qui devrait empêcher le téléchargement d'encore plus de documents et l'augmentation du nombre de langues utilisées, tout cela en utilisant les orthographes correctes. L'étude du contenu de sites haoussa, somali, lingala et isixhosa semble indiquer qu'à l'Ouest et qu'à l'Est du continent les actualités sont favorisées, tandis que pour la zone bantoue on remarque une grande attention pour la religion. Des échantillons d'autres langues africaines confirment cette tendance.

En naviguant sur la Toile le même nombre de jours (quatre) pour le haoussa, le somali, le lingala et l'isixhosa, des corpus linguistiques d'environ 850 000, 300 000, 200 000 et 950 000 mots (*tokens*) ont pu être rassemblés. Dû à des différences entre les structures de ces langues, et pour les langues bantoues à des différences entre les degrés de conjonction, le nombre de mots différents (*types*) est très varié: respectivement 30 000, 40 000, 10 000 et 150 000.

Cette dernière série de mots a ensuite été employée pour la fabrication de correcteurs d'orthographe. Une comparaison de l'efficacité de ces correcteurs d'orthographe, du point de vue de l'utilisateur, en employant la *Déclaration universelle des droits de l'homme* pour les quatre langues de l'échantillon, est présenté dans [16].

Il est clair que pour des langues comme le haoussa et le somali, ainsi que pour des langues bantoues avec une orthographe disjonctive, des logiciels de vérification basés sur des « mots orthographiques » sont tout à fait réalisables. Par contre, pour des langues bantoues avec une orthographe conjonctive, comme l'isixhosa (mais aussi l'isizulu, l'isindebele, le siswati, etc.), un logiciel de vérification ne sera satisfaisant qu'en employant des approches qui tiennent compte de la morphologie de ces langues. Une piste qui semble prometteuse est d'utiliser des machines à états finis (comme cela a été fait pour par exemple le finlandais). Des projets actuels et similaires traitent des langues comme l'amharique, l'arabe et l'isizulu.

En conclusion il est manifeste que les langues africaines méritent et prennent déjà leur place sur la Toile et il est réjouissant que le traitement automatique du langage naturel (TALN) pour les langues africaines soit bien en cours.

Anneleen Van der Veken,
Linguistique africaine, Université libre de Bruxelles, Bruxelles, Belgique.
avdveken@ulb.ac.be

Gilles-Maurice de Schryver,
Département de langues et cultures africaines, Université de Gand, Gand, Belgique.
gillesmaurice.deschryver@ugent.be

Bibliographie

BBC, 2003 : *News in 43 languages, Arabic*, news.bbc.co.uk/hi/arabic/news.

Bisharat, 2003 : *Initiative langues – technologie – développement*, www.bisharat.net.

Chanard (Chr.) et Popescu-Belis (A.), 2001 : « Encodage informatique multilingue : application au contexte du Niger », dans *Les Cahiers du Rifal*, nº 22, p. 33-45.

De Schryver (G.-M.), 2002 : « Web for/as Corpus : A Perspective for the African Languages », dans *Nordic Journal of African Studies*, nº 11 (2), p. 266-282.

De Schryver (G.-M.) et Lepota (B.), 2001 : « The Lexicographic Treatment of Days in Sepedi, or When Mother-Tongue Intuition Fails », dans *Lexikos*, nº 11 (Afrilex-reeks/series 11 : 2001), p. 1-37.

Deutsche Welle, 2003 : *News in 31 languages, Amharic*, www.dwelle.de/amharic/Welcome.html.

Fast Search & Transfer, 2003 : *moteur de recherche « All the Web »*, www.alltheweb.com.

Google, 2003 : *moteur de recherche « Google »*, www.google.com.

Grefenstette (Gr.), 2002 : « The WWW as a Resource for Lexicography », dans Corréard (Marie-Hélène), éd., *Lexicography and Natural Language Processing : A Festschrift in Honour of B.T.S. Atkins*, Euralex, p. 199-215.

Grefenstette (Gr.) et Nioche (J.), 2000 : « Estimation of English and non-English Language Use on the WWW », dans *Proceedings of Riao 2000*, Paris, 12-14 Avril 2000.

Grimes (B. F.) et Grimes (J. E.), éd., 2000 : *Ethnologue. Languages of the World, 14th Edition*, Dallas : SIL International. Voir aussi www.ethnologue.com

IPPmedia, 2003 : *Alasiri, Afternoon news from Dar es Salaam, Tanzania*, ippmedia.com/alasiri.htm.

Prinsloo (D. J.) et De Schryver (G.-M.), 2001 : « Corpus applications for the African languages, with special reference to research, teaching, learning and software », dans *Southern African Linguistics and Applied Language Studies*, nº 19 (1-2), p. 111-131.

Prinsloo (D. J.) et De Schryver (G.-M.), 2002 : « Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages », dans *Nordic Journal of African Studies*, nº 11 (2), p. 249-265.

Prinsloo (D. J.) et De Schryver (G.-M.), 2003 : « Towards Second-Generation Spellcheckers for the South African Languages », dans De Schryver (Gilles-Maurice), éd., *Tama 2003 South Africa : Conference Proceedings*, Pretoria : $(SF)^2$ Press, pp. 135-141.

SomiTek (Somali Information Technology), 1999 : *Hikaadiye, a complete wordprocessor with Somali and English spell checking capability*, www.somitek.com.

United Nations Organisation, 1948 : « Universal Declaration of Human Rights », Adopted and proclaimed by General Assembly resolution 217 A (III) of 10 December 1948, www.unhchr.ch/udhr.

Wolff (E. H.), 1991 : « Standardization and varieties of written Hausa (West Africa) », dans Von Gleich (U.) et Wolff (E. H.), éd., *Standardization of national languages. Symposium on language standardization, 2-3 February 1991*, Hamburg : UEI reports, Unesco Institute for Education, p. 21-32.