

Guillaume SEGERER
UMR 8135 - LLACAN
"Langage, langues et cultures d'Afrique Noire"
Directrice : Martine Vanhove

CR2
Section 34
(n° d'agent : 00004459)

Rapport d'activités
(septembre 2007 - septembre 2008)

17 février 2009

***Message préliminaire:** Ce rapport est rédigé par devoir malgré les difficultés actuelles que traversent les chercheurs en France et dans un état d'indignation exceptionnel par rapport à la dégradation graduelle de nos conditions de travail, l'aggravation de la pénibilité de notre métier de chercheur multitâche, l'incertitude qui pèse depuis plus d'un an sur le devenir du CNRS et de l'Université ainsi que sur le statut de la profession pour laquelle nous avons été recrutés et avons souhaité rentrer dans la fonction publique, et enfin par rapport aux propos calomnieux tenus par le président de la République à propos de la recherche et de l'université françaises.*

Ce rapport est établi au titre de mon statut d'ingénieur d'études **détaché dans le corps des chargés de recherche** pour la période du 01/09/2007 au 31/08/2008 et de **chercheur statutaire** depuis le 01/09/2008, soit, pour ce qui concerne le rapport, pour la période du 1er au 30 septembre 2008.

A. RAPPORT D'ACTIVITE

A.1. CURRICULUM VITAE

ETAT-CIVIL

Nom Prénom : **SEGERER Guillaume**
Date et lieu de naissance : 13 juillet 1965 à Paris 14e
Adresse professionnelle : CNRS-UMR 8135 LLACAN
"Langage, langues et cultures d'Afrique Noire"
BP 8, 7 rue Guy Môquet, 94801 Villejuif Cedex
Téléphone : bureau: 01-49-58-36-96, personnel: 01-43-48-17-22
email : segerer@vjf.cnrs.fr

TITRES ET DIPLOMES UNIVERSITAIRES

Doctorat de Sciences du Langage	(Paris 3)	2000
<i>Description de la langue bijogo (Guinée Bissau)</i> , sous la direction de France Cloarec-Heiss. Jury composé de : France Cloarec-Heiss, Raphaël Kabore, Serge Sauvageot, Denis Creissels, Konstantin Pozdniakov. Thèse soutenue en Sorbonne le 03/04/2000.		
DEA de Sciences du Langage	(Paris 3)	1996
Maîtrise de Sciences du Langage	(Paris 3)	1995
Maîtrise de Linguistique Africaine	(Paris 3)	1995
Licence de Sciences du Langage	(Paris 3)	1994
Licence de Linguistique Africaine	(Paris 3)	1994

FONCTIONS AU LLACAN

Chargé de recherches (CR2)	depuis 09/2008
Ingénieur d'études, détaché dans le corps des chargés de recherche	09/2007 - 08/2008
Ingénieur d'études (IE2)	2002-2007
Informaticien (CDD)	11/2001-06/2002
Informaticien (CDD)	09/2000
Informaticien (CDD)	07/1999-09/1999
Informaticien (CDD)	05/1998-12/1998
Informaticien (CDD)	05/1997-10/1997
Informaticien (Contrat Emploi Solidarité)	07/1996-12/1996
Informaticien (Contrat Emploi Solidarité)	01/1996-06/1996

A.2. RECHERCHE SCIENTIFIQUE

Préambule

A l'heure où ces lignes sont écrites, je suis théoriquement « intégré dans le corps des chargés de recherche ». Cependant, aucune notification formelle officielle ne m'a été faite. Cette intégration fait suite à une période de détachement, qui m'a été accordé pour une durée d'un an alors que ma demande portait sur une période de deux ans. En outre, pour pouvoir prétendre à cette intégration, j'ai dû rédiger un rapport sur les résultats de mes recherches à l'issue de mon détachement, rapport qui devait être soumis au printemps 2008, soit 6 mois après le début du détachement (supposé au départ durer 2 ans). Aujourd'hui je rédige un rapport « à 4 ans » sur mon activité de chercheur alors que mes fiches de paie mentionnent toujours le grade d'IE2. Ce rapport concerne donc théoriquement la période du 1er septembre 2007 au 30 septembre 2008, mais il fera également état d'activités plus récentes.

Présentation

Mes recherches ont pour objet principal un groupe d'une cinquantaine langues parlées sur les côtes de l'ouest de l'Afrique, entre le Sénégal et le Liberia, et connues sous l'appellation de langues **atlantiques**. Quelques-unes de ces langues sont maintenant bien décrites, comme le peul ou le wolof, mais la plupart sont encore très pauvrement documentées. Mon travail suit donc plusieurs directions :

- la description de langues, à partir d'enquêtes de terrain
- la collecte de la documentation existante sur ces langues
- l'harmonisation des données lexicales pour comparaison
- la comparaison et la reconstruction dans plusieurs sous-groupes atlantiques
- le traitement quantitatif des données linguistiques

Dans l'optique comparatiste qui est la mienne, j'accorde une importance particulière aux aspects quantitatifs de l'étude des structures des langues. En effet, la masse des données disponible a considérablement augmenté ces dernières années, et les possibilités de traitement automatique ont été largement simplifiées et rendues facilement accessibles. Et puis, mon expérience comme ingénieur m'a permis d'acquérir des compétences utiles dans le traitement de l'information et les bases de données. C'est donc tout naturellement que je cherche maintenant à mettre au point des outils qui puissent me permettre d'étendre la portée et l'efficacité de mes travaux comparatistes.

a. La description

Ma thèse, soutenue en 2000, portait sur la description de la langue *bijogo*, qui constitue à elle seule un rameau isolé du groupe atlantique. Aujourd'hui, mon intégration dans le corps des chargés de recherche me permet de reprendre le travail de terrain. Ainsi, avec ma collègue Stéphane Robert, nous avons effectué en janvier-février 2009 une mission de terrain au sud du Sénégal, pour travailler sur la langue *kerak*. Cette langue peut être présentée comme une variété de *joola*, le *joola* étant, selon les auteurs, une langue unique, un continuum dialectal ou un sous-groupe comportant entre 6 et 15 langues différentes. C'est aussi pour éclaircir cette position confuse du *joola* que nous avons choisi le *kerak*, qui au sein de l'ensemble *joola* était réputé marginal, bien que n'ayant jamais fait l'objet d'investigations poussées. En outre, le *kerak*, comme la plupart des variétés de *joola*, est menacé à court terme : la langue n'est plus parlée que dans un seul village (Kabrousse), qui compte environ 2500 habitants. Cette description vise donc non seulement à remplir une lacune importante de la documentation, mais aussi à mieux comprendre les rapports génétiques entre les langues de cette zone, qui est,

linguistiquement, l'une des plus morcelées du continent : sur un espace correspondant à un département français coexistent au moins une vingtaine de langues, dont la plupart appartiennent, comme le kerak, au groupe BAK des langues atlantiques.

Je suis également engagé dans le projet Sénélangues, porté par Stéphane Robert et soumis en novembre dernier à l'ANR, qui a précisément pour objectif la description et la documentation du patrimoine linguistique du Sénégal.

b. La comparaison

Parallèlement aux projets plus ponctuels de description, je mène depuis de nombreuses années des travaux de comparaison et de reconstruction, prioritairement dans le domaine atlantique, mais avec, en arrière-plan, des vues plus vastes sur le phylum Niger-Congo dans son ensemble.

Il y a dans cette activité une dimension technique indispensable. Concrètement, cela consiste à réunir et organiser la documentation existante, en assurer la saisie et la numérisation, et surtout permettre l'exploitation de ces données en les organisant de manière homogène. Un des premiers résultats concrets de cet ambitieux effort est la mise en ligne, au printemps 2008, d'une vaste base de données bibliographique (<http://sumale.vjf.cnrs.fr/Biblio/index.html>) comportant près de 25000 références sur les langues africaines et assortie d'une interface de recherche multi-critère ainsi que d'un module cartographique. J'ai également mis au point des outils d'aide à la décision en matière de comparaison lexicale, que j'ai testés et utilisés sur deux sous-ensembles atlantiques : les langues *joola* (voir plus haut) et les langues *tenda* (parlées dans l'extrême sud-est du Sénégal). Ces outils, qui font l'objet d'un développement permanent, visent à combiner la méthode comparative classique et la puissance de calcul désormais accessible.

Quant à l'aspect proprement scientifique, il ne peut se développer pleinement que sur le long terme, mais déjà trois pistes sont explorées :

- D'une part, les outils mentionnés ci-dessus permettent d'appliquer de manière systématique la méthode comparative à des ensembles de langues pour lesquels cela n'avait jamais été tenté (j'ai mentionné les ensembles *joola* et *tenda*, mais d'autres suivront). Il importe de signaler ici que les langues atlantiques présentent des traits de structure qui rendent la comparaison parfois extrêmement problématique. Par exemple, si toutes les langues atlantiques connaissent la classification nominale (c'est-à-dire la répartition des noms en diverses classes morphologiquement marquées et provoquant des phénomènes d'accord), les détails des systèmes de classes varient fortement d'une langue à l'autre. Ainsi, les marques de classes sont tantôt préfixées (bijogo...), tantôt suffixées (peul...), et parfois n'apparaissent même plus de façon manifeste sur le nom (wolof). En outre, les paradigmes de marques de classes peuvent avoir subi des processus d'unification touchant les consonnes (*tenda*) ou les voyelles (wolof, *joola*), ce qui fait qu'entre deux langues ayant adopté des stratégies inverses, plus rien n'est comparable.

- D'autre part, la disponibilité sous forme numérique de vastes corpus lexicaux permet une approche plus quantitative des données linguistiques, que j'ai abordée suivant deux angles distincts.

c. L'approche quantitative des données linguistiques

Il s'agit ici d'effectuer des mesures quantitatives sur d'importantes quantités de données, qui permettent de faire émerger des structures dont l'échelle est habituellement indécélable par les méthodes d'examen habituelles. Une bonne image pour illustrer cette approche est celle de l'archéologie aérienne : vu du sol, un champ labouré est un champ labouré. Mais sur des photos prises d'assez haut, on peut parfois y déceler des traces d'anciennes constructions

romaines, par exemple¹. Ainsi, avec mon collègue Konstantin Pozdniakov, nous avons mis en évidence, par des méthodes statistiques, des restrictions sur les combinaisons de consonnes au sein des séquences -CVC- : les séquences mettant en jeu des consonnes de même point d'articulation sont statistiquement rares. Ces restrictions, que nous pensions au départ caractéristiques des langues atlantiques, se sont révélées opérantes pour toutes les langues que nous avons examinées, accédant ainsi au statut de trait universel statistique. Ces résultats ont été publiés dans la prestigieuse revue *Linguistic Typology*².

L'approche quantitative que je commence à développer est également orientée vers la compréhension des phénomènes de nature sémantique. Avec Martine Vanhove et Bruno Gaume, nous avons soumis l'application PROX (développée à Toulouse par Bruno Gaume pour modéliser l'espace sémantique du français à partir des relations de synonymie³) à un ensemble de données lexicales représentant le groupe atlantique dans son ensemble. Cette expérience a révélé de profondes ressemblances de structure entre les graphes issus de relations synonymiques et les graphes issus de relations polysémiques. Ces résultats ont été exposés au 7e congrès de l'ALT (*Association for Linguistic Typology*).

Positionnement dans l'unité

Tous mes travaux s'inscrivent dans le thème 2 du laboratoire, intitulé « Genèse des langues », et plus particulièrement dans l'opération de recherche « Comparatisme : problématique et enjeux ».

Publications (sept. 2007 - sept. 2008)

5. ARTICLES DANS DES REVUES SANS COMITÉ DE LECTURE, ARTICLES DE SYNTHÈSE ET CONTRIBUTIONS À DES OUVRAGES DE SYNTHÈSE

- 2008 : Bijogo, in Tröbs, Holger, Eva Rothmaler & Kerstin Winkelmann, *La qualification dans les langues africaines / Qualification in African languages*, Köln : Rüdiger Köppe Verlag.

7. AUTRES (COMMUNICATIONS À DES COLLOQUES, NON PUBLIÉS)

- 2008 : Londres, SOAS, ESF Exploratory Workshop : Documenting convergence and diversity - Mandé and Atlantic languages in contact, 5-9 septembre 2008, **invité**.
– *Peut-on faire l'histoire du contact linguistique ?*
- 2008 : Porquerolles, Ecole d'été du Cercle Linguistique de l'INALCO, 25-30 mai 2008, **invité**.
– *Bases de données et méthodes statistiques*
- 2008 : Lyon, colloque 'New Directions in Historical Linguistics' organisé par J.M. Hombert dans le cadre du projet OMLL de l'ESF, 12-14 mai 2008, **invité**.
– *Niger-Congo as a playground for Lexical Comparison*
- 2007 : Paris, ALT 7 (7e colloque de l'Association for Linguistic Typology), 25-28 septembre.
– (avec Martine Vanhove, Karine Duvignau et Bruno Gaume) *From lexical semantics to semantic networks: a computational and typological study*

¹ voir par exemple le site de Ribemont en Picardie : <http://www.imagesdepicardie.com/portrait/agache/ribemont.htm>.

² POZDNIAKOV, Konstantin & Guillaume SEGERER, 2007. Similar Place Avoidance: A Statistical Universal. *Linguistic Typology* 11-2, pp. 307-348.

³ <http://erss.irit.fr:8080/graph/>

A.3 – ENSEIGNEMENT, FORMATION ET DIFFUSION DE LA CULTURE SCIENTIFIQUE

J'ai organisé, le 24 septembre 2007, un atelier « Typologie des langues africaines », adjacent au 7e congrès de l'ALT qui s'est tenu à Paris du 25 au 28 septembre 2007. 8 communications y étaient présentées, dont 4 par des conférenciers invités et 4 par des auteurs sélectionnés.

B. OBJECTIFS

Mes objectifs pour les prochaines années sont en fait contenus dans la description de mes activités récentes. En effet, les domaines abordés, les méthodes mises en oeuvre et les outils développés représentent des champs d'activité relativement nouveaux.

a. Description. La description du *kerak*, qui vient de commencer, sera poursuivie. Une première mission a permis de mettre au jour les structures les plus saillantes de la langue : bases de la phonologie, morphologie nominale, éléments de conjugaison. Mais de nombreux points demeurent intouchés. En fonction de l'avancée de ce projet, il pourra être envisagé, au cours des 4 années à venir, de démarrer la description d'une autre langue, parmi les langues atlantiques les moins connues : *sua*, *nalu*, *eramme* par exemple. Ma participation au projet **Sénélangues**, soumis à l'ANR par S. Robert, me conduira, si celui-ci est retenu, à participer à la description d'autres langues du groupe BAK (*joola*, *balante*, *manjaku*), notamment par un travail de synthèse de la documentation existante mais également, en cas de besoin, par des enquêtes de terrain complémentaires.

b. Comparaison. Les données obtenues à l'occasion des missions de terrain viendront enrichir le corpus déjà très important élaboré sur les langues atlantiques. Les outils déjà développés seront plus systématiquement exploités et appliqués à ce corpus, mais ils seront enrichis de nouvelles fonctionnalités. Ces deux parties du travail de comparaison (extension du corpus, développement d'outils) pourraient être grandement facilitées par l'obtention d'un financement ANR, sollicité dans le cadre de l'appel à projet SHS franco-allemand 2009 (il s'agit du projet RefLex, voir ci-dessous).

a&b. Rédaction d'un ouvrage sur les langues BAK. Les travaux menés en description et en comparaison me permettront de rédiger un ouvrage de référence sur les langues BAK, qui représentent un ensemble particulier au sein du groupe atlantique. C'est l'ensemble le plus diversifié, mais c'est aussi, géographiquement, l'ensemble central. Il semble également, d'après ce que l'on peut actuellement supposer sur la nature linguistique d'un hypothétique proto-atlantique, que ce soit également un ensemble assez conservateur. Sa connaissance détaillée est donc de première importance pour l'établissement des relations génétiques entre les langues atlantiques.

c. Linguistique quantitative. Les résultats obtenus par des méthodes de traitement de masse (statistiques ou autres) m'encouragent à poursuivre dans cette voie. De nombreuses pistes sont encore à explorer. Par exemple, les travaux sur la combinatoire consonantique peuvent être étendus en prenant en compte des variables telles que la longueur vocalique, qui semble jouer un rôle majeur par rapport aux contraintes évoquées ci-dessus : d'après les premières mesures, de part et d'autre d'une voyelle longue, la combinaison des consonnes est beaucoup plus libre qu'autour d'une voyelle brève, dans les langues qui connaissent cette opposition. Les

conséquences théoriques de cette différence de comportement entre voyelles longues et brèves peuvent être importantes, et concernent aussi la sphère cognitive : à partir de quelle durée une émission vocalique a-t-elle pour effet d'effacer, en quelque sorte, le souvenir de la consonne qui la précède, de façon à ce que la consonne suivante puisse être « librement » choisie ? Est-ce que ce type de contrainte, mesurable, pourrait être un moyen de différencier des voyelles phonologiquement longues et des voyelles brèves réalisées allongées, par exemple en position accentuée ?

Ici aussi, les recherches bénéficieront de la réalisation du projet RefLex, qui va maintenant être détaillé.

d. RefLex : un projet ambitieux à vocation internationale. Dans le cadre de l'appel à projet SHS franco-allemand 2009⁴, je vais soumettre en avril prochain, avec mon collègue Gerritt Dimmendaal de l'Université de Cologne, un projet dénommé RefLex (pour *Reference Lexicon*). L'objectif principal de ce projet est de mettre en ligne, accessible à tous, un lexique de référence pour les langues d'Afrique. Par « lexique de référence », j'entends une base de données lexicales issue de documents publiés, dans laquelle chaque fiche lexicale individuelle sera absolument identifiable et vérifiable sur le document d'origine, dont une version numérisée sera également consultable. Le principe est de pouvoir disposer, à terme et en un endroit unique, de toutes les ressources lexicales existantes pour une langue ou un groupe de langue. Ce projet est extrêmement ambitieux, mais des spécialistes de toute l'Europe ont déjà répondu favorablement. Il s'agit d'un projet essentiellement coopératif, et si j'en suis l'instigateur, je n'ai pas particulièrement vocation à le diriger dans l'avenir.

Les objectifs scientifiques sont de trois types :

- établir une méthodologie de partage et d'unification des données
- permettre des avancées significatives dans les études comparatives des langues africaines
- faire émerger de nouvelles problématiques liées à la profusion d'informations lexicales standardisées

Parallèlement à la constitution du corpus lui-même, les membres du projet travailleront, ensemble ou séparément, à la mise au point d'outils de traitement de ces données. Ces outils pourront servir à la recherche de données, à des traitements statistiques basés sur la phonologie, la morphologie, la sémantique, à la cartographie, et à d'autres fonctions à découvrir. L'important est qu'un outil développé par une équipe membre devra pouvoir être exploité par une autre équipe, qui pourra à son tour faire évoluer l'outil, etc.

Les retombées attendues sont de plusieurs types. D'une part, toutes les équipes concernées constitueront un réseau à l'échelle européenne qui pourra développer d'autres projets coopératifs. A cette occasion, il est permis d'espérer qu'un véritable standard de structuration des données lexicales pourra voir le jour. D'autre part, les possibilités ouvertes par ce vaste corpus permettront des avancées significatives dans les disciplines où la comparaison joue un rôle important : typologie, classification, reconstruction.

⁴ <http://www.agence-nationale-recherche.fr/AAPProjetsOuverts?NodId=17&lngAAPIId=241>