# JEP-TALN-RECITAL 2012

JEP : Journées d'Études sur la Parole
TALN : Traitement Automatique des Langues Naturelles
RECITAL : Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues

---

Actes de la conférence conjointe JEP-TALN-RECITAL 2012

Atelier TALAf 2012: Traitement Automatique des Langues Africaines

---

**Éditeurs**
Chantal Enguehard
Mathieu Mangeot
Gilles Sérasset

4 – 8 Juin 2012
Grenoble, France

Des versions imprimées de ces actes peuvent être achetées auprès de :

GETALP-LIG
Laurent Besacier
BP 53
38041 Grenoble Cedex 9
France
Laurent.Besacier@imag.fr

# Actes de l'atelier sur le traitement automatique des langues africaines : écrit et oral

## TALAf-2012

Organisé au sein de la conférence JEP-TALN 2012
Le 8 juin 2012 à Grenoble, France

*Mathieu Mangeot*[1] *Chantal Enguehard*[2]
(1) GETALP-LIG, BP 53 F-38042 Grenoble Cedex 9
(2) LINA, BP 92208, F-44322 Nantes Cedex 03
Mathieu.Mangeot@imag.fr, Chantal.Enguehard@univ-nantes.fr

# Préface

## 1 Motivations et objectifs

Les recherches en traitement automatique des langues africaines sont actuellement à l'orée de développements majeurs. Les efforts de reconnaissance des langues nationales et de standardisation des différents alphabets commencent à porter leurs fruits. Au Niger, par exemple, les alphabets des langues fulfulde, haoussa, kanouri, songhai-zarma et tamajaq ont été définis par des arrêtés du gouvernement en 1999. Par ailleurs, des collègues formés dans les pays du Nord reviennent dans leur pays avec la volonté de continuer les recherches sur les langues locales.

Pour autant, les langues nationales de la plupart des pays d'Afrique sont peu dotées : les ressources électroniques disponibles sont rares, mal distribuées, voire inexistantes. Seules sont accessibles les fonctions d'édition et d'impression, ce qui rend difficile l'usage de ces langues à l'écrit. Au moment où il est question de les introduire dans le système éducatif, de créer des normes d'écriture standardisées et stabilisées et surtout de développer leur usage dans l'administration et la vie quotidienne, le développement et la distribution d'outils dédiés ces langues s'imposent comme une nécessité cruciale.

Développer le traitement automatique de langues africaines nécessite l'élaboration de ressources qui seront les fondements à partir desquels des traitements plus élaborés peuvent être construits. Il apparaît indispensable de constituer en premier lieu des corpus écrits et oraux annotés aussi larges que possibles. À partir de tels corpus, il est possible d'extraire des exemples pour aider à la constitution de dictionnaires ou de mettre au point des modèles de langage pour la reconnaissance vocale. Toutefois, la constitution de tels corpus reste une entreprise délicate dans le contexte de langues peu dotées car, d'une part les transcriptions souffrent du manque de standardisation de la langue et, d'autre part l'enrichissement de corpus reste très onéreux.

Des astuces peuvent parfois être inventées pour pallier le manque de ressources. Par exemple, s'il n'existe pas de corpus oraux avec transcriptions, il est possible de constituer un corpus oral de textes lus.

Enfin, il y a lieu de prendre en compte les contraintes socio-économiques s'exerçant sur la population des locuteurs : les ressources économiques sont limitées, les ressources

humaines qualifiées sont rares, les recherches sont sporadiques et isolées, les résultats confidentiels et parcellaires. Il est donc nécessaire de définir des méthodologies économes en coût d'achat de logiciels et en temps de travail qualifié visant à produire des résultats pérennes, partagés et faciles à enrichir. La constitution de ressources linguistiques de manière générale, et plus encore pour les langues africaines devrait donc respecter plusieurs principes : utilisation d'outils gratuits en source ouverte, définition et utilisation de standards (ISO, Unicode), transfert de connaissances entre les collègues des pays du Nord et du Sud, disponibilité des ressources sous licence ouverte (Creative Commons), etc.

Cet atelier a pour but d'effectuer un état des lieux des travaux de constitution de ressources linguistiques de base (dictionnaires, corpus oraux et écrits), de mettre au point des méthodologies simples et économes d'élaboration de ressources, d'échanger sur les techniques permettant de se passer de certaines ressources inexistantes et d'envisager la direction des futurs travaux dans le domaine.

## 2 Présentation des articles

L'atelier a reçu douze soumissions. Onze articles ont été rédigés en français et un en anglais.

Parmi ces articles, cinq ont été acceptés en première lecture, et cinq acceptés après révision. Parmi ceux-ci, huit articles portent sur l'écrit et deux sur l'oral.

De plus, Mame Thierno Cissé, Professeur à l'Université Cheikh Anta Diop de Dakar, conférencier invité, interviendra pour présenter une base de données lexicale multifonctionnelle : le dictionnaire unilingue wolof et bilingue wolof-français.

La diversité linguistique est présente puisque quatorze langues figurent dans les articles acceptés : ahmarique, amazighe, arabe, bambara, français, haoussa, ikota, kanouri, mbochi, soŋay-zarma, swahili, tamajaq, wolof, yorouba.

Les auteurs se répartissent entre huit pays : Burkina-Faso (1), Canada (1), Ethiopie (1), France (16), Mali (1), Maroc (2), Niger (10), Sénégal (2), Tunisie (2).

Les articles acceptés se regroupent autour de trois thèmes principaux :

### 2.1 Traitement de l'oral

– Hadrien Gelas, Solomon Teferra Abate, Laurent Besacier et François Pellegrino *Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées à morphologie riche*

Cet article traite de la reconnaissance automatique de la parole pour l'amharique et le swahili, deux langues peu dotées à morphologie riche en utilisant des unités de découpage au niveau du morphème et de la syllabe.

– Annie Rialland, Martial Embanga Aborobongui, Martine Adda-Decker et Lori Lamel *Mbochi : corpus oral, traitement automatique et exploration phonologique*

Cet article décrit la constitution d'un corpus oral en langue mbochi, langue bantou parlée

au Congo-Brazzaville. Le corpus a été transcrit puis aligné automatiquement en mots et en segments phonémiques afin de permettre des études acoustico-phonétiques et phonologiques à grande échelle.

## 2.2 Dictionnaires et systèmes d'écriture

– Abdoulkarim Chérif Ari, Arimi Boukar, Kevin Anthony Jarrett, Maï Moussa Maï, Manoua Djibir, Taweye Aïchéta Chégou Koré *Élaboration d'un dictionnaire bilingue kanouri-français*

Cet article présente la langue kanouri avec sa place dans les différentes classifications, sa typologie et son système verbal. Il présente également le dictionnaire kanouri-français de 6 000 entrées élaboré lors du projet SOUTÉBA puis informatisé lors du projet DiLAF.

– Chantal Enguehard, Soumana Kané, Mathieu Mangeot, Issouf Modi et Mamadou Lamine Sanogo *Vers l'informatisation de quelques langues d'Afrique de l'Ouest*

Cet article présente le projet DiLAF qui vise à convertir des dictionnaires éditoriaux au format XML et à les mettre à disposition en ligne sur une plate-forme spécialisée. Il s'agit de dictionnaires bilingues langue africaine-français : haoussa-français, kanouri-français, soŋay zarma-français, tamajaq-français et bambara-français.

– Bernard Gautheron et Antonia Simon-Colazo *La transcription phonétique au bout des doigts, claviers et polices ergonomiques pour la transcription en API*

Le but de cet article est de promouvoir des outils ergonomiques qui facilitent la transcription phonétique manuelle pour les langues qui ne disposent pas encore de traitement automatique. L'utilisation d'un clavier ergonomique et d'une fonte phonétique spécifique à l'alphabet phonétique international (API) et dédiée à chaque langue facilite l'accès à tous les signes API nécessaires.

– Rahma Sellami, Fatiha Sadat et Lamia Hadrich Belguith *Extraction de lexiques bilingues à partir de Wikipédia*

Cet article présente une approche d'extraction de lexiques bilingues pour les paires de langues arabe-français et yorouba-français à partir de l'encyclopédie en ligne Wikipédia.

## 2.3 Analyse lexicale et syntaxique

– Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean et Emmanuel Schang *Décrire la morphologie des verbes en ikota au moyen d'une métagrammaire*

Dans cet article, le concept de métagrammaire introduit par Candito pour les grammaires d'arbres adjoints décrivant la syntaxe du français et de l'italien, est appliqué à l'ikota, langue bantoue parlé au Gabon. Le formalisme XMG (eXtensible MetaGrammar) est utilisé pour décrire les variations morphologiques des verbes.

– Abdou Mijinguin et Harouna Naroua *Règles de formation des noms en haoussa*

Cet article présente quelques caractéristiques du fonctionnement lexical du haoussa comme les règles de formation de mots haoussa à partir des racines. Cette analyse a

permis aux auteurs de formuler plusieurs règles de flexion et de dérivation pouvant être utilisées pour construire des outils de traitement automatique.

– Mar Ndiaye et Chérif Mbodj *Vers un analyseur syntaxique du wolof*

Cet article présente un projet d'analyseur syntaxique du wolof (parlé au Sénégal, en Mauritanie et en Gambie) basé sur l'outil FIPS développé au LATL à Genève à partir de grammaires GB.

– Fatima Zahra Nejme et Siham Boulaknadel *Formalisation de l'amazighe standard avec NooJ*

Dans la suite des travaux de standardisation de l'amazighe effectués au Maroc par l'Institut Royal de la Culture Amazighe (IRCAM), cet article présente la construction d'un module Nooj de formalisation de règles morphologiques pour la catégorie nom permettant de générer son genre, son nombre, et son état.

## 3 Comité de programme

Laurent Besacier (LIG, Grenoble, France)

Mame Thierno Cissé (ARCIV, Université Cheikh Anta Diop, Dakar, Sénégal)

Chantal Enguehard (LINA, Nantes, France)

Gil Francopoulo (Tagmatica, Paris, France)

Hadrien Gelas (DDL, Lyon, France)

Mathieu Mangeot (LIG, Grenoble, France)

Chérif Mbodj (Centre de Linguistique Appliquée de Dakar, Sénégal)

Kamal Naït-Zerrad (INALCO, Paris, France)

Harouna Naroua (Université Abdou Moumouni, Niamey, Niger)

Pascal Nocera (Université d'Avignon, France)

Guy De Pauw (Université d'Anvers, Belgique)

Francois Pellegrino (DDL, Lyon, France)

Mamadou Lamine Sanogo (INSS, Ouagadougou, Burkina-Faso)

Gilles Sérasset (LIG, Grenoble, France)

## 4 Conclusion

Le nombre important de soumissions dans la thématique de l'atelier montre que la nécessité du traitement automatique des langues africaines est toujours d'actualité et que des travaux de recherche sont en cours. En revanche, les travaux restent épisodiques, éparpillés et espacés dans le temps. Il apparaît donc nécessaire de regrouper ces efforts en mettant en place, par exemple, des entrepôts de données libres sous licence ouverte (Creative Commons) comme dans le projet DiLAF. Les savoirs et savoirs-faire doivent également être capitalisés pour resservir pour d'autres langues et d'autres contextes.

# Table des matières